PREFIX SEMANTIC DYNAMICS IN LONG-FORM LANGUAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-form responses from large language models (LLMs) can remain locally coherent while becoming semantically repetitive or increasingly detached from the original prompt as generation proceeds. We introduce the Prefix Semantic Dynamics (PSD) method to characterize and monitor this process during text generation. Specifically, PSD constructs a semantic trajectory from the normalized embeddings of successive cumulative text prefixes. This trajectory is further characterized by three geometric descriptors: correlation dimension D_2, semantic recurrence ρ, and tortuosity τ, which describe its expansion, recurrence, and path geometry, respectively. We evaluate PSD on 720 text trajectories spanning natural, generated, and controlled text, where its geometric descriptors capture systematic differences in both semantic repetition and expansion. We then introduce PSD Control, which uses the evolving prefix trajectory as feedback to guide generation. Across 80 LongFact prompts, PSD Control consistently shifts text generation toward trajectories that are less recurrent, more expansive, and better aligned with the prompt than standard continuation. These results demonstrate that PSD provides a new representation that can both characterize semantic development and guide subsequent generation.

1 INTRODUCTION

In long-form generation, we usually expect large language models (LLMs) to write long responses that remain relevant to the prompt as they unfold. However, a response can remain coherent from one passage to the next while gradually becoming repetitive (Holtzman et al., 2020; Welleck et al., 2020) or drifting beyond our intended scope (Spataru et al., 2024; Xu et al., 2023). We therefore need to know when these changes emerge and how they develop through the rest of the response.

Existing metrics measure lexical agreement (Papineni et al., 2002; Lin, 2004), learned semantic similarity (Zhang et al., 2020; Sellam et al., 2020), factual support (Min et al., 2023; Wei et al., 2024), or uncertainty and consistency across responses (Farquhar et al., 2024; Manakul et al., 2023). These metrics usually provide a static assessment, often after generation is complete. Tracking how meaning develops dynamically can reveal different patterns and help adjust subsequent generation if needed.

We therefore introduce Prefix Semantic Dynamics (PSD) to measure semantic development through a growing response. At each checkpoint, PSD embeds the cumulative prefix so far into a normalized vector as the prefix state. Repeating this process gives us a semantic trajectory that shows how the response develops (Figure 1). Drawing on correlation and recurrence analysis (Grassberger & Procaccia, 1983; Eckmann et al., 1987; Marwan et al., 2007; Kantz & Schreiber, 2003), we describe this trajectory using correlation dimension D_2, semantic recurrence ρ, and tortuosity τ. These quantities respectively measure finite-scale expansion, the concentration of nearby prefix states, and accumulated local change relative to endpoint change.

We first evaluate PSD on 720 text trajectories to examine how its descriptors reflect text content and organization. Experiments show that different forms of degeneration leave distinct patterns in the trajectory. Repetitive content pulls later prefixes back toward similar semantic states, while bland text adds relatively little new semantic structure as the response grows. These differences become clearer over longer prefixes, as natural text continues to explore new semantic regions while degenerate text remains increasingly concentrated around states it has already visited. The trajectory therefore

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

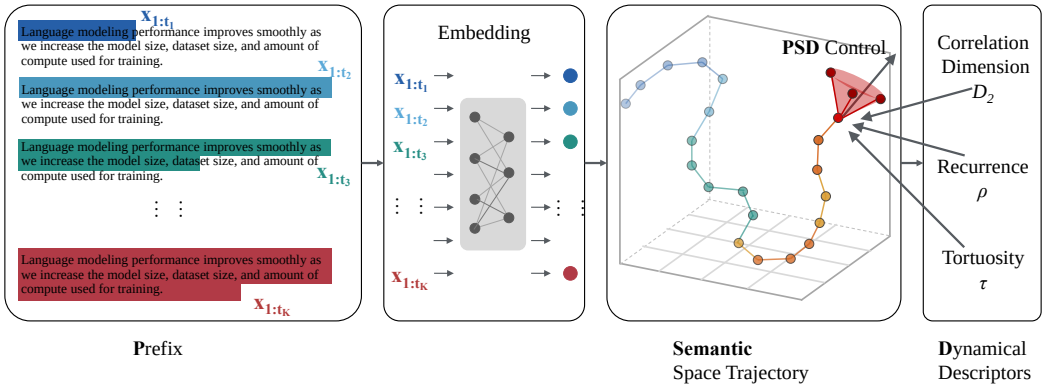

Figure 1: **Prefix Semantic Dynamics.** Each state represents all text from the beginning of a response through one checkpoint. Normalized prefix embeddings form a trajectory summarized by D_2, ρ, and τ. PSD Control evaluates candidate extensions in this state space and selects a segment to append to the response.

captures how semantic content unfolds across the response, including both where it moves and how often it returns to earlier regions.

We further introduce PSD Control to use the semantic trajectory as feedback during generation. At each round, four actuators propose candidate segments that continue the discussion, introduce a prompt-relevant direction, add detail, or return attention to the prompt. The controller evaluates how each candidate would extend the current trajectory, selects one segment, and repeats the process, so that each generation step is informed by the semantic content accumulated throughout the response. On 80 previously unused LongFact prompts with two generator scales and two open representation models, PSD Control reduces recurrence, increases D_2, and improves prompt alignment compared with standard continuation. We further use the same trajectory state for direct Pareto selection among candidate segments, which produces a larger reduction in recurrence and greater semantic expansion than the phase policy while maintaining similar alignment. These trends remain when trajectories are evaluated with a different representation model, and the resulting responses also show greater relevance and content-word coverage.

2 Related Work

Semantic representations and uncertainty. Universal Sentence Encoder, Sentence-BERT, and SimCSE represent text as vectors that can be compared by meaning (Cer et al., 2018; Reimers & Gurevych, 2019; Gao et al., 2021). BERTScore and BLEURT use learned representations to evaluate generated text (Zhang et al., 2020; Sellam et al., 2020). Semantic entropy and related graph or kernel methods measure uncertainty through semantic relationships among sampled responses (Farquhar et al., 2024; Nikitin et al., 2024; Qiu & Miikkulainen, 2024; Jiang et al., 2024). Our PSD instead follows the ordered semantic states of cumulative prefixes within a single response.

Correlation structure in language. Correlation dimension describes how the fraction of nearby state pairs changes with neighborhood radius (Grassberger & Procaccia, 1983). Recurrence plots locate nearby states along a trajectory and have long been used to study temporal organization in dynamical systems (Eckmann et al., 1987; Webber & Zbilut, 1994; Marwan et al., 2007). More recently, correlation dimension has been estimated from LLM next-token log-probability vectors (Du & Tanaka-Ishii, 2025). Our PSD applies these geometric measurements to the semantic trajectory formed by cumulative prefixes.

Generation control. Controllable generation includes gradient steering (Dathathri et al., 2020), future discriminators (Yang & Klein, 2021), expert and anti-expert decoding (Liu et al., 2021), and history-aware decoding (Xu et al., 2023). Tree of Thoughts, Self-Refine, and Reflexion use multiple model calls to explore, evaluate, or revise outputs (Yao et al., 2023; Madaan et al., 2023; Shinn

et al., 2023). Our PSD Control uses the cumulative-prefix semantic trajectory itself as feedback for selecting the next segment.

3 Prefix Semantic Dynamics

3.1 Cumulative-prefix states

Consider a text of length T and checkpoints $0 < t_1 < \cdots < t_K = T$. At checkpoint k, the cumulative prefix $x_{1:t_k}$ contains every token from the beginning through t_k. A representation model E maps this variable-length prefix to a fixed-dimensional vector, which we normalize:

$$z_k = \frac{E(x_{1:t_k})}{\|E(x_{1:t_k})\|_2}. \tag{1}$$

The ordered sequence $Z = (z_1, \ldots, z_K)$ is the semantic trajectory. Successive states track how new content changes the response accumulated so far. Algorithm 1 combines measurement and control; Appendix Algorithm 2 gives standalone PSD measurement.

3.2 Finite-scale trajectory descriptors

We use cosine distance $d_{\cos}(z_i, z_j) = 1 - \langle z_i, z_j \rangle$. Correlation dimension and semantic recurrence are obtained from the empirical correlation integral:

$$C(\epsilon) = \frac{2}{K(K-1)} \sum_{1 \le i < j \le K} \mathbf{1}[d_{\cos}(z_i, z_j) \le \epsilon]. \tag{2}$$

This quantity is the fraction of prefix-state pairs within a semantic neighborhood of radius ϵ. Its dependence on radius describes the distribution of pairwise semantic distances across scales.

Correlation dimension. Let $C_+(\epsilon)$ be the fraction of strictly positive pairwise distances that do not exceed ϵ. Zero-distance pairs enter neither count. Its log-log slope measures finite-scale expansion:

$$D_2 = \frac{\mathrm{d} \log C_+(\epsilon)}{\mathrm{d} \log \epsilon}. \tag{3}$$

We sample $C_+(\epsilon)$ on a fixed quantile grid and fit the main diagnostic atlas over $0.10 \le C_+ \le 0.50$. Appendices B.3 and C.3 give alternative diagnostic windows and controller fitting settings, respectively.

Semantic recurrence. At a fixed radius, the correlation integral gives the proportion of nearby prefix-state pairs:

$$\rho = C(\epsilon^*). \tag{4}$$

A larger ρ indicates that more pairs occupy the same local semantic neighborhood. We use $\epsilon^* = 0.1$ for the main diagnostic backend. The radius is fixed within each backend and calibrated when changing representations.

Trajectory tortuosity. We compare accumulated local change with the change between the first and last states using

$$\tau = \frac{\sum_{k=1}^{K-1} (1 - \langle z_k, z_{k+1} \rangle)}{1 - \langle z_1, z_K \rangle}. \tag{5}$$

This cosine-distance form of tortuosity sums changes between successive states and expresses them relative to the endpoint change. It complements D_2 and ρ by retaining the order of movement along the path. Its interpretation depends on the form of semantic development under study.

Unit normalization gives $\|z_i - z_j\|_2^2 = 2d_{\cos}(z_i, z_j)$, so radius remapping preserves neighborhood membership. All reported slopes use cosine distance; Appendix B.3 gives the corresponding scaling relations.

3.3 Control from the prefix trajectory

At generation round r, PSD Control computes $s_r = (D_2, \rho, \tau, A_q)$ from the committed prefix, where

$$A_q = \cos(E(q), z_K) \tag{6}$$

measures prompt alignment. Algorithm 1 evaluates four actuators $\mathcal{A}$: Continue, Branch, Densify, and ReAnchor. The selected segment extends the committed prefix.

The *phase-based* policy diagnoses a phase from state history and selects a Pareto candidate accordingly. *PhaseAgnosticPareto*, labeled PSD (phase-agnostic) in figures, uses an equal-weight rank sum within the same Pareto set, omitting phase labels, running medians, and descriptor trends. Both share actuators, candidate count, and Continue-only warmup. Appendix C.1 gives phase rules; Appendix C.3 gives objectives, tie-breaking, and the policy-specific preference for higher τ.

Algorithm 1 Prefix Semantic Dynamics and Generation Control

Require: Prompt q, initial prefix x, generator G, representation E, rounds R, policy p
Initialize state history $H \leftarrow [\,]$.
for $r = 1, \ldots, R$ **do**
  Embed and normalize each cumulative prefix $x_{1:t_k}$ to form $Z = (z_1, \ldots, z_K)$ (Eq. 1).
  Compute C, C_+, and (D_2, ρ, τ) using Eqs. 2–5.
  Set $s_r = (D_2, \rho, \tau, A_q)$ using Eq. 6; append s_r to H.
  $\phi_r \leftarrow \pi_{\text{phase}}(H)$ for the phase-based policy; otherwise $\phi_r \leftarrow \varnothing$.
  $\mathcal{A}_r \leftarrow \{\text{Continue}\}$ during warmup; otherwise $\mathcal{A}_r \leftarrow \mathcal{A}$.
  for each $a \in \mathcal{A}_r$ (in parallel) **do**
    Generate $c_a \leftarrow G(q, x, a, \phi_r)$; compute $s_{r,a}$ from the extended prefix $x \| c_a$.
  end for
  Select $a_r \leftarrow \pi_p(\{(a, s_{r,a})\}_{a \in \mathcal{A}_r}, \phi_r)$.
  Append the selected segment: $x \leftarrow x \| c_{a_r}$.
end for
return Completed text x and recorded PSD trajectory.

4 Experimental Protocol

4.1 Diagnostic benchmark

The diagnostic benchmark contains 720 text trajectories (Table 1). Human passages from SEP, Gutenberg, Wikipedia, and academic sources provide natural-language references, while Python code supplies structured non-prose text. Source-matched LLM generations allow comparisons within broad domains. Additional generations use normal, bland, incoherent, and repetitive prompts.

We use controlled perturbations to examine how specific changes to a text appear in its trajectory. Shuffling changes discourse order, repeated blocks introduce recurring content, and bland passages limit semantic development. Random-name streams provide weak discourse structure, while synthetic trees introduce an explicit branching structure. We construct checkpoints by recursive boundary splitting with a target step of 100 tokens and use `text-embedding-3-large` as the main diagnostic backend. We further vary the prefix horizon, token step, D_2 fitting window, embedding dimension, and representation backend to examine how these choices affect the measurements.

4.2 Generation control and output evaluation

The initial controller study uses 20 LongFact-style prompts (Wei et al., 2024), ten generation rounds, and five policies: Continue, AlwaysBranch, AlwaysDensify, AlwaysReAnchor, and PSD Control. Its

Table 1: Diagnostic benchmark. Controller prompts are counted separately.

Text group	Included classes	Texts
Human prose	SEP, Gutenberg, Wikipedia, academic	200
Structured text	Python code	50
Manufactured controls	Shuffled, repetitive, bland, random names, synthetic tree	400
Source-matched LLM	Matched SEP, Gutenberg, Wikipedia	30
Prompt-regime LLM	Normal, bland, incoherent, repetitive	40

100 final policy–prompt states and 740 candidate continuations characterize the changes induced by each actuator and the selections made by the phase-based policy.

The expanded study uses the same 80 previously unused LongFact prompts in four configurations. It combines Qwen3.5-9B and Qwen3.5-122B-A10B-FP8 generators with two online representations: the dedicated embedding model Qwen3-Embedding-8B and the normalized LAST-pooled hidden state of Harrier-OSS-v1-27B. Each run starts from a shared 360-token seed and adds up to 300 tokens per round for ten rounds, with checkpoints at a 100-token target step. The phase rules and actuator prompts are fixed from the initial study.

The policies include Continue, PSD Control, PhaseAgnosticPareto, TextOnlyRouter, Continue with repetition penalty 1.1, and three complete shuffled-phase replicas. TextOnlyRouter receives the prompt, the latest 900 words of the prefix, and its four candidate texts. It uses the same actuator set and candidate count as PSD Control, with one additional deterministic routing call and no PSD values. Together, these policies compare phase-based selection, direct geometric selection, routing from recent text, and token-level repetition suppression.

Prompts are the unit of statistical inference. We average each descriptor over the ten rounds within a prompt, then estimate policy differences and 95% intervals using 10,000 paired prompt-level bootstrap resamples. All four configurations use the same 80 prompts. Final-state results are secondary. Final-output measures include prompt relevance under both representations, prompt content-word coverage after stopword removal and Porter stemming, the fraction of exactly repeated sentences, and response length. A separate method-blinded, retrieval-backed audit of OpenAI-generated outputs evaluates factual support.

5 Results and Discussion

5.1 Semantic repetition and discourse organization have distinct geometric signatures

Figure 2(a) compares human prose with source-matched LLM generations. Each curve gives the fraction of nearby prefix-state pairs as the semantic neighborhood grows. Human and generated text follow different mean curves within the same broad domains, showing differences in how their prefix states are distributed across semantic scales.

Figure 2(b) compares code and normal LLM generations with controlled text regimes. Repetitive, bland, and tree controls accumulate a larger fraction of nearby state pairs at small radii, indicating concentrated trajectories. Code and normal LLM generations show steeper growth over their scaling ranges, reflecting greater semantic expansion. This scale-dependent growth is captured by D_2.

Recurrence maps locate these similarities in time (Figure 2c–f). Human SEP and code examples have localized near-diagonal structure; normal LLM generation shows dispersed off-diagonal recurrence, and repetitive text forms a broad region of similar states. The scalar ρ measures their overall frequency, while the maps locate recurrence along the response.

5.2 Descriptor profiles distinguish forms of semantic development

The descriptor profiles connect these trajectory patterns to recognizable properties of text (Figure 3). Repetition produces high ρ, bland passages have low D_2, and shuffling discourse changes the relation between accumulated local change and endpoint change. Human prose, code, source-matched LLM

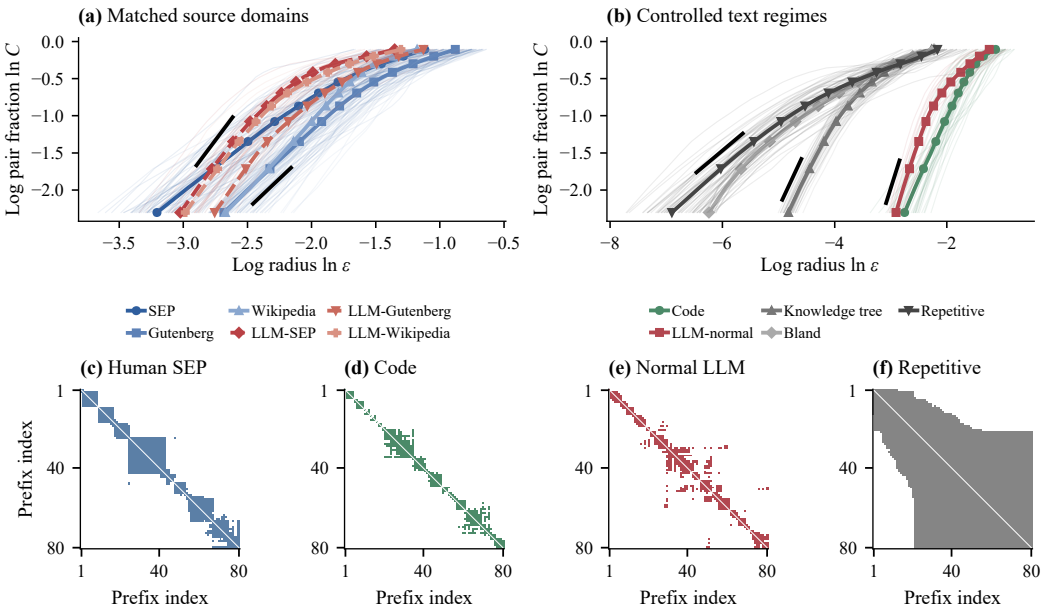

Figure 2: **Semantic structure across scales and along the response.** (a,b) Individual trajectories (faint) and group mean correlation-integral curves (bold) at 80 prefixes. The fitted log-log slope gives D_2. (c–f) Representative recurrence maps locate similar states in time; colored cells have cosine distance at most 0.05. This visualization threshold differs from the scalar descriptor's radius of 0.1.

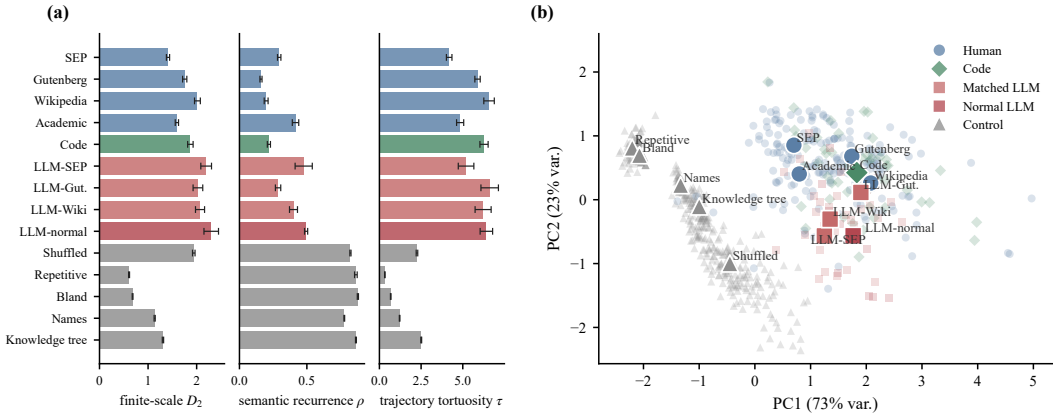

Figure 3: **Descriptor profiles and their projection.** (a) Group means of D_2, ρ, and τ, with standard errors across texts. (b) PCA of the same descriptors. The atlas contains 690 trajectories; the other 30 prompt-induced LLM degeneration outputs appear separately in the appendix.

text, and normal LLM generations have different group means across the three descriptors. The differences between human and generated text within matched source domains show that cumulative-prefix geometry captures aspects of semantic development beyond broad domain membership.

Increasing the prefix horizon reveals how this geometry develops (Appendix Figure A1). Human prose, code, and normal LLM generations generally show increasing D_2 and τ. Degeneration controls retain comparatively low expansion and high recurrence: adding text mainly adds states to an already concentrated region. PSD thus distinguishes continued text production from continued semantic expansion.

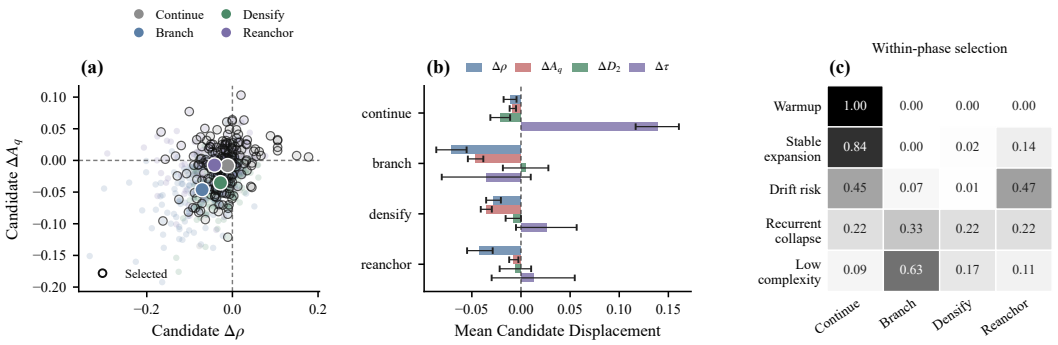

Figure 4: **Candidate effects and phase-conditioned selection.** (a) All 740 candidate displacements from the initial 20-prompt study; selected candidates are outlined and large markers show prompt-averaged actuator effects. (b) Mean actuator displacements with ± 1.96 SEM across prompt-level means. (c) Within-phase action fractions over the same run's 200 committed rounds, showing observed phases only.

Checkpoint spacing and fitting radius resolve different aspects of this development. Repetitive and bland controls remain in the low-expansion range across the tested D_2 windows, while the relative slopes of some natural-text groups vary with the radius range. Tortuosity is particularly sensitive to checkpoint spacing: the human–code ordering changes as the token step increases. These observations identify the scale of each comparison. D_2 describes neighborhood growth over the fitted radii, and τ describes the local changes resolved by the chosen checkpoints.

5.3 Trajectory feedback guides semantic development

The prefix state exposes available directions of intervention. The initial 20-prompt study shows how candidate extensions of a committed prefix induce different displacements (Figure 4a). Branching reduces recurrence more strongly, while re-anchoring preserves more prompt alignment (Figure 4b). The phase policy selects Continue in 51% of committed rounds; branching is most frequent in the low-complexity phase, while re-anchoring is most frequent in drift risk (Figure 4c). Appendix Figures A9 and A10 give the final-state frontier, complete fixed-policy results, and action counts.

Control acts throughout generation. Figure 5 follows 80 prompts over ten rounds with the 122B generator and Qwen states. After the first two rounds, recurrence rises under Continue while PSD Control maintains a lower-recurrence path; alignment increases under both trajectory-based selectors. Each point measures the committed cumulative prefix. Appendix Figure A8 gives all four configurations.

Expansion and prompt alignment improve together. Across all four generator–representation configurations on the 80 new LongFact prompts, PSD Control reduces trajectory-average recurrence by 0.0334–0.0764, increases prompt alignment by 0.0086–0.0139, and increases D_2 by 0.0321–0.1021 relative to Continue (Table 2; Appendix Table 6). All twelve paired prompt-level 95% bootstrap intervals exclude zero.

Each rollout is also evaluated with both representations, yielding eight rollout–evaluator combinations. Recurrence, alignment, and expansion retain the same directions of change in all eight, with 23 of the 24 intervals excluding zero. Both the representation used for feedback and the other representation therefore register the semantic change induced by control.

Continuous descriptors support candidate selection. Using candidate-extended PSD vectors without a phase table, PhaseAgnosticPareto reduces recurrence and increases alignment and D_2 relative to Continue in all eight rollout–evaluator combinations; all 24 intervals exclude zero. Compared with the phase policy, it achieves lower recurrence and higher D_2, with similar observed alignment.

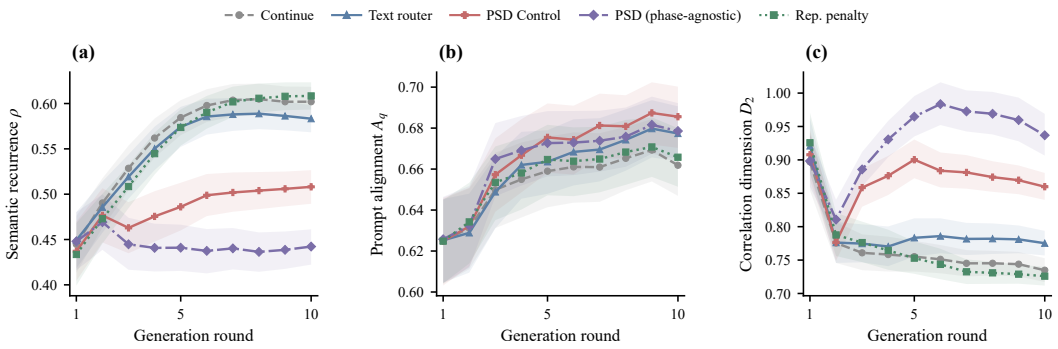

Figure 5: **Semantic development under trajectory feedback.** Mean (a) recurrence, (b) prompt alignment, and (c) D_2 over ten rounds, with pointwise 95% prompt-bootstrap intervals. All five policies use the same 80 prompts with Qwen3.5-122B-A10B-FP8 generation and Qwen3-Embedding-8B states. The phase-agnostic selector uses continuous PSD vectors without the phase table.

Table 2: Paired ten-round mean changes from Continue on 80 prompts per configuration, evaluated with the online representation. Bold marks the most favorable mean per metric and configuration. Appendix Tables 7–8 report 95% prompt-bootstrap intervals; Figure A7 gives the graphical comparison.

		Qwen3.5-9B		Qwen3.5-122B	
Policy	Change	Qwen	Harrier	Qwen	Harrier
Text router	$\Delta\rho \downarrow$	+0.0026	-0.0009	-0.0110	-0.0227
	$\Delta A_q \uparrow$	+0.0000	-0.0008	+0.0059	+0.0046
	$\Delta D_2 \uparrow$	+0.0248	+0.0013	+0.0264	+0.0163
PSD Control	$\Delta\rho \downarrow$	-0.0334	-0.0395	-0.0764	-0.0618
	$\Delta A_q \uparrow$	**+0.0139**	+0.0114	**+0.0127**	+0.0086
	$\Delta D_2 \uparrow$	+0.0673	+0.0321	+0.1021	+0.0499
PSD (phase-agnostic)	$\Delta\rho \downarrow$	**-0.0720**	**-0.0566**	**-0.1182**	**-0.0798**
	$\Delta A_q \uparrow$	+0.0132	**+0.0130**	+0.0109	**+0.0133**
	$\Delta D_2 \uparrow$	**+0.1434**	**+0.0407**	**+0.1644**	**+0.0715**
Rep. penalty	$\Delta\rho \downarrow$	-0.0006	+0.0008	-0.0073	+0.0059
	$\Delta A_q \uparrow$	-0.0011	-0.0035	+0.0031	+0.0025
	$\Delta D_2 \uparrow$	+0.0029	+0.0023	+0.0001	+0.0064

With the same actuator set and candidate count, PSD Control has lower recurrence and higher prompt alignment and D_2 point estimates than the recent-context TextOnlyRouter in all eight rollout–evaluator combinations. The paired intervals exclude zero in the favorable direction in seven recurrence comparisons, seven alignment comparisons, and all eight D_2 comparisons. The router selects from the prompt, recent prefix, and candidate texts.

The completed responses retain more prompt content. PSD Control produces shorter responses with higher prompt relevance and content-word coverage. Relevance, measured by the representation not used for online selection, increases by 0.0095–0.0165 across the four configurations. Prompt content-word coverage rises by 0.0264–0.0487, while responses are 42–62 words shorter on average.

ContinueRepPenalty reduces the fraction of exactly repeated sentences by 0.0341–0.1532, while its trajectory-average recurrence change ranges from -0.0073 to $+0.0059$ with the online representation (Table 2). PSD Control reduces trajectory recurrence across all four configurations, while its changes in exact sentence repetition vary across configurations (Table 10). The two measurements capture different aspects of repetition: repeated sentences count exact textual reuse, whereas semantic recurrence measures proximity among cumulative-prefix states.

The separate OpenAI factual-support audit, with 79 complete prompt pairs, estimates a PSD Control–Continue difference of $+0.0239$, with a 95% interval of $[-0.0574, 0.1050]$. The factual-support

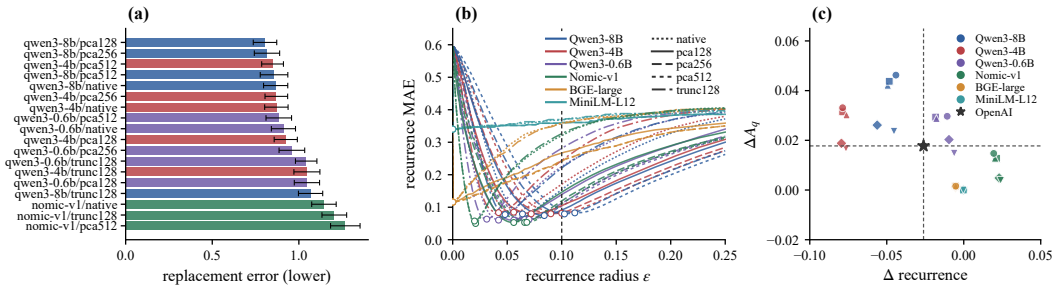

Figure 6: **Open-representation validation on 100 policy–prompt trajectories.** (a) The 18 lowest replacement errors, with ± 1.96 SEM; all 29 variants appear in Appendix Figure A12. (b) Recurrence MAE across radii; color identifies the model and line style the dimensional variant. The vertical line marks radius 0.1. (c) Controller-effect alignment.

difference remains unresolved. The demonstrated output gains concern prompt relevance and content-word coverage.

5.4 Different representations capture common directions of semantic change

Qwen and Harrier evaluators agree on the PSD Control–Continue effect direction for ρ, A_q, D_2, and raw τ in all four rollout configurations (16 comparisons). Spearman correlations across prompt–method observations range from 0.345 to 0.665; Kendall agreement in method ordering ranges from 0.500 to 1.000. Common effect directions coexist with differences in magnitudes and policy rankings.

On 100 policy–prompt trajectories, we also compare six open embedding models and their dimensional variants. Qwen3-Embedding-8B and Qwen3-Embedding-4B (Zhang et al., 2025) most closely reproduce the OpenAI descriptor organization; Nomic (Nussbaum et al., 2024), BGE-large-en-v1.5 (Xiao et al., 2023), and MiniLM (Wang et al., 2020) show larger differences. Appendix C.9 reports descriptor-wise ranks.

Radius calibration reduces recurrence MAE from 0.154 to 0.079 for Qwen3-Embedding-8B and from 0.171 to 0.085 for Qwen3-Embedding-4B. Figure 6 compares descriptor replacement, recurrence calibration, and control-effect alignment across the original model and dimensional variants. The prefix protocol transfers across semantic spaces, with radii and descriptor magnitudes interpreted within each space.

6 Conclusion

PSD represents long-form generation through cumulative-prefix geometry. The 720-trajectory benchmark connects this geometry to semantic expansion, repetition, and discourse organization. Across 80 new prompts, two generator scales, and two representations, trajectory feedback reduces recurrence and increases expansion and prompt alignment. Direct Pareto selection confirms that this continuous state can guide subsequent generation.

7 Limitations

PSD descriptor values depend on the representation and measurement settings, with recurrence radii requiring calibration when changing backends. PSD Control adds inference cost through candidate generation and prefix embedding. Our control experiments focus on long-form responses to LongFact prompts; further evaluation on other tasks and generator families would help assess the broader applicability of trajectory feedback.

AI USE STATEMENT

Generative AI tools are part of the experimental system studied in this work: they produce the long-form generations and candidate continuations described in Sections 4 and 5.3. We also used generative AI tools to refine and polish the language of this manuscript and to assist with debugging code. We take full responsibility for verifying all experimental protocols, code, outputs, quantitative results, citations, and claims, and for the entire content of this work.

ETHICS STATEMENT

The study uses public text, generated text, and synthetic controls, and does not involve private personal data or human-subject experiments. Prefix-state control may improve monitoring of long generations, and the same capability could be applied to undesirable content generation. We discuss these uses and release considerations in Appendix C.11.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 define the cumulative-prefix construction, descriptors, controller, baselines, and evaluation protocol. The appendix provides the full data inventory, phase and selector rules, statistical procedures, robustness analyses, model identifiers, calibration details, runtime measurements, and additional results. The anonymized supplementary package contains the measurement and controller code, cached descriptors, prompts, run manifests, figure-level tables, and scripts used to reproduce the reported analyses.

REFERENCES

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 169–174, 2018.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*, 2020.

Xin Du and Kumiko Tanaka-Ishii. Correlation dimension of auto-regressive large language models. *arXiv preprint arXiv:2510.21258*, 2025.

Jean-Pierre Eckmann, S. Oliffson Kamphorst, and David Ruelle. Recurrence plots of dynamical systems. *Europhysics Letters*, 4(9):973–977, 1987.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. doi: 10.1038/s41586-024-07421-0.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6894–6910, 2021.

Peter Grassberger and Itamar Procaccia. Measuring the strangeness of strange attractors. *Physica D: Nonlinear Phenomena*, 9(1–2):189–208, 1983.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020.

Mingjian Jiang, Yangjun Ruan, Prasanna Sattigeri, Salim Roukos, and Tatsunori Hashimoto. Graph-based uncertainty metrics for long-form language model generations. In *Advances in Neural Information Processing Systems*, 2024.

Holger Kantz and Thomas Schreiber. *Nonlinear Time Series Analysis*. Cambridge University Press, 2 edition, 2003.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 6691–6706, 2021.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, 2023.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Norbert Marwan, M. Carmen Romano, Marco Thiel, and Jürgen Kurths. Recurrence plots for the analysis of complex systems. *Physics Reports*, 438(5–6):237–329, 2007.

Microsoft. Harrier-OSS-v1-27B Model Card. Hugging Face model card, 2026. URL `https://huggingface.co/microsoft/harrier-oss-v1-27b`.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for LLMs from semantic similarities. In *Advances in Neural Information Processing Systems*, 2024.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Xin Qiu and Risto Miikkulainen. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space. In *Advances in Neural Information Processing Systems*, 2024.

Qwen Team. Qwen3.5-9B and Qwen3.5-122B-A10B-FP8 Model Cards. Hugging Face model cards, 2026. URL `https://huggingface.co/Qwen/Qwen3.5-9B`. The Qwen3.5-122B-A10B-FP8 card is available at `https://huggingface.co/Qwen/Qwen3.5-122B-A10B-FP8`.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2019.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7881–7892, 2020.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023.

Ava Spataru, Eric Hambro, Elena Voita, and Nicola Cancedda. Know when to stop: A study of semantic drift in text generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2024.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788, 2020.

Charles L. Webber, Jr. and Joseph P. Zbilut. Dynamical assessment of physiological systems and states using recurrence plot strategies. *Journal of Applied Physiology*, 76(2):965–973, 1994.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. Long-form factuality in large language models. In *Advances in Neural Information Processing Systems*, 2024.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*, 2020.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-Pack: Packaged resources to advance general chinese embedding, 2023. URL `https://arxiv.org/abs/2309.07597v1`.

Nan Xu, Chunting Zhou, Asli Celikyilmaz, and Xuezhe Ma. Look-back decoding for open-ended text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1039–1050, 2023.

Kevin Yang and Dan Klein. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 3511–3535, 2021.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, 2023.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

A MEASUREMENT PIPELINE AND REPRESENTATIVE TRAJECTORIES

Algorithm 2 Prefix Semantic Dynamics

Require: Text $x_{1:T}$, representation E, checkpoints $\{t_k\}_{k=1}^{K}$, radius grid, fitting window, recurrence radius ϵ^*

1: **for** $k = 1, \ldots, K$ **do**
2: $\quad z_k \leftarrow E(x_{1:t_k})/\|E(x_{1:t_k})\|_2$
3: **end for**
4: Compute $C(\epsilon)$ from unordered pairs using Equation 2.
5: Fit D_2 from positive-distance fractions C_+, evaluate $\rho = C(\epsilon^*)$, and compute τ using Equations 3–5.
6: **return** $Z = (z_1, \ldots, z_K)$ and (D_2, ρ, τ)

Figure 2c–f in the main text shows the four representative recurrence maps.

B ADDITIONAL EXPERIMENTAL DETAILS

B.1 DETAILED DIAGNOSTIC BENCHMARK INVENTORY

Table 3 expands the 720-trajectory diagnostic benchmark. The subsets are grouped by experimental role. Natural-language references test ordinary discourse; code supplies structured non-prose text; manufactured controls impose known trajectory changes; source-matched generations reduce broad topic mismatch; and prompt-induced LLM regimes test whether generation conditions move the trajectory in predicted directions.

Source-matched LLM generations. The source-matched subsets are generated counterparts to the human source families. Prompts request a long-form passage in the same broad domain and discourse style as SEP-like exposition, Gutenberg-like literature, or Wikipedia-like encyclopedic writing. Generated and human texts are truncated and measured by the same cumulative-prefix pipeline. The comparison does not remove every content difference, but it reduces the explanation that the observed displacement is caused only by broad topic.

Manufactured controls. Each control anchors one interpretable direction. Shuffling preserves local lexical material while disrupting discourse order. Repetition revisits a small semantic neighborhood. Bland text suppresses specificity and expansion. Random-name streams provide local variation without sustained discourse, and synthetic trees impose simple hierarchical branching. These controls are diagnostic perturbations rather than natural corpora. Their purpose is to connect D_2, ρ, and τ to known changes in how a text unfolds.

B.2 REPRODUCIBILITY, COMPUTE, AND ASSETS

The supplementary package contains the PSD measurement code, controller runner, plotting scripts, cached descriptors, figure-level tables, prompts, and run manifests. Model calls are separated from analysis. Plotting and statistical scripts read cached CSV files, so all reported comparisons can be recomputed without repeating generation. The main diagnostic pipeline uses `text-embedding-3-large`, recursive boundary splitting, and a 100-token target checkpoint step. Open-representation experiments use separately recorded model revisions and radii.

No model is trained. The main cost comes from long-form generation, candidate generation, and cumulative-prefix embedding. Descriptor computation, bootstrap inference, and plotting are CPU-feasible from cached outputs. The expanded local study used four 48-GB A6000 GPUs for tensor-parallel generation and representation inference. Table 12 reports measured per-response end-to-end wall-clock times, and Table 5 gives the exact rollout counts.

All natural-language and code sources are public sources or derived prompt and metadata records. When redistribution terms restrict raw source text, the package supplies prompts, descriptors, and

Table 3: **Detailed diagnostic benchmark inventory.** Text counts refer to the cached PSD benchmark used for the main descriptor and robustness figures. Human, code, and manufactured-control corpora are measured as fixed text trajectories; LLM corpora are generated long-form outputs measured with the same cumulative-prefix pipeline.

Family	Subset	n	Experimental role
Human prose	SEP	50	expository natural-language reference
Human prose	Gutenberg	50	literary natural-language reference
Human prose	Wikipedia	50	encyclopedic natural-language reference
Human prose	Academic	50	technical/academic prose reference
Structured text	Python code	50	structured non-prose baseline
Manufactured controls	Shuffled text	200	order-disrupted semantic-control trajectory
Manufactured controls	Repetitive text	50	high-recurrence degeneration probe
Manufactured controls	Bland generic text	50	low-specificity degeneration probe
Manufactured controls	Random-name stream	50	locally varied but weakly structured control
Manufactured controls	Synthetic knowledge tree	50	hierarchical structured-control trajectory
Source-matched LLM	Matched SEP	10	generated counterpart to SEP-like prompts
Source-matched LLM	Matched Gutenberg	10	generated counterpart to literary prompts
Source-matched LLM	Matched Wikipedia	10	generated counterpart to encyclopedic prompts
LLM prompt regimes	Normal	10	default long-form generation regime
LLM prompt regimes	Bland	10	prompt-induced low-specificity regime
LLM prompt regimes	Incoherent	10	prompt-induced drift/incoherence regime
LLM prompt regimes	Repetitive	10	prompt-induced repetition regime

reconstruction scripts. Generated outputs, synthetic controls, controller logs, and figure-level CSV files are indexed by run manifests.

B.3 Prefix construction and finite-scale robustness

Every main PSD state embeds a complete cumulative prefix $x_{1:t_k}$. Recursive chunking only selects stable checkpoint boundaries; it does not replace the cumulative prefix with an isolated chunk. Figure A2 varies the token step. Figure A3 shows all pairwise descriptor projections, and Figure A4 varies the D_2 fit window.

The fitted D_2 is an empirical finite-scale statistic of the observed trajectory. It is not an estimate of an asymptotic attractor dimension. All reported radii and slopes use cosine distance. Under unit normalization, $d_E = \sqrt{2d_{\cos}}$; corresponding neighborhoods contain the same pairs, and the Euclidean scaling slope is twice the cosine-distance slope. A common distance rescaling shifts the log-distance axis; the fitted slope is unchanged when the fit window is remapped by the same scale. The window, checkpoint schedule, and representation can still alter finite-sample estimates, which motivates the sensitivity analyses below.

The main descriptor atlas in Figure 3 uses the $0.10 \leq C \leq 0.50$ fitting window, corresponding to the blue bars in Figure A4. The gray "Original mid-slope" bars use $0.15 \leq C \leq 0.85$. Both are fits to the same cached diagnostic spectra. Figure A2 instead uses the earlier boundary-chunking benchmark, with a different corpus composition and text lengths, and the original mid-slope estimator. Its curves average method summaries within each family. The step-100 points therefore do not reproduce the individual text-group means in Figure 3.

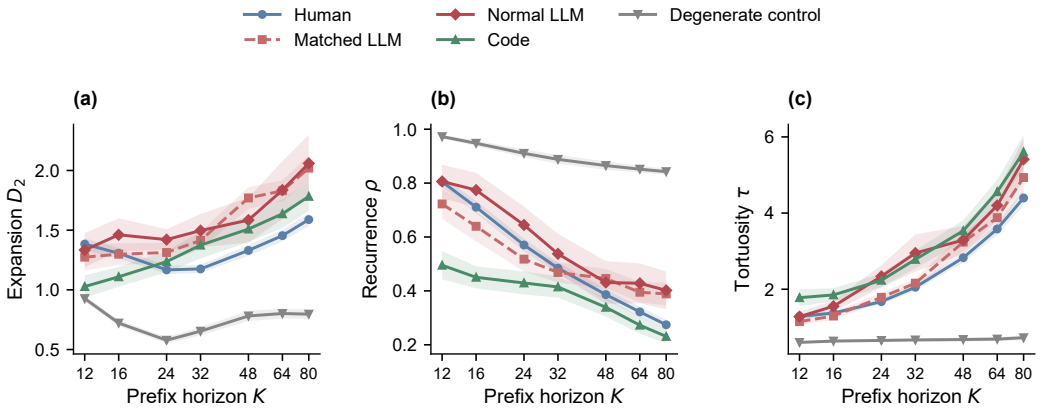

Figure A1: **Development of geometry over prefix horizons.** Descriptors are recomputed as more cumulative prefixes are observed. Natural text develops greater expansion and accumulated path change; degeneration controls retain low expansion and high recurrence. Bands show mean ± 1.96 standard errors across eligible trajectories.

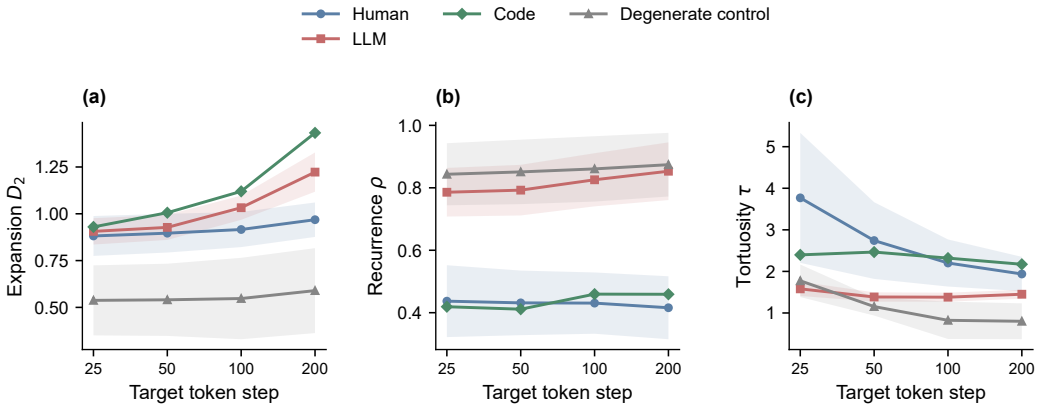

Figure A2: **Token-step robustness on the separate boundary-chunking benchmark.** D_2 uses the original $0.15 \leq C \leq 0.85$ window. Curves summarize method means within each family across checkpoint resolutions; they are not the step-100 values of the main descriptor atlas. Bands summarize variation across method summaries within each family; a single-method family has no band.

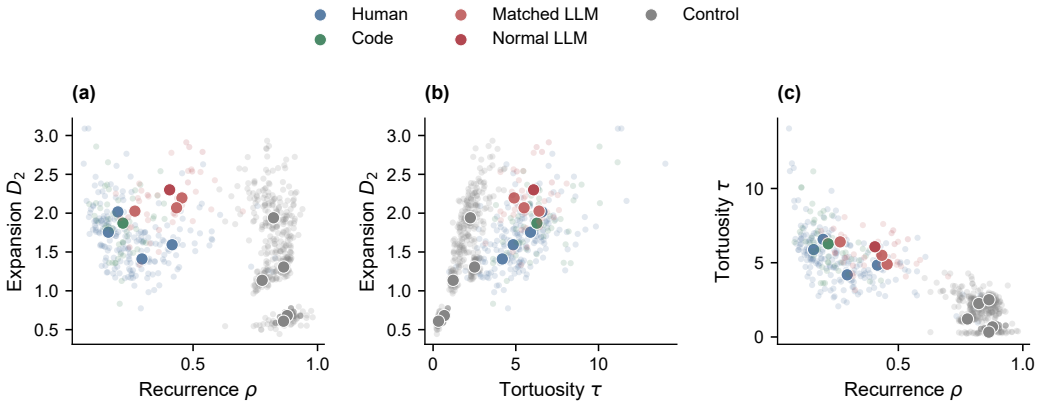

Figure A3: **Pairwise descriptor projections.** The three pairwise views of D_2, ρ, and τ expose the benchmark organization without relying on PCA. Light points are trajectories; large markers are subset means.

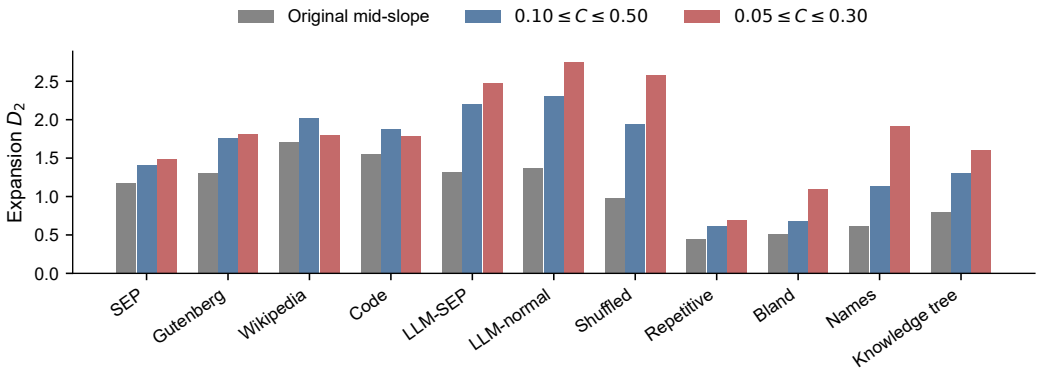

Figure A4: D_2 **fit-window sensitivity.** Blue bars use the main atlas window $0.10 \leq C \leq 0.50$; gray "Original mid-slope" bars use $0.15 \leq C \leq 0.85$. Alternative fitting windows change magnitudes and some natural-text orderings; repetitive and bland controls remain in the low-expansion range.

B.4 Prompt regimes move LLM trajectories

The prompt-regime study asks whether generation conditions move trajectories before online control is introduced. Figures A5 and A6 compare normal, bland, incoherent, and repetitive prompting at the correlation-integral and prefix-horizon levels. These comparisons show how prompt conditions change cumulative-prefix geometry and motivate evaluating prompt actuators during generation.

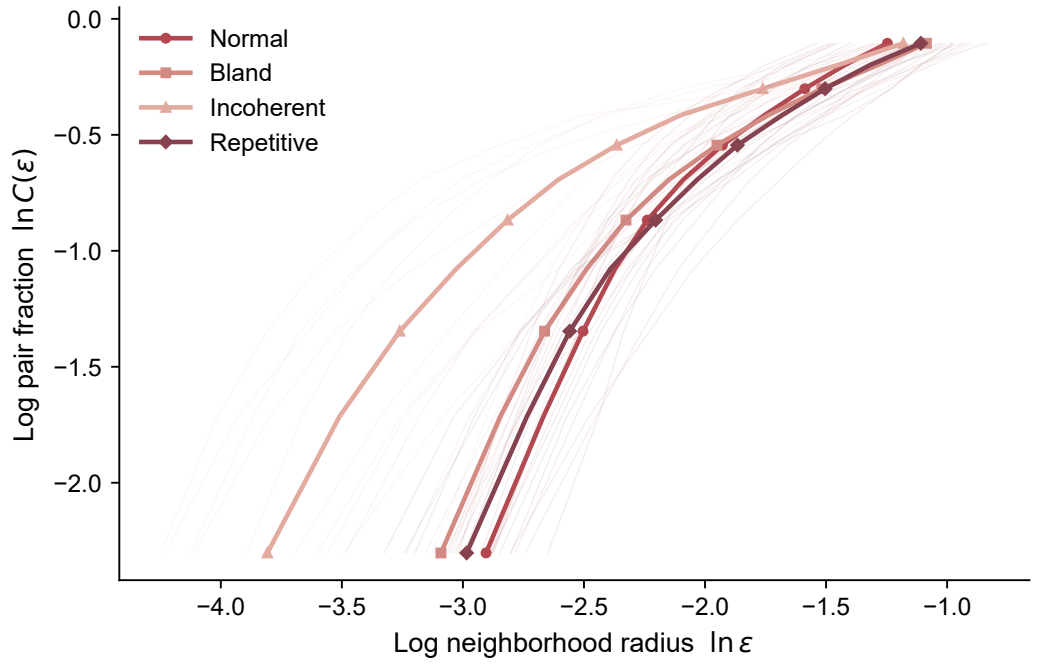

Figure A5: **Prompt-regime correlation-integral spectra.** Prompt-induced LLM regimes differ in their log-log $C(\epsilon)$ curves before scalar descriptor fitting.

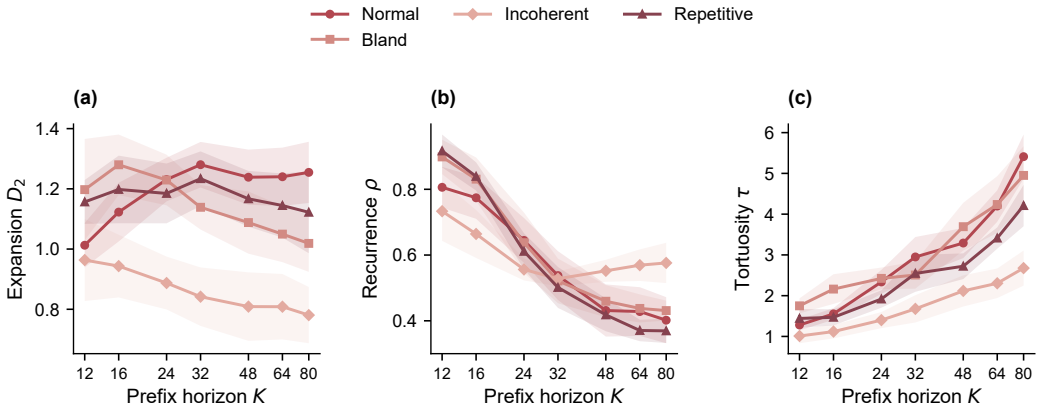

Figure A6: **Prompt-regime horizon trends.** Regime differences persist as additional cumulative prefixes enter the trajectory.

C PSD Control Details

Algorithm 1 in the main text gives the shared generation loop. This appendix specifies its phase diagnosis, warmup, actuator instructions, and candidate-selection policies.

C.1 Phase diagnosis and selection policy

At each round, PSD Control computes D_2, ρ, τ, and $A_q = \cos(E(q), z_K)$ from the committed cumulative-prefix trajectory. The phase rule compares the current descriptor state with prior states from the same response. After four candidate continuations are generated in parallel, the controller forms a Pareto set and applies the phase-specific selection rule in Table 4. The phase label is an interpretable policy input; the continuous trajectory state is the underlying representation.

Table 4: **PSD Control phase diagnosis and actuator selection.** At each round, PSD descriptors are computed on the current cumulative-prefix trajectory. The controller compares the current state with the median of previous states for the same generation and then chooses among Pareto-filtered candidate continuations.

Diagnosed phase	Trigger rule	Selection rule among candidate continuations
Warmup	$K < K_{\text{warmup}}$ or fewer than two observed states	Continue to accumulate a stable prefix trajectory
Drift risk	$A_q^{(r)} - A_q^{(r-1)} < -0.008$	Choose the Pareto candidate with highest A_q
Recurrent collapse	$\rho^{(r)} > \text{median}(\rho^{(<r)})$ and $D_2^{(r)} < \text{median}(D_2^{(<r)})$	Choose the Pareto candidate with lowest ρ
Low complexity	$D_2^{(r)} < \text{median}(D_2^{(<r)})$	Choose the Pareto candidate with highest D_2
Bland linear	$\tau^{(r)} < \text{median}(\tau^{(<r)})$ and $\rho^{(r)} > \text{median}(\rho^{(<r)})$	Choose the Pareto candidate with highest τ
Stable expansion	none of the above	Prefer continuation within the Pareto set; otherwise maximize A_q

C.2 Actuator instructions

The following instruction is appended to the prompt and current-prefix context for each actuator.

Table 5: **Controller and representation-validation studies.** Prompt counts are statistical units within a configuration. The four expanded configurations reuse the same 80 prompts and are not pooled as 320 independent observations.

Study	Generator / representation	Size	Purpose
Initial controller study	GPT generation / OpenAI embedding	20 prompts	fixed-policy frontier and actuator mechanism
Open representation replacement	six open embedding models and variants	100 trajectories	descriptor ranking and radius calibration
Open representation controller	GPT generation / Qwen3-Embedding-8B	20 prompts	open-representation controller replication
Additional OpenAI control	GPT-5-mini / text-embedding-3-large	80 prompts	separate factual-support audit; 79 common complete cases
Expanded control, cell 1	Qwen3.5-9B / Qwen3-Embedding-8B	80 prompts	frozen controller and matched baselines
Expanded control, cell 2	Qwen3.5-9B / Harrier-OSS-v1-27B	80 prompts	generator-matched representation replication
Expanded control, cell 3	Qwen3.5-122B-A10B-FP8 / Qwen3-Embedding-8B	80 prompts	larger-generator replication
Expanded control, cell 4	Qwen3.5-122B-A10B-FP8 / Harrier-OSS-v1-27B	80 prompts	larger-generator representation replication
Cross-evaluator analysis	Qwen and Harrier remeasurement of all four cells	8 evaluator cells	effect-direction and ordering transfer

Continue CONTINUE: Continue naturally from the current prefix for one short step. Preserve prompt relevance, avoid restarting, avoid outlining, and add new material without excessive semantic jumping.

Branch BRANCH: Open one new but prompt-relevant semantic branch. The branch should be distinct from the recent prefix, connected to the original prompt, and concrete enough to support continued generation. Do not list many unrelated branches.

Densify DENSIFY: Add compact, specific, prompt-relevant details in the current semantic neighborhood: names, dates, mechanisms, measurements, distinctions, and concrete consequences. Avoid generic filler and avoid changing topic.

Reanchor RE-ANCHOR: Reconnect the continuation to the original prompt's central question, scope, and constraints. Do not restart the answer; steer it back while adding new content.

C.3 Expanded 80-prompt cross-backend study

Prompt set and statistical unit. We selected 80 previously unused prompts from the LongFact concept and object collections with a fixed seed. The same prompts were used in all four generator–online-representation configurations. Each configuration contains 80 prompts, eight methods, 640 final texts, and 6,400 round states. The complete matrix contains 2,560 nonempty, hash-unique final texts. Prompt is the unit of inference. The four configurations test replication under different systems and are not pooled as 320 independent prompts.

Generators and representations. The generators are `Qwen/Qwen3.5-9B` and `Qwen/Qwen3.5-122B-A10B-FP8` (Qwen Team, 2026), both run with thinking disabled and a 16,384-token context. Online semantic states come from `Qwen/Qwen3-Embedding-8B` (Zhang et al., 2025) or the normalized LAST-pooled hidden state of `microsoft/harrier-oss-v1-27b` (Microsoft, 2026). Qwen is a dedicated embedding model. Harrier is used as a hidden-state representation backend. Its role is therefore representation replication, not a claim that it is a standard sentence embedder. The model revisions, dimensions, pooling configuration, and service settings are recorded in the released registry.

Table 6: PSD Control minus Continue on the same 80 prompts per configuration. Entries are paired ten-round mean differences [95% bootstrap interval]. Qwen denotes Qwen3-Embedding-8B; Harrier denotes Harrier-OSS-v1-27B with LAST pooling. The 122B generator is Qwen3.5-122B-A10B-FP8.

Generator / state	$\Delta\rho$	ΔA_q	ΔD_2
9B / Qwen	-0.0334 $[-0.0482, -0.0183]$	$+0.0139$ $[+0.0093, +0.0187]$	$+0.0673$ $[+0.0501, +0.0847]$
9B / Harrier	-0.0395 $[-0.0549, -0.0245]$	$+0.0114$ $[+0.0084, +0.0146]$	$+0.0321$ $[+0.0200, +0.0442]$
122B / Qwen	-0.0764 $[-0.0895, -0.0631]$	$+0.0127$ $[+0.0086, +0.0171]$	$+0.1021$ $[+0.0861, +0.1185]$
122B / Harrier	-0.0618 $[-0.0738, -0.0493]$	$+0.0086$ $[+0.0057, +0.0116]$	$+0.0499$ $[+0.0396, +0.0599]$

Table 7: Matched-policy comparisons with Qwen3.5-9B. All entries are paired ten-round differences from Continue [95% prompt-bootstrap interval], using the online representation as evaluator; $n = 80$ per policy and configuration. PSD Pareto denotes PhaseAgnosticPareto.

Policy	$\Delta\rho$	ΔA_q	ΔD_2
Qwen3-Embedding-8B			
Text router	$+0.0026$ $[-0.0104, +0.0155]$	$+0.0000$ $[-0.0045, +0.0046]$	$+0.0248$ $[+0.0109, +0.0392]$
PSD Control	-0.0334 $[-0.0482, -0.0183]$	$+0.0139$ $[+0.0093, +0.0187]$	$+0.0673$ $[+0.0501, +0.0847]$
PSD Pareto	-0.0720 $[-0.0908, -0.0537]$	$+0.0132$ $[+0.0076, +0.0186]$	$+0.1434$ $[+0.1245, +0.1627]$
Rep. penalty	-0.0006 $[-0.0133, +0.0125]$	-0.0011 $[-0.0050, +0.0025]$	$+0.0029$ $[-0.0092, +0.0151]$
Harrier-OSS-v1-27B (LAST)			
Text router	-0.0009 $[-0.0156, +0.0144]$	-0.0008 $[-0.0038, +0.0023]$	$+0.0013$ $[-0.0090, +0.0112]$
PSD Control	-0.0395 $[-0.0549, -0.0245]$	$+0.0114$ $[+0.0084, +0.0146]$	$+0.0321$ $[+0.0200, +0.0442]$
PSD Pareto	-0.0566 $[-0.0714, -0.0427]$	$+0.0130$ $[+0.0099, +0.0163]$	$+0.0407$ $[+0.0307, +0.0509]$
Rep. penalty	$+0.0008$ $[-0.0140, +0.0154]$	-0.0035 $[-0.0064, -0.0006]$	$+0.0023$ $[-0.0071, +0.0113]$

Qwen uses recurrence radius 0.064, selected on the original 20-prompt calibration set and frozen before the new prompts. Harrier uses recurrence radius 0.036, fixed before the 80-prompt rollouts. This is a frozen-backend replication without a separate held-out calibration split.

Generation and policy settings. Each prompt receives one shared 360-token seed. All methods then run for ten rounds with at most 300 new tokens per round. Prefix checkpoints target 100 tokens, and warmup ends at $K = 8$. PSD Control, PhaseAgnosticPareto, TextOnlyRouter, and each shuffled-phase replica generate one candidate for each of the four actuator types per controlled round.

PhaseAgnosticPareto removes phase text from candidate prompting and does not use the phase label, running medians, or descriptor trends. It Pareto-filters the candidate-extended trajectories under lower ρ and higher D_2, τ, and A_q. Each surviving candidate receives an equal-weight rank sum across these four quantities. Ties follow the fixed order Continue, Branch, Densify, ReAnchor.

TextOnlyRouter receives the original prompt, the latest 900 words of the current prefix, and the four candidate texts. It receives no PSD quantity or phase label. The same generator chooses one actuator at temperature 0 and top-p 1, with thinking disabled. This adds one short router call per controlled

Table 8: Matched-policy comparisons with Qwen3.5-122B-A10B-FP8. All entries are paired ten-round differences from Continue [95% prompt-bootstrap interval], using the online representation as evaluator; $n = 80$ per policy and configuration. PSD Pareto denotes PhaseAgnosticPareto.

Policy	$\Delta\rho$	ΔA_q	ΔD_2
Qwen3-Embedding-8B			
Text router	-0.0110 $[-0.0239, +0.0014]$	$+0.0059$ $[+0.0019, +0.0101]$	$+0.0264$ $[+0.0134, +0.0397]$
PSD Control	-0.0764 $[-0.0895, -0.0631]$	$+0.0127$ $[+0.0086, +0.0171]$	$+0.1021$ $[+0.0861, +0.1185]$
PSD Pareto	-0.1182 $[-0.1339, -0.1029]$	$+0.0109$ $[+0.0067, +0.0152]$	$+0.1644$ $[+0.1474, +0.1814]$
Rep. penalty	-0.0073 $[-0.0198, +0.0052]$	$+0.0031$ $[-0.0003, +0.0063]$	$+0.0001$ $[-0.0121, +0.0124]$
Harrier-OSS-v1-27B (LAST)			
Text router	-0.0227 $[-0.0353, -0.0099]$	$+0.0046$ $[+0.0014, +0.0079]$	$+0.0163$ $[+0.0081, +0.0245]$
PSD Control	-0.0618 $[-0.0738, -0.0493]$	$+0.0086$ $[+0.0057, +0.0116]$	$+0.0499$ $[+0.0396, +0.0599]$
PSD Pareto	-0.0798 $[-0.0910, -0.0687]$	$+0.0133$ $[+0.0099, +0.0167]$	$+0.0715$ $[+0.0623, +0.0806]$
Rep. penalty	$+0.0059$ $[-0.0050, +0.0166]$	$+0.0025$ $[-0.0001, +0.0051]$	$+0.0064$ $[-0.0016, +0.0142]$

round. Because the router truncates prefixes longer than 900 words, it is a matched recent-context baseline, not a full-prefix information-equivalent control.

ContinueRepPenalty uses the same continuation instruction with repetition penalty 1.1, fixed before evaluation. Three ShuffledPhase methods run complete autoregressive trajectories. They retain warmup and deterministically replace each later diagnosed label with one of three prompt-round-specific shuffled labels. These runs test whether labels change behavior; they do not by themselves prove that the phase mapping is optimal.

Finite-sample implementation. The controller implementation fits scaling after excluding zero-distance pairs. With $d_{ij} = d_{\cos}(z_i, z_j)$, the fitted pair fraction is

$$C_+(\epsilon) = \frac{\sum_{i<j} \mathbf{1}[0 < d_{ij} \le \epsilon]}{\sum_{i<j} \mathbf{1}[d_{ij} > 0]}.$$

Thus $C_+ = C$ when all pair distances are positive. Recurrence uses the all-pair fraction C in Equation 2, including zero-distance pairs. Non-finite distances are excluded before either computation. The controller scaling fit uses finite positive pair distances at quantiles $0.10, 0.18, 0.26, 0.34, 0.42, 0.50, 0.58, 0.66, 0.74, 0.82, 0.90$. A fit needs at least eight positive pairs and four distinct radii with empirical fractions between 0.02 and 0.98. Tortuosity is undefined when endpoint cosine distance is below 10^{-12}. Undefined descriptors are omitted only from the corresponding descriptor average.

Inference. The primary endpoint is the prompt-level mean of each descriptor across ten generation rounds. We compute method differences within prompt, then draw 10,000 paired bootstrap resamples to obtain 95% intervals. Paired sign-flip tests use 100,000 randomizations. Recurrence and prompt alignment are co-primary quantities; D_2 measures expansion, and raw τ remains descriptive. Final-state results provide a secondary endpoint. Only 16 of 51,200 round rows have undefined D_2, all at round 1 with $K = 4$, where too few points are available for the scaling fit.

Table 9 shows that PSD Control versus Continue has the intended direction in every evaluator cell for recurrence, alignment, and D_2. PhaseAgnosticPareto reproduces all three effects with the phase table removed. Its recurrence and D_2 changes are stronger than those of the phase-based policy in all eight cells, with small observed alignment differences. The result strengthens the general control

Table 9: **Trajectory-average robustness across four rollouts and two evaluators.** Each comparison contains eight evaluator cells formed by remeasuring four generator–online-representation rollouts with Qwen and Harrier. Sign counts cells with the favorable point-estimate direction; CI counts prompt-level 95% intervals that exclude zero in that direction.

Metric	Favorable sign	Favorable CI	Mean-effect range
PSD Control vs Continue			
recurrence ρ	8/8	7/8	$[-0.0764, -0.0136]$
prompt alignment A_q	8/8	8/8	$[+0.0076, +0.0139]$
expansion D_2	8/8	8/8	$[+0.0320, +0.1021]$
PhaseAgnosticPareto vs Continue			
recurrence ρ	8/8	8/8	$[-0.1182, -0.0283]$
prompt alignment A_q	8/8	8/8	$[+0.0063, +0.0159]$
expansion D_2	8/8	8/8	$[+0.0407, +0.1644]$
PSD Control vs TextOnlyRouter			
recurrence ρ	8/8	7/8	$[-0.0654, -0.0129]$
prompt alignment A_q	8/8	7/8	$[+0.0001, +0.0139]$
expansion D_2	8/8	8/8	$[+0.0152, +0.0757]$

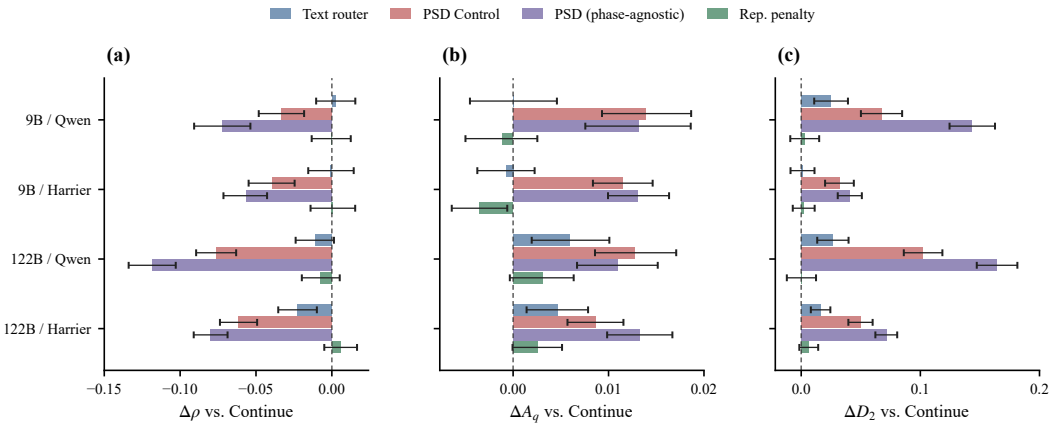

Figure A7: **Policy comparisons across generators and representations.** Graphical companion to Table 2. Paired ten-round mean changes from Continue in (a) recurrence, (b) alignment, and (c) D_2, with 95% prompt-bootstrap intervals. Each configuration uses the same 80 prompts and its online representation for evaluation. The two PSD selectors and text router use four candidates per round; repetition-penalized Continue uses one.

conclusion: cumulative-prefix geometry is an actionable state that can support multiple policies, including a simpler continuous selector.

Against TextOnlyRouter, PSD Control has lower recurrence and higher alignment and D_2 point estimates in all eight cells. Seven recurrence intervals, seven alignment intervals, and all eight D_2 intervals exclude zero in the favorable direction. Thus a no-PSD recent-context router and equal candidate budget do not reproduce the same geometric operating point.

C.4 Final-output comparisons

We evaluate final outputs separately from the state variables used for selection. Prompt relevance is the cosine similarity between prompt and response under each evaluator representation. To reduce same-representation circularity, the main text emphasizes the evaluator that did not supply the online state. Prompt content-word coverage removes a standard English stopword list, applies Porter stemming, and measures the fraction of prompt content stems present in the response. Repeated sentences measures the fraction of exactly repeated sentences. Table 10 compares the same three

Table 10: Final-output comparisons with Continue. Entries give the minimum and maximum configuration-level mean paired differences, not confidence intervals. Prompt relevance uses eight evaluator cells; text-only metrics use four configurations, each with 80 prompts. Arrows indicate the preferred direction. Content-word coverage removes stopwords and uses Porter stems. Per-configuration estimates and 95% prompt-bootstrap intervals are provided in `data/final_output_effects.csv`.

Metric	PSD Control	Phase-agnostic Pareto	ContinueRepPenalty
Prompt relevance ↑	$[+0.0095, +0.0266]$	$[+0.0027, +0.0212]$	$[-0.0313, +0.0066]$
Content-word coverage ↑	$[+0.0264, +0.0487]$	$[+0.0171, +0.0475]$	$[-0.0103, +0.0017]$
Repeated sentences ↓	$[-0.0144, +0.0878]$	$[-0.0891, -0.0087]$	$[-0.1532, -0.0341]$

quantities for all three policies. These quantities are output diagnostics, not replacements for factual verification.

PSD Control improves prompt relevance and content-word coverage across the tested configurations, while its changes in exact sentence repetition have mixed directions. PhaseAgnosticPareto has favorable mean changes in all three quantities across the tested configurations. ContinueRepPenalty consistently reduces exact sentence repetition, with mixed changes in relevance and coverage. The ranges summarize effect magnitudes; statistical uncertainty is reported separately for each configuration in the accompanying CSV.

The separate factual-support audit evaluates an additional 80-prompt GPT-5-mini run with `text-embedding-3-large`, listed separately from the initial 20-prompt study in Table 5. It is method-blinded and retrieval-backed, using `gpt-4.1-2025-04-14` to assess sampled claims; it does not evaluate the Qwen rollouts. The reported analysis uses the 79 prompts with complete audit results across all four policies. One TextOnlyRouter output lacked auditable retrieval sources and its prompt was excluded from this common comparison set. Within that set, the PSD Control–Continue sampled support-rate difference is $+0.0239$ with 95% interval $[-0.0574, +0.1050]$. The interval is too wide to establish improvement or non-inferiority. This audit leaves the factual-support difference unresolved.

C.5 Round-wise control in all four expanded configurations

Figure A8 extends the main-text process view to every generator–representation combination and to expansion. Each curve averages measurements from the same 80 prompts; bands resample complete prompt trajectories 10,000 times and give pointwise 95% intervals. No smoothing or interpolation is applied. At round 1, the D_2 fit is undefined for one prompt in the 9B/Qwen Pareto curve and one in the 9B/Harrier router curve; these two points use 79 finite observations. Every other displayed point uses 80. Undefined fits are not imputed.

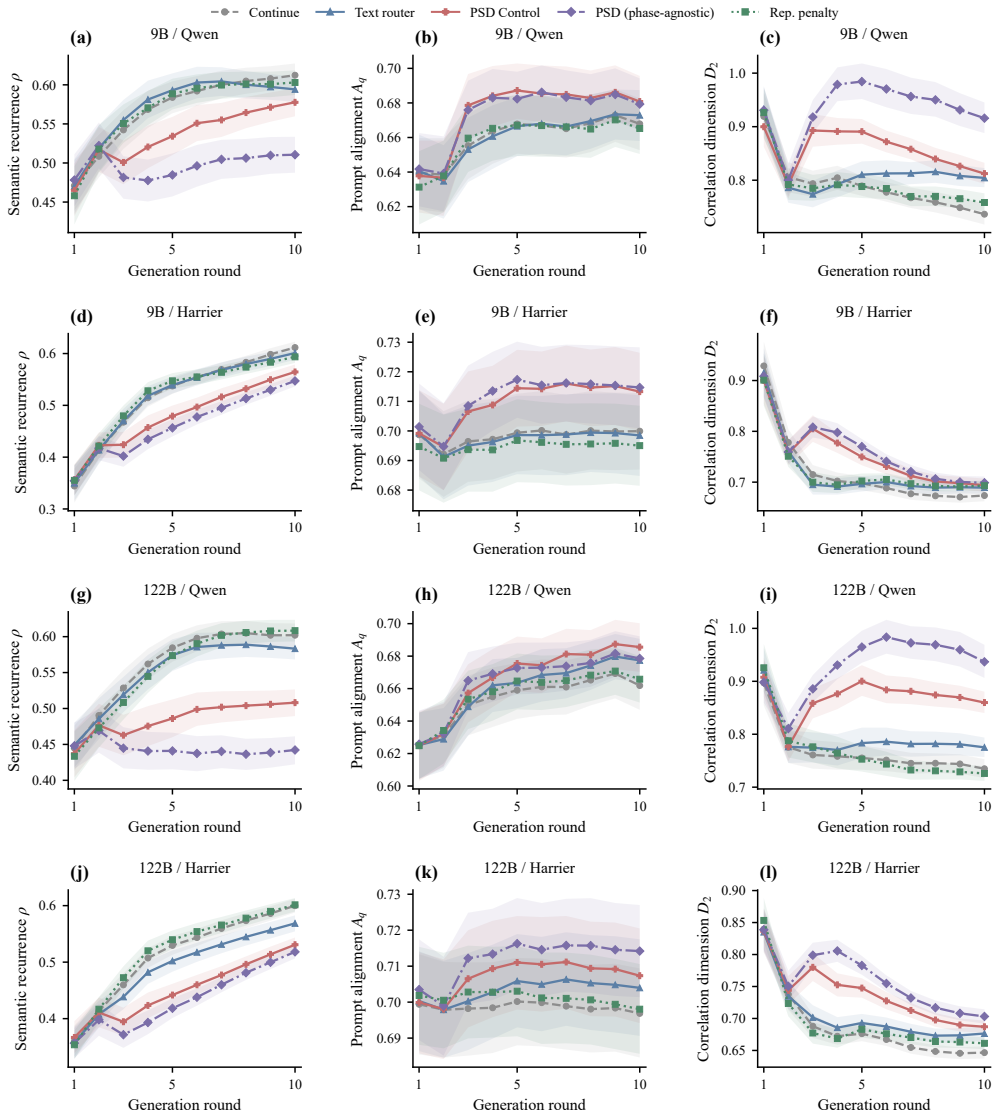

Figure A8: **The generation process across all expanded configurations.** Columns show recurrence, prompt alignment, and expansion; rows identify the generator and online representation. Evaluation uses the online representation. Colors and markers are shared with Figure 5.

C.6 PHASE-USE AND THRESHOLD AUDITS

Only the drift-risk rule contains a fixed numerical cutoff, $A_q^{(r)} - A_q^{(r-1)} < -0.008$. Table 11 replays the policy under $\pm 10\%$ and $\pm 20\%$ changes. Across all non-warmup rounds, labels retain 96.2–99.4% agreement and replayed actions retain 97.2–99.8% agreement. Sensitivity concentrates near the decision boundary, where label agreement is lower. This is expected for a thresholded discretization and motivates the phase-agnostic continuous selector.

Table 11: **Sensitivity to the fixed drift threshold** -0.008**.** Ranges span the four expanded rollout configurations. The near-boundary subset contains rounds within 20% of the original threshold.

Change	All rounds		Near boundary	
	Label	Action	Label	Action
-20%	96.2–97.3%	97.4–98.3%	39.5–47.1%	54.1–71.1%
-10%	97.8–98.6%	98.5–99.2%	65.8–73.5%	78.4–86.8%
$+10\%$	98.1–99.4%	98.8–99.8%	67.6–89.5%	81.1–97.4%
$+20\%$	96.8–97.7%	97.2–98.9%	52.9–60.5%	61.4–81.6%

Three complete autoregressive shuffled-phase replicas cover all 80 prompts in each configuration. Shuffling changes action distributions (total-variation distance 0.033–0.113) and the recurrence-alignment operating point. The phase-based mapping often improves alignment; recurrence and D_2 comparisons are mixed. These results establish that labels affect actions without establishing the taxonomy's causal superiority. PhaseAgnosticPareto confirms that trajectory control remains effective without discrete labels.

C.7 ACTUATOR EFFECTS IN THE INITIAL CONTROLLER STUDY

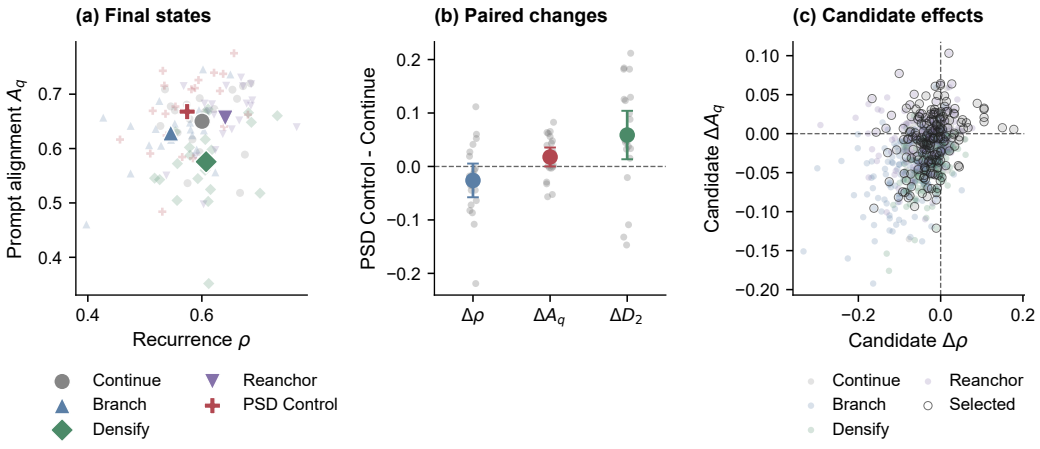

Figure A9: **Actuator effects in the initial controller study.** (a) Final states under fixed and phase-based policies form a recurrence–alignment frontier. (b) Paired differences from Continue compare policies on the same 20 prompts. (c) Candidate continuations are shown by their next-state displacements $(\Delta\rho, \Delta A_q)$; outlined points were selected. Large markers are exact method means; paired-effect bars show mean ± 1.96 prompt-level standard errors.

C.8 CANDIDATE-LEVEL MECHANISM

Figure A11 follows one response through the phase-based controller. The descriptors are recomputed after every committed segment, and the policy alternates between continuation and corrective actions as the prefix changes. The example illustrates online feedback rather than retrospective rescoring.

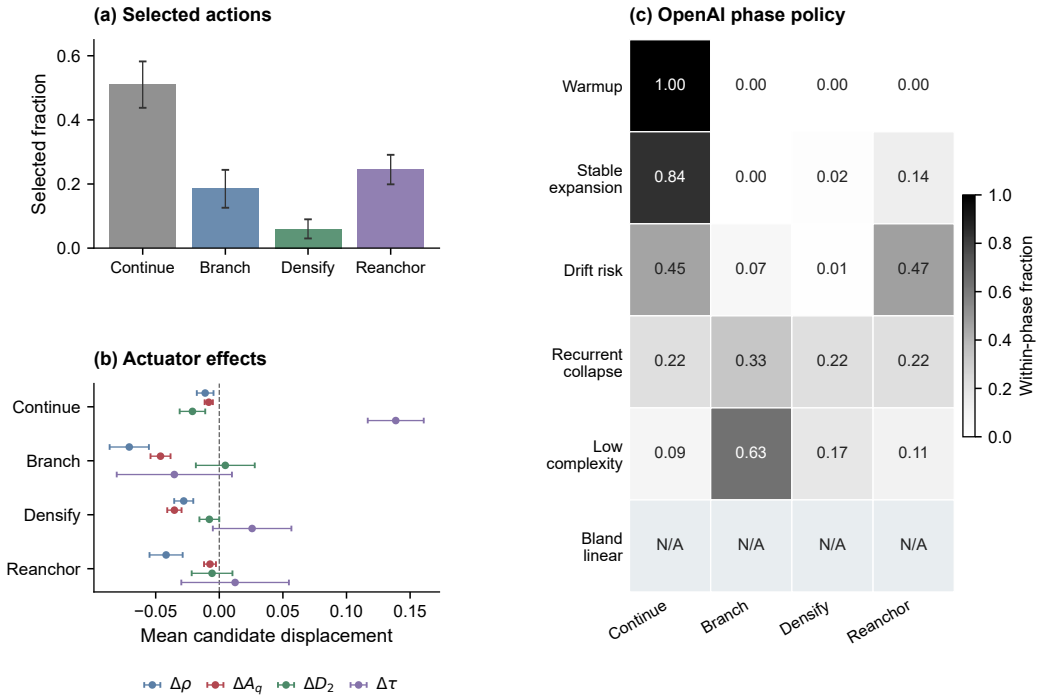

Figure A10: **State-dependent actuator selection in the initial 20-prompt study.** (a) Selected-action fractions summarize the use of each actuator. (b) Candidate displacement profiles show mean changes in ρ, A_q, D_2, and τ. (c) Selection fractions conditioned on the diagnosed phase relate actuator choice to the observed prefix state. All panels use the same OpenAI run. Action fractions use 200 committed rounds; displacement bars show mean ± 1.96 prompt-level standard errors. N/A denotes an unobserved phase.

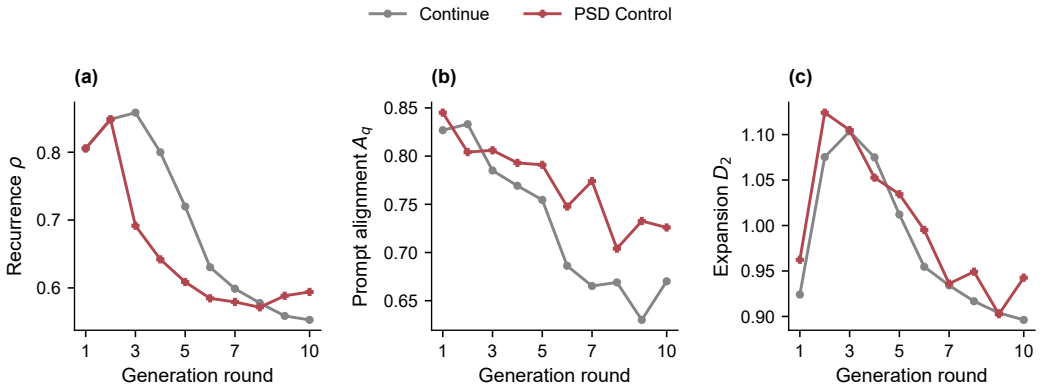

Figure A11: **Controller case study.** Round-wise descriptors for one prompt show how the observed cumulative-prefix trajectory and selected actuator evolve together.

C.9 OpenAI–Qwen controller replication and representation transfer

The open-representation analysis contains two complementary tests. First, the same 100 method-prompt trajectories are re-embedded with Qwen3-Embedding-8B, Qwen3-Embedding-4B, Qwen3-Embedding-0.6B, `nomic-embed-text-v1`, `bge-large-en-v1.5`, and `all-MiniLM-L12-v2`, including native and dimension-matched variants. Qwen3-Embedding-8B and Qwen3-Embedding-4B have the lowest joint replacement errors. Radius calibration reduces recurrence MAE from 0.154 to 0.079 for Qwen3-Embedding-8B and from 0.171 to 0.085 for Qwen3-Embedding-4B. Weaker representations show larger distortions or saturation.

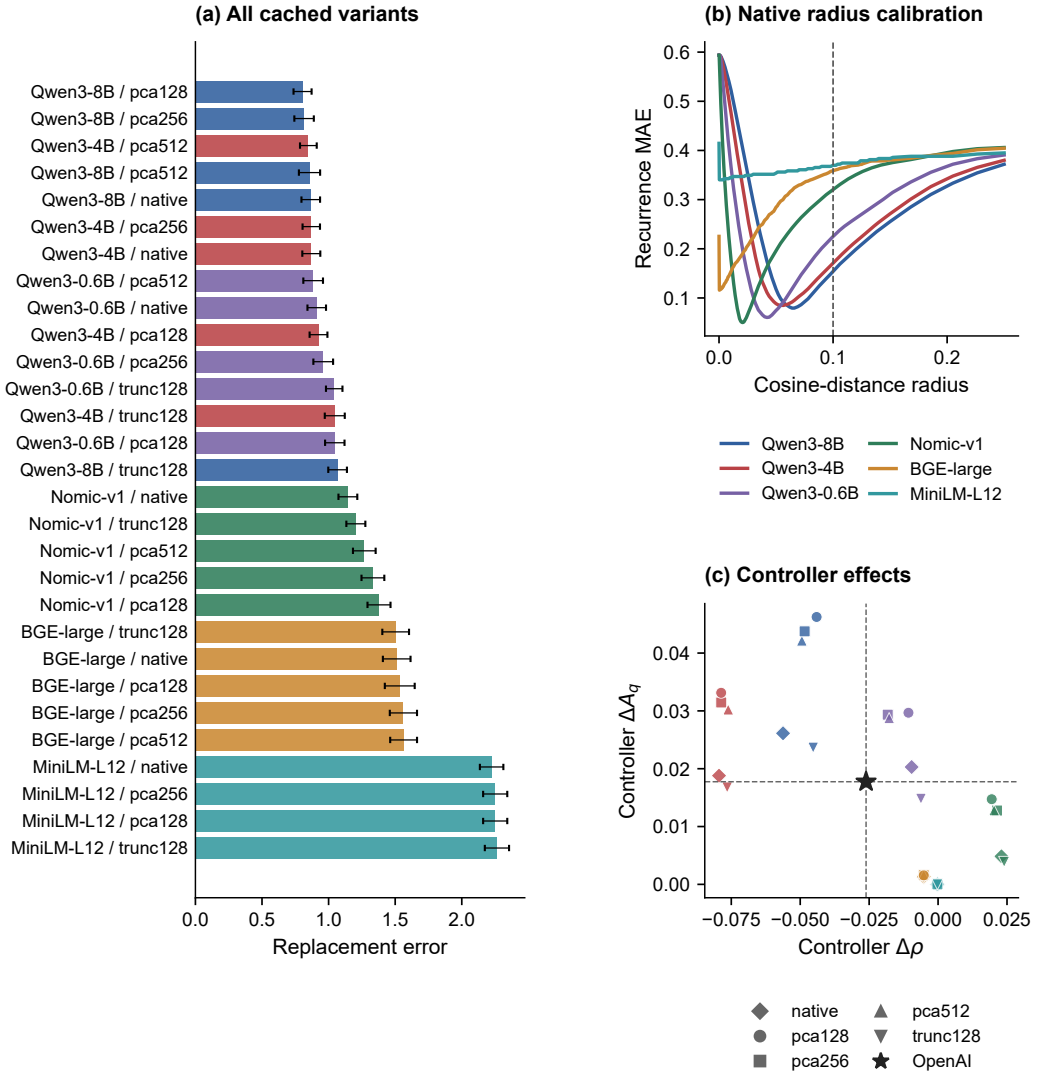

Figure A12: **Open-representation validation.** (a) All 29 cached model/dimensional variants ranked by descriptor replacement error against OpenAI; bars show mean ±1.96 SEM. (b) Native-variant recurrence MAE for all six backends across radii, complementing the full variant curves in main Figure 6b. (c) All finite cached controller-effect points, with colors identifying backends and symbols identifying dimensional variants.

For Qwen3-Embedding-8B, trajectory-level Spearman correlations are 0.685 for D_2, 0.476 for ρ, 0.444 for A_q, and 0.312 for τ; Kendall agreement across the five policies is 0.8, 1.0, 0.8, and 0.2, respectively.

Second, the controller is rerun with Qwen3-Embedding-8B. Figure A13 compares paired effects with the OpenAI reference. Recurrence, alignment, and D_2 retain their direction. Tortuosity changes direction across these two controller runs, identifying it as the least portable numerical descriptor in the 20-prompt experiment.

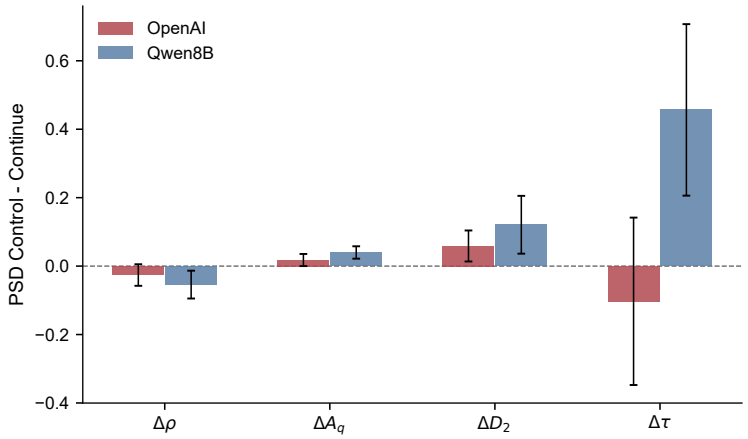

Figure A13: **OpenAI–Qwen paired controller effects.** The recurrence, prompt-alignment, and D_2 directions reproduce under Qwen3-Embedding-8B; the tortuosity effect is backend-sensitive.

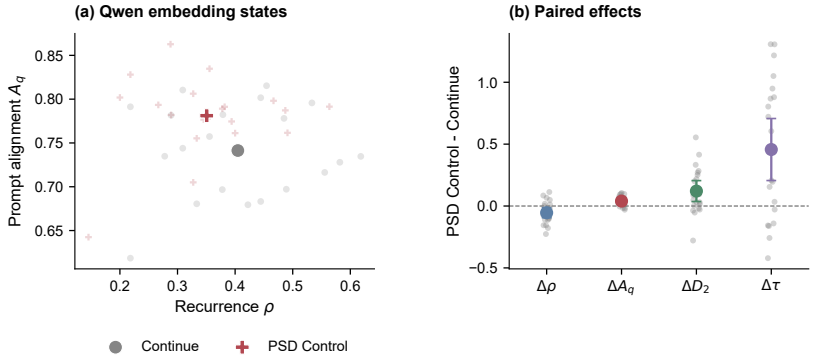

Figure A14: **Qwen controller validation.** An open representation reproduces a recurrence-alignment frontier while requiring its own numerical radius.

The expanded Qwen-Harrier matrix reinforces the same distinction. Re-evaluating each rollout in both representations preserves the PSD Control versus Continue direction for all four descriptors in all 16 paired representation comparisons. Prompt-method Spearman agreement ranges from 0.345 to 0.665, and eight-method Kendall ordering ranges from 0.500 to 1.000. These are meaningful but incomplete agreements. PSD is portable as a prefix-only protocol; descriptor values, radii, magnitudes, and some policy rankings remain properties of the chosen representation.

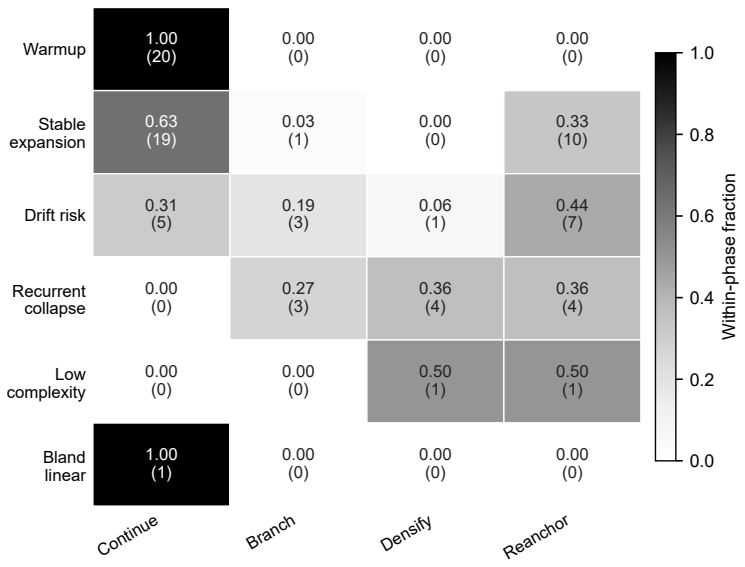

Figure A15: **Qwen phase-policy matrix.** Cells show phase-conditioned actuator frequencies under the Qwen controller run.

C.10 Runtime and deployment interpretation

PSD Control does not require a person to interrupt generation. The four candidate calls are issued in parallel, their extended prefix states are scored, and one segment is committed. Table 12 reports the measured wall-clock cost for ten rounds. With the 122B generator and Qwen representation, median wall time is 130.4 seconds for PSD Control and 97.7 seconds for Continue, a ratio of 1.33. The ratio measures latency under parallel execution, not total inference compute. Monitoring without candidate generation requires only cumulative-prefix embedding and is substantially cheaper.

Table 12: **Measured wall-clock time for a complete ten-round response.** Candidate generation is parallelized across four workers. Times include generation and online representation calls; they do not imply equal total inference compute.

Generator	State	Continue (s)		PSD Control (s)	
		p50	p95	p50	p95
9B	Qwen	29.3	31.0	39.9	42.2
9B	Harrier	59.7	77.7	121.1	139.2
122B	Qwen	97.7	128.9	130.4	236.4
122B	Harrier	126.7	132.4	227.0	235.7

C.11 Broader impacts

Trajectory-level monitoring can help identify repetition, drift, and loss of specificity before a long response is complete, and can support more reliable human oversight of long-form systems. The same feedback mechanism could also be used to sustain prompt adherence in misleading or otherwise harmful generation. Representation-dependent thresholds may behave differently across domains, languages, and writing styles, creating false alarms or missed failures if deployed without calibration. High-stakes use should therefore calibrate the representation on the target domain, retain the generated trajectory for audit, and combine PSD with independent factuality, safety, and fairness evaluations. This work releases a measurement and prompt-selection framework, not a deployed decision system or a new generative model.