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

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

S1 Complete Prompts and Generation Configuration

These supplementary materials accompany Periodic Locking and Topic Retention in Long Dialogues Between Two Small Language Model Agents. The original observational dataset contains 621 complete dialogues of 200 messages each and one incomplete dialogue of 161 messages. The prompts, parameters, metrics, and file identifiers below are drawn from the saved configuration, raw logs, and reproducible analyses. Message numbers index the alternating sequence across both agents A and B. All experimental prompts and generated messages were in Chinese. English prompt translations and lexical glosses in this document are provided for interpretation; exact reproduction must use the original Chinese strings in the archived configuration and analysis code, not these English translations.

S1.1 System Prompts and Fixed Opening Scenarios

Complete system prompt for the free condition (English translation): You are communicating with another conversational partner. Respond naturally to the other person and continue the conversation. The opening information provides only the initial background. Reply with only 1 to 3 sentences each time; keep it brief and do not provide an outline.

Complete system prompt for the topic-anchored condition (English translation): You are communicating with another conversational partner. Respond naturally to the other person and continue the conversation. The opening information provides only the initial background. Reply with only 1 to 3 sentences each time; keep it brief and do not provide an outline. Throughout the conversation, always stay focused on the opening information.

Table S1. The five fixed fictional opening scenarios, translated from Chinese. Both original prompt conditions use the complete corresponding scenario. Exact original strings are archived in runs/qwen_duet_v1/config.json.

Scenario ID	Name	Opening scenario (English translation)
0	Balcony	The following is a fictional scenario: an east-facing balcony holds three pots of mint, one of which has a blue label.
1	Reading room	The following is a fictional scenario: a reading room has four rows of bookshelves and a green chair beside the window.
2	Park	The following is a fictional scenario: a small park has two paths, a wooden table in the center, and an osmanthus plant nearby.
3	Town	The following is a fictional scenario: a town named Qinglan has seven bridges, and its residents enjoy exchanging used books on weekends.
4	Notebook	The following is a fictional scenario: a person has prepared a blank notebook, written the number 37 on the first page, and has not yet decided what to record.

The first user message for A consists of the opening scenario, a newline, and the instruction translated as "Please start a conversation based on this." The first user message for B consists of the same scenario, two newlines, the prefix translated as "The other person says:", and A's first reply. Thereafter, each agent records its own output as assistant and its partner's output as user. A always speaks first; agent identity therefore cannot be separated from the initial input structure or speaking order. The exact Chinese prompts and construction rules are retained in runs/qwen_duet_v1/config.json and run_dialogues.py.

S1.2 Parameters, Versions, and Original Seeds

Table S2. Generation parameters and execution environment for the observational experiment.

Item	Setting
Generative model	Qwen/Qwen3.5-0.8B
Model revision	2fc06364715b967f1860aea9cf38778875588b17
Loaded parameter count	752,393,024; frozen weights
Precision / attention	BF16 / SDPA
GPU	NVIDIA GeForce RTX 3070 Ti Laptop GPU
PyTorch / Transformers	2.13.0+cu130 / 5.15.0
temperature / top_p / top_k	0.8 / 0.95 / 20
repetition_penalty / thinking	1.0 / disabled
Maximum new tokens per message / context limit	128 / 32,768
Maximum input tokens among complete messages	25,979
History and cache	Full history; separate cache for each agent; no silent truncation
Input template	Append-only concatenation of official template fragments, retaining historical generated tokens and empty think markers
Semantic encoder	BAAI/bge-small-zh-v1.5
Encoder revision	7999e1d3359715c523056ef9478215996d62a620
Encoding	CLS vectors, L2 normalization, float32, no retrieval instruction; 512-token limit; no input truncation

In the original design, replicate index r starts at 0. Each block contains the five opening scenarios crossed with two prompt conditions. Python random.Random(20260926+r) shuffles the 10 combinations within each block. Scenario index j ranges from 0 to 4 as listed in Table S1. The two conditions with the same scenario and replicate index share an episode seed.

$$\text{seed}(r,j) = 20260926 + 100r + j;\ \text{sampling_seed}(r,j,t) = 1000 \times \text{seed}(r,j) + t,\ t = 1,\ldots,200.$$

This seed scheme supports reproduction and paired comparisons; identical random seeds under different prompts do not imply identical generated text. The planned collection comprised 1,000 episodes and was stopped after 621 complete dialogues. The first 62 complete blocks contain 620 episodes and 310 matched pairs. Only one episode, the free park condition, was completed in block 63. The complete original messages

were generated from 26 September 2026 at 19:41:57 to 3 October 2026 at 08:24:57, China Standard Time (UTC+08:00).

S2 Metrics, Equations, Denominators, and Lexical Dictionaries

S2.1 Repetition, Periodicity, and Near-Repetition Confirmation

Let y_t denote the raw text. Normalized text x_t is obtained through NFKC normalization, conversion to lowercase, and removal of all whitespace, while retaining punctuation. The primary strict-periodicity metric directly compares y_t without normalization or similarity scoring.

$$Q(p,w) = 1\{y_t = y_{t-p} \text{ for every } t = 201-w,...,200\};\ p^* = \min\{p \in \{1,...,10\}: Q(p,40) = 1\}.$$

If no period p satisfies the criterion, p* is set to 0. Reference messages used in a tail-window comparison may precede the window. Period 1 means that every visible reply is identical; period 2 means that two replies alternate. Finite textual periodicity does not require identical growing inputs, output tokenizations, or hidden states.

$$R_{2,200} = (1/200)\ \Sigma_{t=3}^{200}\ 1\{x_t = x_{t-2}\};\ R_{2,198} = (1/198)\ \Sigma_{t=3}^{200}\ 1\{x_t = x_{t-2}\}.$$

The overall self-repetition rate in the main text uses a denominator of 200 and assigns 0 to the first two messages, maintaining consistency with the original trajectory summaries. The denominator of 198 includes only comparable positions. Across all 621 episodes, the corresponding mean rates are 72.6948% and 73.4291%. All final 20 messages have reference messages, so their rate of 94.5974% is unaffected by this choice. Repetition of any earlier message indicates whether the current x_t equals any preceding x_u, where $u < t$.

$$s_t = (v_t \cdot v_{t-2})/(\|v_t\|\ \|v_{t-2}\|),\ t \geq 3;\ I_t(\theta) = 1\{s_t > \theta\}.$$

Here v_t is the character-bigram frequency vector of the normalized text. Similarity is set to 0 if either vector has zero norm; the first two messages do not qualify as near-repetition. The primary threshold is $\theta = 0.90$. For m consecutive positions satisfying $I_t = 1$, the onset τ of the first qualifying run and its confirmation position c are defined below. The primary analysis uses $m = 20$.

$$\tau = \min\{a: \Sigma_{u=a}^{a+m-1}\ I_u = m\};\ c = \tau + m - 1.$$ If no confirmation occurs, the bounded score is 201.

Recovery is defined as at least one occurrence of $s_t \leq \theta$ after confirmation and through message 200. The primary near-repetition criterion was confirmed in 587 episodes, at positions 27 to 186, with a median confirmation position of 61. Even the shortest remaining observation period contained 14 messages. The numbers of episodes with at least 20, 50, and 100 subsequent messages were 586, 583, and 542, respectively; none showed recovery.

S2.2 Novelty, Opening-Scenario Coverage, and Semantics

$$N_t = |G(x_t) \setminus (\cup_{u=1}^{t-1}\ G(x_u))| / \max(1, |G(x_t)|).$$

$G(x)$ is the set of distinct character bigrams in a text. N_t measures new textual fragments relative to the preceding history of that episode; it does not assess factual accuracy or communicative value. It is set to 0 for an empty set. As the cumulative historical vocabulary grows, new fragments become harder to introduce; N_t and strict repetition are therefore reported separately. Character length is the number of characters in the raw text. A length-cap event is identified from the recorded finish_reason=length.

Table S3. English glosses of the Chinese topic terms and fact-pattern entries. A topic mention requires any topic term; fact-term coverage is the number of matched regular-expression entries divided by the number of entries for that scenario. Semicolons separate entries; alternatives within an entry do not increase its count. These glosses are not executable English regexes. Exact Chinese terms and regular expressions are archived in deep_analysis/prepare.py.

Scenario	Topic terms: English glosses	Fact-pattern entries: English glosses
Balcony	balcony; mint; blue; label; facing east	facing east / eastward / east-facing; three pots / 3 pots; mint; blue followed by label (up to 5 intervening characters), or its shorter form (up to 3)
Reading room	reading room; bookshelf; green; chair; beside the window	reading room; four rows / 4 rows; bookshelf; green followed by chair (up to 5 intervening characters), or the contracted green-chair form; beside the window
Park	park; path; wooden table; osmanthus	park; two paths / 2 paths (up to 5 intervening characters between the numeral-and-classifier and path); wooden table; osmanthus
Town	Qinglan; small town; seven bridges; used books; exchange books	Qinglan; seven [bridge classifier] / 7 [bridge classifier]; bridge; weekend; used books
Notebook	notebook; first page; 37; thirty-seven	notebook; blank; first page; 37 / thirty-seven

Topic terms and fact-term regular expressions are matched by occurrence in normalized text; they do not detect negation, numerical relations, or truth. The string "37" may also match within longer numbers. These are therefore lexical-coverage measures, not factual-retention rates. Politeness, closing, and action expressions are searched in the raw text. English glosses of the original politeness terms are: thanks; gratitude; wish you (informal); wish you (formal); look forward to; next time; talk again; goodbye; bye-bye; good wishes; blessings. Closing-expression glosses are: goodbye; bye-bye; next time again; talk again; that is all for today; let us stop here; take my leave; see you later; good night. Action-loop expression glosses are: let us begin; let us set off; let us get started; let us act; ready; are you ready; start now; let us start; start together. These glosses are explanatory only; the exact Chinese alternation patterns are preserved in deep_analysis/prepare.py.

Semantic similarity to the opening scenario is the dot product of the unit-normalized BGE message embedding and the corresponding scenario embedding. The semantic margin is this similarity minus the largest similarity to any of the other four scenarios. A nearest-scenario hit occurs when the largest of the five similarities corresponds to the episode's own scenario. This direct nearest-vector measure differs from topic decoding, which requires a trained classifier. BGE encoded 27,865 distinct raw replies and five opening scenarios, yielding 27,870 distinct inputs.

The early predictors are means of 22 message-level metrics over the first 20 messages: lexical similarity to the opening scenario; lag-1 similarity; lag-2 similarity; lag-2 set Jaccard similarity; repetition of any earlier message; exact lag-1 repetition; exact lag-2 repetition; exact lag-2 repetition after retaining only alphanumeric characters; new-bigram proportion; character count; topic mention; fact-term coverage; politeness expressions; closing expressions; action-loop expressions; question marks; length-cap events; local near-repetition; the number of new bigrams adopted from the partner; semantic similarity to the opening scenario; semantic margin; and nearest-scenario hits. Every feature was checked by reaggregation from the message-level table and excludes information from message 21 onward.

The implemented local near-repetition feature near90 is defined as max(lag1 cosine, lag2 cosine) > 0.90, whereas the primary event uses only lag-2 similarity. "New bigrams adopted from the partner" counts distinct bigrams in the current message that the speaker has not previously used and that appear in the partner's immediately preceding reply. Alphanumeric filtering uses Unicode isalnum and therefore includes Chinese characters. The original lexical_seed feature compares a message with the direct concatenation of normalized initial_a and initial_b, so the same opening scenario occurs twice; the BGE opening-similarity

feature uses a single copy of the scenario. First-20-message means exclude undefined lag-similarity/Jaccard positions. For exact-repetition Boolean features, lag-1 is set to 0 at the first message, and lag-2 and alphanumeric lag-2 are set to 0 at the first two messages.

S2.3 Post Hoc Sensitivity to Numeric Term Boundaries

Within the notebook scenario only, comprising 124 episodes and 24,800 messages, the topic-dictionary string "37" was replaced by the numeric-boundary expression (?<!\d)37(?!\d). The other Chinese entries, glossed as "notebook," "first page," and "thirty-seven," were unchanged. This variant removes matches to the substring 37 in longer numbers such as "1937." It still does not assess negation, relations, or factual truth, and it is not a semantic parser for all numerical expressions. The original primary metric remains unchanged.

Table S4. Topic-term boundary sensitivity for the notebook scenario. Values are topic-term hit proportions (%); "boundary" denotes the numeric-boundary variant.

Window	Original: all	Boundary: all	Original: free	Boundary: free	Original: anchored	Boundary: anchored
First 20	50.44	50.32	49.19	49.03	51.69	51.61
Final 20	43.35	42.14	35.08	33.47	51.61	50.81
All 200	43.98	42.93	37.43	36.00	50.54	49.86

The boundary variant removed 3, 30, and 261 original topic hits from the first 20, final 20, and all 200 messages, respectively. In the final 20 messages, the topic-anchored minus free difference changed from 16.5323 to 17.3387 percentage points. The 62-pair bootstrap 95% interval for the boundary variant was [1.2097, 33.0645] percentage points. This is a post hoc, single-scenario sensitivity analysis without multiplicity correction and does not replace the BH inference across 14 outcomes in the complete design. Recalculations for the original and boundary definitions, all three windows, both conditions, and fact-term coverage are saved in notebook_number_boundary_summary.csv and notebook_number_boundary_paired.csv.

S3 Complete Paired Inference for the Topic-Anchored Prompt

All differences are defined as topic-anchored minus free. Differences are first computed within the 310 scenario-and-seed matched pairs; the five differences within each of the 62 replicate blocks are then resampled together. The tables report metric-specific 95% percentile intervals from 20,000 block-bootstrap resamples, two-sided p values from 50,000 whole-block sign flips, and BH-adjusted q values across the 14 exploratory comparisons. The intervals are not adjusted for simultaneous coverage; family-level inference is based on the BH q values. The computational random seed is 20261008.

Table S5. Block-paired inference for proportion metrics. Group means are percentages; differences and intervals are percentage points. Self-repetition retains the denominator of 200.

Metric	Free	Anchored	Difference	95% interval	p	q
Exact self-repetition	72.12	73.25	1.13	[-2.21, 4.49]	0.5152	0.7484
Local near-repetition	75.49	76.71	1.22	[-1.45, 3.89]	0.3839	0.7484
Topic mention, first 20	48.74	51.92	3.18	[0.18, 6.15]	0.0428	0.5997
Topic mention, final 20	39.68	44.45	4.77	[-2.26, 11.81]	0.1949	0.7484
Fact-term coverage, final 20	11.16	11.95	0.79	[-1.30, 2.95]	0.4750	0.7484
Politeness terms, final 20	17.10	17.50	0.40	[-5.08, 5.97]	0.9087	0.9786

Metric	Free	Anchored	Difference	95% interval	p	q
Length-cap events	5.83	4.20	-1.62	[-4.43, 1.21]	0.2683	0.7484
Periodicity in final 40	98.71	97.74	-0.97	[-3.23, 1.29]	0.5880	0.7484
Nearest-scenario hits, final 20	64.44	61.84	-2.60	[-9.42, 4.08]	0.4632	0.7484

Table S6. Block-paired inference for count and semantic metrics. Differences and intervals retain the original units of each metric.

Metric	Free	Anchored	Difference	95% interval	p	q
Distinct normalized replies (count)	45.50	44.67	-0.83	[-4.92, 3.17]	0.6920	0.8074
Bounded near-repetition onset (message)	56.47	53.47	-3.00	[-10.10, 4.22]	0.4157	0.7484
Scenario semantic similarity, first 20	0.5137	0.5199	0.0062	[-0.0013, 0.0138]	0.1119	0.7484
Scenario semantic similarity, final 20	0.4588	0.4543	-0.0044	[-0.0202, 0.0111]	0.5864	0.7484
Semantic margin, final 20	0.0443	0.0444	0.0000	[-0.0145, 0.0148]	0.9953	0.9953

The smallest adjusted q value was 0.5997, for topic-term occurrence in the first 20 messages; none of the 14 outcomes reached 0.05. The earlier analysis used 10,000 block-bootstrap resamples, whereas the final main text and the tables here use the 20,000-resample recalculation. The intervals differ slightly because of Monte Carlo variation: for overall self-repetition, the earlier and recalculated intervals were [−2.19, 4.43] and [−2.21, 4.49] percentage points, respectively. Using the 198 comparable positions as the denominator gives free and topic-anchored means of 72.8527% and 73.9899%, a difference of 1.1372 percentage points, and a 20,000-resample block interval of [−2.2369, 4.5357]; the p and q values are unchanged.

S4 Robustness, Algorithm Comparisons, and Temporal Sensitivity

S4.1 Periodicity Prevalence and Levels of Measurement

Table S7. Strict-periodicity prevalence in different analysis sets. Wilson intervals treat episodes as binomial units; block intervals retain within-replicate dependence.

Analysis set	Periodic / episodes	Rate (%)	Wilson 95% CI (%)	Block 95% CI (%)
All complete dialogues	610/621	98.23	[96.86, 99.01]	[97.10, 99.19]
Complete balanced blocks	609/620	98.23	[96.85, 99.01]	[97.10, 99.19]
Never reached length cap	518/525	98.67	[97.27, 99.35]	[97.68, 99.45]

Block resampling for all 621 episodes uses 63 available replicate clusters, the last of which contains only one episode. The balanced 620-episode analysis uses 62 complete blocks of 10 episodes. The third row is a subset selected for never reaching the per-message 128-token cap during the entire dialogue; it is not a randomized intervention on the length cap.

Table S8. Periodicity rates across five scenarios and two original prompt conditions. Intervals are episode-level Wilson intervals.

Scenario	Original condition	Periodic / episodes	Rate (%)	95% CI (%)
Balcony	Topic-anchored	61/62	98.39	[91.41, 99.71]
Balcony	Free	62/62	100.00	[94.17, 100.00]
Notebook	Topic-anchored	61/62	98.39	[91.41, 99.71]
Notebook	Free	61/62	98.39	[91.41, 99.71]
Park	Topic-anchored	62/62	100.00	[94.17, 100.00]
Park	Free	62/63	98.41	[91.54, 99.72]
Reading room	Topic-anchored	57/62	91.94	[82.47, 96.51]
Reading room	Free	61/62	98.39	[91.41, 99.71]
Town	Topic-anchored	62/62	100.00	[94.17, 100.00]
Town	Free	61/62	98.39	[91.41, 99.71]

Table S9. Sensitivity of raw-string periodicity to the tail-window size and maximum period. Each cell is the number detected among 621 episodes.

Tail window	$p \leq 2$	$p \leq 4$	$p \leq 10$	$p \leq 20$
20	581	607	612	613
40	579	605	610	610
60	575	600	604	604
80	573	595	599	599
100	560	579	583	583

The primary criterion, $p \leq 10$ over a 40-message window, identified 610 periodic episodes using either raw strings or NFKC-normalized text. With raw strings, the numbers of episodes with minimum periods $p = 1$, 2, 4, 6, and 8 were 299, 280, 26, 3, and 2, respectively. Normalization moved only one episode from $p = 2$ to $p = 1$. Applying the same criterion to output token-ID sequences also identified 610 episodes, but the $p = 1$ and $p = 2$ counts were 287 and 292, with all other counts unchanged. Twelve episodes had visible-string period 1 but token period 2, demonstrating that identical text can have different tokenizations. Their episode IDs were 57, 71, 156, 222, 249, 294, 329, 332, 398, 422, 445, and 458.

S4.2 Similarity Thresholds and Stopping Costs

Table S10. Similarity-threshold sensitivity for the 20-consecutive-message rule. Near-repetition requires $s > \theta$; recovery requires any subsequent $s \leq \theta$.

θ	Confirmed episodes	Episodes recovering later	Median confirmation	Avoidable messages (%)
0.80	591	1	60	64.26
0.85	589	0	60	64.00
0.90	587	0	61	63.57
0.95	586	0	62	63.15
0.99	583	3	62	62.44

Table S11. Retrospectively avoidable costs at $\theta = 0.90$. Values are proportions (%) with block-bootstrap 95% CIs.

Run length m	Messages	Recorded generation time	Output tokens
3	71.96 [70.40, 73.49]	89.65 [88.03, 90.95]	76.08 [74.57, 77.58]
5	70.87 [69.30, 72.40]	89.13 [87.46, 90.43]	74.85 [73.35, 76.27]
10	68.31 [66.80, 69.76]	87.97 [86.21, 89.37]	72.07 [70.59, 73.45]
20	63.57 [62.17, 64.99]	85.77 [83.90, 87.29]	66.87 [65.50, 68.19]

Cost accounting retains the triggering message and removes only the suffix after confirmation. The 20-message rule corresponds to 78,953 messages, 3,395,705 output tokens, and 482,021 seconds of recorded generation time. The proportions are calculated from the original trajectories, exclude detection overhead and model-loading time, and are not measured savings from a rerun. Because this rule detects lag-2 repetition, it can miss longer periods. All 25 combinations of thresholds 0.80, 0.85, 0.90, 0.95, and 0.99 with consecutive-run lengths 3, 5, 10, 20, and 30 are provided in stopping_threshold_sensitivity.csv.

S4.3 Topic Decoding and Early Prediction

Table S12. Five-scenario decoding from the final two messages. Accuracies and intervals are percentages. The original 621-episode OOF intervals use block resampling of fixed predictions; the five repeated five-fold partitions use the balanced 620-episode sample.

Method	OOF accuracy	95% CI	Mean across five partitions	Range across five partitions
TF-IDF + LR	64.57	[61.03, 68.10]	63.65	62.74–64.35
TF-IDF + linear SVM	64.09	[60.32, 67.79]	63.65	61.77–65.00
BGE + LR	71.82	[68.46, 74.92]	72.68	71.61–73.87

Each fold keeps entire replicates separate. TF-IDF vocabularies and classifiers are fitted only on training folds; BGE is a frozen external encoder. The same five opening-scenario templates occur across all folds. The results therefore measure discrimination among these five categories and should not be interpreted as generalization to unseen topics. The intervals resample fixed out-of-fold (OOF) predictions and do not propagate model-refitting uncertainty. Multiple fold partitions are not independent datasets.

Table S13. Prediction of near-repetition confirmation by message 80 using the first 20 messages in the original 621 episodes. Balanced accuracy uses a threshold of 0.5.

Algorithm	AUC	AUC 95% CI	Balanced accuracy	Brier score
Gradient boosting	0.688	[0.640, 0.736]	0.588	0.178
Logistic regression	0.720	[0.676, 0.765]	0.546	0.169
Random forest	0.697	[0.648, 0.743]	0.555	0.172

Table S14. Simple baselines evaluated on the same balanced 620-episode sample. The positive class is 461/620 = 74.35%.

Method	AUC	AUC 95% CI	Balanced accuracy	Brier score
22-feature LR	0.720	[0.675, 0.765]	0.546	0.169
Lag-2-similarity-only LR	0.698	[0.655, 0.739]	0.500	0.176
Training-fold class prior	0.480	[0.434, 0.527]	0.500	0.191

The AUC difference between the full logistic regression (LR) model and the lag-2-similarity-only LR model was 0.0225, with a paired block-bootstrap 95% interval of [−0.0100, 0.0542] based on fixed OOF predictions. These results do not establish that the larger feature set outperforms the simple baseline. At a threshold of 0.5, the full LR model produced TP = 445, FN = 16, TN = 20, and FP = 139, corresponding to 96.53% sensitivity and 12.58% specificity; the single-feature LR model classified every episode as positive. LR using only topic/condition identities had an AUC of 0.5619. The training-fold class-prior baseline can have an OOF AUC different from 0.5 because the class prior varies slightly across folds.

For label permutation, the 62 complete outcome blocks of 10 cells were reassigned across replicates while preserving topic-condition positions, within-block pairing, and grouped validation. Standardization and five-fold LR fitting were repeated for each permutation. Across 200 permutations, the mean AUC was 0.5156, the central 95% range was [0.4518, 0.5758], and the maximum was 0.5889. The observed AUC was 0.7202, yielding permutation $p = (1 + 0)/(200 + 1) = 0.004975$. This test evaluates episode-level association beyond fixed design-cell identity.

Table S15. Sensitivity of early classification to grouped partitions and time ordering. The temporal split uses 450 training episodes and 170 later test episodes.

Method	Mean AUC, 20 partitions	AUC range, 20 partitions	Later-test AUC	Later-test balanced accuracy
22-feature LR	0.721	0.707–0.733	0.640	0.537
Lag-2-similarity-only LR	0.694	0.687–0.699	0.674	0.500
Scenario/condition-only LR	0.548	0.519–0.570	0.478	0.500

The 20 partitions use GroupKFold randomization seeds 0 to 19 and keep replicates separate throughout. The temporal sensitivity analysis fixes training blocks at 0 to 44 and test blocks at 45 to 61. It is a retrospective later-period test within the same known generation batch. Its denominator differs from the confirmation-time regression analysis in the main text, which used the original 621 episodes and a 171-episode later test set.

Table S16. Distinguishing prediction of confirmation from prediction of repetition onset. All models are refitted with grouped five-fold validation on the selected samples.

Sample rule	Episodes	22-feature AUC	Single-feature AUC	Screenable at message 20
Complete balanced sample	620	0.720	0.698	Yes
No three consecutive near-repetitions by message 20	592	0.713	0.679	Yes
Post hoc exclusion of sustained-run onset ≤ 20	574	0.695	0.667	No

Among all 621 episodes, sustained near-repetition had already begun by message 20 in 46 episodes; the earliest onset was message 8 and the earliest confirmation was message 27. Features from the first 20 messages may therefore contain emerging repetition signals. The second row excludes episodes using only three consecutive near-repetitions already visible at message 20. The third row requires subsequent trajectories to establish the onset and is a post hoc sensitivity analysis, not an online screening rule.

S4.4 Sensitivity to Collection Order

The first 31 complete blocks and the final 31 complete blocks each contain 310 episodes. Their strict-periodicity rates were 98.0645% and 98.3871%, respectively: later minus earlier = 0.3226 percentage points, with a block-bootstrap 95% interval of [−1.6129, 2.5806]. Differences in bounded near-repetition onset, overall self-repetition, and final-20-message semantic similarity also did not reach 0.05 after BH correction across the four comparisons. Cumulative periodicity rates for the first 100, 200, 300, 400, 500, 600, and 620 episodes were 97.0%, 97.5%, 98.0%, 98.0%, 98.4%, 98.17%, and 98.23%, respectively. These results describe stability within the current batch and do not correct for the absence of a predetermined collection-stopping rule.

S5 Branch-Intervention Protocol for Frozen Prefixes

S5.1 Prefix Selection and Sampling Seeds

The protocol locked_prefix_branch_intervention_v1 was written and assigned a saved SHA-256 hash on 8 October 2026 before new generation began. Within each original scenario × prompt-condition cell, complete episodes were first screened for raw-text equality between messages 61 to 80 and the same speakers' previous replies, messages 59 to 78. The eligible episode with the smallest episode_id was selected. Selection used only the original first 80 messages and completion status, without access to new branch outputs. Ten prefixes cover all original design cells. Each prefix is continued under two new seeds, each seed is crossed with five conditions, and each branch generates 12 messages, six per agent.

Table S17. Frozen prefixes, original seeds, and two new continuation base seeds. The final column gives the number of eligible episodes in the original design cell.

Prefix ID	Original episode ID	Scenario / original condition	Original episode seed	New base seeds b_0 / b_1	Eligible episodes
0	16	Balcony / free	20261026	2026100800 / 2026100801	40
1	9	Balcony / anchored	20260926	2026100810 / 2026100811	46
2	2	Reading room / free	20260927	2026100820 / 2026100821	43
3	0	Reading room / anchored	20260927	2026100830 / 2026100831	38

Prefix ID	Original episode ID	Scenario / original condition	Original episode seed	New base seeds b_0 / b_1	Eligible episodes
4	10	Park / free	20261028	2026100840 / 2026100841	50
5	3	Park / anchored	20260928	2026100850 / 2026100851	54
6	4	Town / free	20260929	2026100860 / 2026100861	41
7	1	Town / anchored	20260929	2026100870 / 2026100871	48
8	52	Notebook / free	20261430	2026100880 / 2026100881	41
9	5	Notebook / anchored	20260930	2026100890 / 2026100891	37

$$b(p,r) = 2026100800 + 10p + r;\ \text{sampling_seed}(p,r,k) = 1000 \times b(p,r) + k,\ r \in \{0,1\},\ k = 1,\ldots,12.$$

The five conditions within the same prefix and continuation seed share step-level random seeds to form paired comparisons. Condition order is shuffled within each prefix × continuation-seed block using Python random.Random(20261008) and saved. The two continuation seeds are repeated measurements on the same history, not two independent histories. The ten prefixes correspond to eight original episode seeds: the free and topic-anchored reading-room prefixes share one original seed, as do the two town prefixes.

S5.2 Five Treatment Conditions and the Complete Appended Instruction

Table S18. Branch conditions. The original model and generation settings are retained except for the listed changes.

Condition	History treatment	Sampling or instruction change
baseline	Full original 80-message prefix plus growing continuation	temperature=0.8, top_p=0.95, top_k=20, repetition_penalty=1.0
penalty115	Full history	Only repetition_penalty changes, to 1.15
short8	Original system prompt + the agent's initial user message + four most recent assistant/user pairs	Rolling last-eight-message window; no textual summary; original output tokens retained
instruction	Full history	Fixed anti-repetition instruction appended to both system prompts
official_sampling	Full history	temperature=1.0, top_p=1.0, top_k=20, presence_penalty=2.0, repetition_penalty=1.0

Complete anti-repetition text appended to the system prompt, preceded by one newline (English translation): In the following conversation, do not repeat or merely rephrase anything you or the other person has already said. Each time, add a specific new detail within the opening scenario or ask a question that has not yet been discussed, so that the conversation continues to move forward. The exact Chinese instruction is preserved in paper/supplement_experiments/protocol.json.

For each generation, short8 selects the agent's four most recent complete assistant/user pairs and separately retains its initial user message. B's initial user message still contains the opening scenario and A's first utterance. The required cache is rebuilt when the history window changes. This is an intentional history intervention, not an unplanned truncation in the original run.

repetition_penalty follows the Transformers token-repetition penalty and acts on the input and already-generated sequence for that generation call. The presence penalty is implemented by a custom

LogitsProcessor: each distinct token already generated within the current reply has 2.0 subtracted from its next-step logit. The subtraction is not accumulated by frequency, does not penalize input/history tokens, and is applied before temperature and top-k/top-p sampling transformations. Pre-run checks verified that historical tokens are excluded, repeated occurrences of a token incur only one subtraction, and no penalty is applied before any output has been generated.

The recommended-sampling arm follows the non-thinking text-generation recommendations in the public model card at the fixed model revision, verified on 8 October 2026. The per-message 128-token cap, brief-reply instruction, and history protocol retain the settings of this study. Because this arm changes temperature, top_p, and presence penalty together, between-arm differences cannot be attributed to any one of these parameters.

Both agents' caches are cleared before every branch, and their role-specific histories are independently rebuilt from the same original 80-message prefix. System-prompt modifications are applied to both agents. New outputs are not passed between conditions or continuation seeds. The input-token hash for the first baseline continuation is compared with the original input hash at message 81; the new sampling seed does not require reproduction of the original message-81 output. Model weights remain fixed, as do BF16 precision, SDPA, disabled thinking, the per-message maximum of 128 new tokens, and the context limit of 32,768 tokens.

S5.3 Primary and Secondary Endpoints and Inference

$$D = (1/12)\ \Sigma_{k=1}^{12}\ 1\{s_{80+k} \le 0.90\}.$$

The primary endpoint D is the proportion of all 12 continuation messages whose same-speaker lag-2 character-bigram cosine similarity is no greater than 0.90. Higher values indicate more short-term textual change. Each of the first two continuation messages is compared with that speaker's most recent reply in the original prefix. The change proportion in the final six messages and whether all six satisfy the change criterion are secondary endpoints and do not replace the primary endpoint.

Other secondary endpoints include raw and normalized exact lag-2 repetition; the proportion of messages outside the set of original final-20-message raw texts; nearest cosine distance to the original cycle set; whether the final six messages form a raw-text cycle with p = 1 to 4; whether such a cycle is new; new-bigram proportion relative to the full original prefix and preceding continuation; original topic-term coverage; lexical similarity to the opening scenario; character count; length-cap events; empty replies; and recorded elapsed time. Even when short8 removes older text from the model input, novelty is evaluated against the full original prefix, preventing forgotten earlier content from being counted as new information.

Within each condition, the two continuation seeds are averaged within each prefix, followed by an equal-weight mean across the ten prefixes. Differences from baseline retain prefix pairing; prefixes sharing an original seed are resampled together. Inference uses 10,000 bootstrap resamples of the eight original-seed clusters and exact sign-flip enumeration of all $2^8 = 256$ cluster-level sign combinations. The four primary-endpoint contrasts receive joint BH correction; secondary endpoints are exploratory descriptions. Two continuation seeds do not increase the number of independent prefixes, and eight clusters limit test resolution. The experiment estimates the 12-message short-term response of the selected, already locked prefixes, not prevention from the beginning of a dialogue or long-term factual quality.

S5.4 Branch Results and Independent Recalculation

All 100 branches were completed. Each treatment condition contains 10 prefixes × 2 new seeds, yielding 20 branches and 240 new messages. An independent implementation recalculated character bigrams, the primary endpoint, final-window change, departure from the original cycle, and short periodicity directly from the raw JSONL files. Every branch-level result agreed with the original analysis, as did the exact cluster-level sign-flip p values and BH q values for all four primary contrasts.

Table S19. Observed results for the frozen branches. D, original-cycle departure, topic mention, and length-cap columns are percentages; the two final-six columns are branch counts out of 20.

Condition	D	All final six changed	Final-six cycle	Original-cycle departure	Topic mention	Length cap
Baseline	0.00	0/20	20/20	0.00	60.00	5.00
Repetition penalty 1.15	0.00	0/20	19/20	0.00	60.00	5.00
Recent eight	0.00	0/20	20/20	0.00	60.00	5.00
Anti-repetition instruction	0.00	0/20	20/20	0.00	60.00	5.00
Recommended sampling	7.50	1/20	15/20	7.50	62.50	2.92

"All final six changed" requires $s \leq 0.90$ for each of the last six messages. "Final-six cycle" uses the smallest $p \in \{1,2,3,4\}$ for which $y_t = y_{t-p}$ holds at every position t = 87 to 92; reference positions may precede these six messages. "Departure from the original cycle" requires the current text to have similarity no greater than 0.90 to every normalized text in the original final 20 messages. A branch can leave the original cycle and form another cycle, so these three outcomes are not interchangeable.

Table S20. Paired primary-endpoint differences from baseline. Differences and intervals are percentage points. Intervals use the protocol-specified 10,000 bootstrap resamples of eight clusters.

Treatment	Difference in D	95% interval	Exact p	BH q
Repetition penalty 1.15	0.00	[0.00, 0.00]	1.0000	1.0000
Recent eight	0.00	[0.00, 0.00]	1.0000	1.0000
Anti-repetition instruction	0.00	[0.00, 0.00]	1.0000	1.0000
Recommended sampling	7.50	[1.14, 17.08]	0.2500	1.0000

The mean D for the recommended-sampling combination was 7.50%, compared with 0 for the other four arms. Four of its 20 branches had at least one message satisfying $s \leq 0.90$, and one branch satisfied this condition for all final six messages. The exact sign-flip comparison with baseline gave p = 0.25 and q = 1.00 after BH correction across four contrasts. The descriptive change therefore does not establish an intervention effect in this small sample. An eight-cluster percentile-bootstrap interval is not an inversion of the exact test; when nonzero changes are concentrated in a few clusters, the two procedures can yield different boundaries. The other three contrasts and their resampling intervals were all 0, reflecting a degenerate observed distribution in this sample rather than proving a population effect of exactly zero or treatment equivalence.

Recalculating branch topic mentions with the numeric-boundary rule from Section S2.3 reduced the baseline, repetition-penalty, recent-eight, and anti-repetition-instruction arms from 60.0% to 50.0%, and the recommended-sampling arm from 62.5% to 52.5%. The between-arm difference remained 2.5 percentage points. This post hoc description reflects the ability of longer numbers such as "1937" to match the original substring "37"; occurrence of a topic term does not automatically imply retention of the opening scenario's specific facts.

Sampling seeds, speaker order, output-token counts, context limits, and ending records were verified for every continuation message. The first input hashes of all 20 baseline branches matched their corresponding original message-81 inputs. For the repetition-penalty and recommended-sampling arms, first input hashes also matched baseline for the same prefix/seed. Input hashes changed as designed for the instruction and short-history arms. The first generation by each agent in every branch started from an empty cache.

Independent checks confirmed earliest-eligible selection of all ten prefixes, grouping into eight original-seed clusters, all 100 scheduled entries, and the original-data hashes.

The 1,200 continuation messages generated 61,450 output tokens, with a total recorded per-message generation time of 2,625.65 seconds and 0 empty replies. An additional 50,000-resample eight-cluster bootstrap is saved in independent_branch_primary_contrasts.csv as a numerical sensitivity analysis. The contrast table here retains the 10,000-resample intervals specified in the frozen protocol.

After the main generation run, the first reply was replayed once under each of the five conditions for the same balcony prefix and continuation seed, recording the actual generation configuration and installed logits-processor order. All five configurations matched the protocol, and the input hashes and output tokens reproduced the corresponding original branches exactly. Independent small-tensor checks verified the presence-penalty subtraction for positive and negative logits, exclusion of history tokens, and counting each distinct token only once. These five replays are execution checks and are excluded from the 100-branch inferential sample.

S6 Data, Code, and Reproduction Workflow

Reproduction files retain their project-relative directory structure. Raw JSONL records store message text, episode and speaker identifiers, input hashes, generated token IDs, per-message sampling seeds, lengths, finish reasons, and elapsed times. Configuration, schedule, and environment records are stored separately; analysis outputs can be recalculated from the raw records. Model weights are obtained at the stated revisions. Verification hashes for the original weights and generator are provided below.

Table S21. SHA-256 hashes of key files. The Qwen weight file contains 1,746,942,600 bytes.

File	SHA-256
Qwen weights	04b1c301231dd422b8860db31311ab2721511346a32cb1e079c4c4e5f1fe4696
Original generator	d9931d5a2f0bc5598d5812a99219cfb7419d479b391c7c3e568a8fe94e5b5977
BGE weights	354763b9b1357bc9c44f62c6be2276321081ed2567773608c0d0785b61d5a026

Per-file hashes for the raw data are retained in the manifest and data-identity audit files; code hashes and runtime versions are stored separately.

Table S22. Principal reproduction files and their roles. Paths are relative to the project root.

Relative file or directory	Contents and purpose
runs/qwen_duet_v1/config.json, schedule.json, runtime.json	Observational generation parameters, episode ordering, hardware, and versions
runs/qwen_duet_v1/episodes/	Message-level JSONL for 621 complete episodes and one incomplete episode
run_dialogues.py	Original two-context dialogue generator
deep_analysis/prepare.py	Read-only verification and message-/episode-level lexical metrics
deep_analysis/embed.py, compare_methods.py, dynamics.py	Frozen semantic encoding, paired statistics, algorithm comparisons, and sequence analysis
deep_analysis/results/manifest.json	Raw-file SHA-256 hashes and data-structure verification
paper/audit/paper_tables.json, table_inventory.csv	Shared numerical sources for the main text and supplement, CSV inventory, and hashes

Relative file or directory	Contents and purpose
paper/audit/audit_statistics.py	Independent periodicity calculations, intervals, block inference, and permutation baselines
paper/audit/audit_predictive_sensitivity.py	Simple baselines, repeated partitions, temporal validation, and stopping-cost intervals
paper/audit/early_signal_sensitivity.py	Sensitivity to early repetition signals and the confirmation target
paper/audit/notebook_number_boundary_sensitivity.py	Post hoc numeric-boundary sensitivity for the term 37
paper/audit/audit_completed_branches.py	Independent metrics, paired inference, and data-identity verification for the 100 new branches
paper/supplement_experiments/protocol.json, schedule.json	Protocol and schedule frozen before branch generation
paper/supplement_experiments/raw/, source_prefixes/	New branch logs and their original 80-message source prefixes
paper/supplement_experiments/run_branches.py, analyze_branches.py	Branch generation and read-only result recalculation

Statistical recalculation used Python 3.11.15, NumPy 2.4.4, pandas 2.3.3, SciPy 1.17.1, scikit-learn 1.9.0, and statsmodels 0.14.6. The principal original-analysis random seed was 20261003; the sequence and stopping analyses used 73472; the independent supplementary audit used 20261008; and the subsequent predictive sensitivity analysis used 71008. Actual versions and script SHA-256 hashes are saved in analysis_code_identity.json.

To recalculate the existing observational data, run prepare.py, embed.py, compare_methods.py, dynamics.py, and validate_results.py in sequence; the existing frozen embedding cache may be reused. For the independent CPU audit, run audit_statistics.py, audit_predictive_sensitivity.py, early_signal_sensitivity.py, notebook_number_boundary_sensitivity.py, and finish_audit.py in sequence. Branch analysis reads completed logs with analyze_branches.py, followed by independent verification with audit_completed_branches.py; neither step calls the model. To regenerate branches, run_branches.py reads the frozen protocol, verifies source data, and skips fully completed branches. Incomplete branches are not silently overwritten.

S7 Scope of Statistical and Execution Verification

The independent audit checked the hashes of all 622 original log files and rechecked episode/message primary keys, the 200-message completion criterion, strict A/B alternation, schedule/scenario consistency, every sampling-seed formula, output-token counts, input-length limits, and context_reset. The 621 complete episodes contain 124,200 messages; the incomplete episode's 161 messages are excluded from primary inference. Weight-file SHA-256 hashes matched the download-verification record, and the original generator hash matched the original runtime record. The stop flag and state of the original observational run were left unchanged.

The original observations and offline algorithm comparisons are exploratory. Proportion intervals quantify episode-level uncertainty conditional on these five fixed scenarios, this model, and this generation protocol; messages are not independent samples. Stopping collection at 621 complete dialogues was not governed by a predetermined power rule. Earlier/later comparisons and checks on the balanced 620-episode subset do not remove this selection context. Prefix selection, seeds, treatments, and endpoints for the new branches were frozen before new outputs were generated, but their historical prefixes came from already observed original data.

The cache audit selected two positions each from dialogues with periods 1, 2, and 4 and from dialogues without a primary-periodicity event. Each empty cache was rebuilt from the full token history. All eight input hashes and output-token sequences matched the original records. This supports reproducibility of the cache path at the sampled positions; it does not establish that full re-rendering of the official chat template would yield identical results, nor does it exclude an effect of the original append-only protocol.

Grouped validation for supervised models separates replicates, while the fixed scenario categories remain shared across folds. Early prediction uses features visible in the first 20 messages and targets subsequent confirmation, allowing some repetition signals to have emerged beforehand. Sensitivity analyses that select samples using post hoc labels are not online policies. Bootstrap intervals based on fixed OOF predictions and repeated fold partitions describe uncertainty and partition dependence; they do not constitute independent external validation.

Stopping-cost estimates are counterfactual accounting of already generated suffixes. Resolution of near-repetition, departure from the original text set, sustained textual change, entry into a new cycle, and factual accuracy are distinct outcomes. Tail-topic decodability, lexical coverage, and dense-vector similarity do not automatically imply retention of factual relations. Quantitative conclusions are interpreted according to these operational definitions; observed finite string cycles are not equated directly with hidden-state attractors or irreversibility over an infinite future.