RAMBLING ON REPEAT? EXACT FIRST-EXIT LAWS AND MINIMAL INTERVENTION FOR VERBATIM LLM LOOPS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern LLM generations occasionally enter verbatim repetition and run to their output caps, wasting user token budgets, increasing request latency, exhausting computing power, and exposing a system vulnerability. Existing repetition penalties and n-gram bans are global heuristics that distort healthy text but neither estimate a detected loop's persistence nor certify exit. We combine serving-stack signals and rigorous first-exit theory to quantify and control this behavior with guarantees on minimum intervention cost. Temperature-T decoding samples the Boltzmann distribution over token energies, making the runner-up gap Δ an activation energy. Hazard products along one greedy trace give the exact finite-horizon distribution of exit time and phase at any temperature. If a measured lap repeats indefinitely, escape obeys an Arrhenius law governed by the cycle's *minimum* gap, with exits concentrating at that weakest phase. The parameter-free law is validated across 37 orbits from six families up to a 284B MoE model including its 3-bit quantization. Inverting it yields two training-free, event-triggered controllers for a user-specified risk tolerance δ: a provably minimal-temperature pulse that exits within a chosen horizon, and a stepwise minimum-KL guard that contains every detected trajectory within L tokens at minimum compatible distortion, each with successful-exit probability at least $1 - \delta$. Deployed as a vLLM logits processor, the untuned guard raises Qwen3-4B sustained escape from 0.71 to 0.91 and reduces cap exceedance on 710 matched natural-workload prefixes from 30.7% to 5.1%.

1 INTRODUCTION

Most language-model responses terminate normally, but a decoder caught in a verbatim loop can repeat until a hard output cap. This classical failure of open-ended generation is self-reinforcing (Holtzman et al., 2019; Xu et al., 2022; Li et al., 2023). It is a minority event on the modern models in our census, yet recent work still finds repetition–quality trade-offs across open LLMs (Huang et al., 2025), while non-halting and inference-cost attacks expose the resulting token, latency, and compute waste (Hammouri et al., 2025; Dong et al., 2025). We therefore treat exact repetition as a low-rate but high-cost runtime failure.

That rarity is mismatched to global remedies: truncation, increased temperature, repetition penalties, and n-gram bans also alter healthy text and legitimate repetition (Fan et al., 2018; Holtzman et al., 2019; Hewitt et al., 2022; Meister et al., 2023; Keskar et al., 2019). They provide global knobs, not event-level estimates of how long the current loop is likely to persist, where it is likely to exit, or the minimum intervention needed to reduce its persistence risk to a prescribed level.

Operational risk is jointly determined by workload-dependent capture incidence and persistence after capture. Our experiments report the former, while our theory and controllers address the latter. The key observation is that $\mathrm{softmax}(z/T)$ is precisely the Boltzmann distribution over token energies $-z_i$. Generation can therefore be viewed as a stochastic dynamical system whose state is the growing context and whose observable coordinates are the logits: greedy decoding is its zero-temperature flow, verbatim repetition is a symbolic periodic orbit, and temperature sampling introduces state-dependent noise. Open inference exposes this local energy landscape, allowing a server to quantify and constrain a loop's persistence once a matcher confirms repetition, without retraining or perturbing unrelated

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.

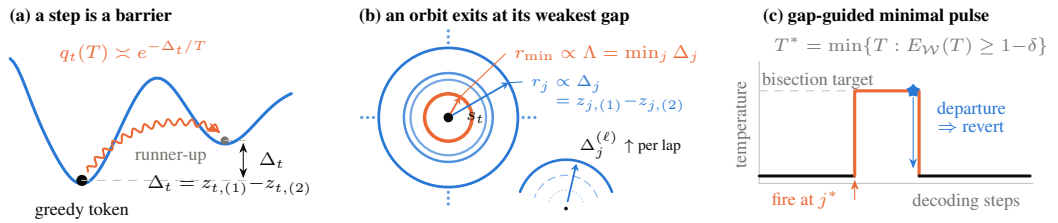


Figure 1: The one-shot exit picture: (a) each gap Δ_t is an activation energy crossed with hazard $q_t(T) \asymp e^{-\Delta_t/T}$ (Thm. 1); (b) concentric rings at various radii encode the P phase gaps (dots omit further phases), and the highlighted $\Lambda = \min_j \Delta_j$ sets the Arrhenius exit scale and phase (Thm. 2); the inset schematizes a generic phase radius expanding lap over lap, i.e., gap-profile deepening before the measured saturation (Sec. 4.1); (c) the pulse fires the minimal temperature at the weakest phase and reverts on departure (Prop. 1).

tokens. Our guarantee is to break each detected repetition trajectory within a prescribed budget at the minimum distortion compatible with the constraint.

We make three major contributions in this work: a conditional law, a cross-model regularity, and a control method derived from that law, as follows:

- **An exact first-exit theory for autoregressive decoding.** Non-asymptotic bounds identify the runner-up gap as a per-step activation energy (Theorem 1); hazard products turn one finite greedy trace into exact finite-horizon survival and exit-location probabilities at every T. Repeating one recorded period indefinitely makes exit time Arrhenius with $\Lambda = \min_j \Delta_j$ and localization at the argmin phase (Theorem 2); hazard summability separates positive non-exit mass from almost-sure departure (Theorem 3), with measured saturation supporting long-lived metastability over the observed horizon.
- **Self-reinforcing persistence after capture: a cross-model regularity.** Repeated traversal raises the gap profile (100% of $\geq$3-lap orbits on GPT-2, median 0.58 nats/lap; positive in 8/9 eligible scaled configurations), deepening the measured escape barrier until saturation. Capture is a minority event on the modern post-trained panel (0%–21% vs GPT-2's 64%), yet every fitted orbit obeys the same Arrhenius exit law, including at 3-bit quantization. Note this contribution concerns persistence conditional on capture while leaving mechanisms of first entry open.
- **A minimum-intervention runtime assurance layer, derived from the law.** Training-free closed forms trigger only on detected risk: a provably minimal-temperature pulse for a detected orbit (Proposition 1), and a minimum-KL guard capping every detected verbatim repetition trajectory at L tokens with probability $1-\delta$ (Proposition 2), shipped as a vLLM logits processor. The primitives trace the low-intervention frontier on GPT-2; on Qwen3-4B at T=0.8, the guard transfers untuned, raising sustained escape 0.71→0.91 (exact McNemar $p = 3.8 \times 10^{-6}$; recapture 0.74→0.69).

2 Related Work

Degeneration, non-halting, and loop guards. Loops under maximization decoding (Holtzman et al., 2019) are linked to Markov inflow, autoregressive reinforcement, training-data repetition, and specialized features (Fu et al., 2021; Xu et al., 2022; Li et al., 2023; Hiraoka & Inui, 2025). Mitigations use truncation, training objectives, or penalties (Fan et al., 2018; Welleck et al., 2019; Keskar et al., 2019), but global penalties trade repetition against quality (Huang et al., 2025). Runtime work studies non-halting fixed points (Hammouri et al., 2025), inference-cost attacks (Dong et al., 2025), and KV-cache guards (Xu et al., 2026). We instead condition on an exact detected re-traversal and next-token distributions to quantify persistence and solve its minimum-KL tail constraint, excluding semantic loops and non-repetitive length attacks.

Controlled decoding and adaptive samplers. Learned controllers use KL-regularized value functions, per-token temperatures, or temperature and top-p schedules (Mudgal et al., 2023; Dhuliawala et al., 2024; Wang et al., 2026), incurring distortion while active. Our detector-triggered closed forms require no training, use one greedy trace, and certify window escape and controlled tails.

Dynamical systems and thermal escape. Language-model dynamics work studies attention, in-context learning, belief states, and entropy drift (Geshkovski et al., 2025; Von Oswald et al., 2023; Shai et al., 2024; Braverman et al., 2020). We study the induced closed loop at the exposed logit–temperature interface. Discretizing classical thermal escape (Kramers, 1940; Hänggi et al., 1990; Freidlin & Wentzell, 2012) makes activation energies, exit laws, and exit localization measurable from logits. Appendix J relates the case study to hallucination detection (Manakul et al., 2023; Farquhar et al., 2024; Ji et al., 2023).

3 Mathematical modeling and implementation

We first model the decoder a stochastic dynamical system whose energy landscape is its logits, then derive the exit laws that price a detected loop and apply those laws to the two controllers the paper ships, finally we close this section with concrete implementation. All mathematical proofs are provided in Appendix A unless stated otherwise.

3.1 LLM decoding as a stochastic dynamical system

Figure 1 summarizes the picture. Let $\mathcal{V}$ be the token vocabulary, with $V = |\mathcal{V}|$. An autoregressive model maps the growing context $x_{<t}$ to logits $z_t \in \mathbb{R}^V$. At temperature $T > 0$, its next-token pickup law is

$$\pi_{t,i}^{(T)} := \Pr_T(W_t = i \mid x_{<t}) = \frac{e^{z_{t,i}/T}}{\sum_{j\in\mathcal{V}} e^{z_{t,j}/T}}, \qquad i \in \mathcal{V}, \tag{1}$$

where W_t is the random next token and a lowercase w_t denotes its realized value. We write $\pi_t := \pi_t^{(1)}$ for the nominal predictive distribution ($T = 1$) and $g_t := \arg\max_i z_{t,i}$ for the deterministic greedy token (ties are resolved by the backend's fixed rule), therefore greedy decoding is the $T \to 0^+$ limit when the maximizer is unique.

The exact Markov state is $x_{<t}$ (equivalently, history plus its KV-cache), while the nominal law π_t is only an observable readout of that state. We compare readouts by the Hellinger distance $d_{\mathrm{Hel}}(\pi_t, \pi_s) := \frac{1}{\sqrt{2}}\|\sqrt{\pi_t} - \sqrt{\pi_s}\|_2$. As two distinct contexts may induce the same predictive distribution π_t while leading to different subsequent generations, recurrence of the observable readouts does not imply recurrence of the underlying state. Therefore we define recurrence at the level of the emitted greedy trace.

A greedy trace (g_t) is said to form a *symbolic cycle* of period P with onset t_0 if (throughout the observation horizon):

$$g_{t+P} = g_t, \qquad \forall\, t \geq t_0.$$

In such a cycle, the emitted token symbols repeat periodically, even though the underlying full context never returns exactly to a previous state. Moreover, the logit profiles observed at corresponding phases may drift from one lap to the next. We refer to this output-periodic object as an *orbit*. The theory developed below studies two scalar quantities along its trace: the runner-up gap Δ_t and the departure hazard $q_t(T)$.

Definition 1 (Gap and hazard)**.** At step t with a unique greedy token g_t, the *runner-up gap* and *departure hazard* are defined as

$$\Delta_t := z_{t,g_t} - \max_{i\neq g_t} z_{t,i}, \qquad q_t(T) := \Pr_T(W_t \neq g_t \mid x_{<t}) = 1 - \pi_{t,g_t}^{(T)}. \tag{2}$$

For control, ν_t denotes the modified next-token distribution.

Lemma 1 (Monotone thermal noise)**.** *For fixed non-constant logits with a unique maximum, $T \mapsto q(T)$ is strictly increasing on $(0,\infty)$, with $q(T) \to 0$ as $T \to 0^+$ and $q(T) \to 1 - 1/V$ as $T \to \infty$.*

3.2 The Arrhenius law

Theorem 1 (Discrete Arrhenius law[1])**.** *For every $T > 0$,*

$$\tfrac{1}{2}\, e^{-\Delta/T} \;\leq\; q(T) \;\leq\; (V-1)\, e^{-\Delta/T}, \tag{3}$$

[1]Throughout the paper, we omit the step index t from the gap Δ_t and hazard $q_t(T)$ whenever discussion of a particular step is irrelavent.

and consequently $\lim_{T\to 0^+} T \ln q(T) = -\Delta$.

It is not hard to see from reaction-rate theory (Hänggi et al., 1990) that the runner-up gap is exactly the activation energy of the step: the log hazard is linear in inverse temperature with slope $-\Delta$ (up to a bounded prefactor). And unlike activation energies in physics, Δ here is directly observable at every step from the logits.

3.3 THE EXIT PROBLEM FOR PERIODIC ORBITS

Choose a reference index t_{ref} on a detected period-P greedy orbit and save its token indices $x_{<t_{\text{ref}}}$. From this prefix, greedily record H reference steps $g_1, \ldots, g_H$ and their full-logit rows, using local indices $t = 1, \ldots, H$. By sampling from the same saved prefix with temperature-T produces random tokens W_t. Define the first-exit time $\tau := \inf\{t \geq 1 : W_t \neq g_t\}$ and survival function $S_T(t) := \Pr_T(\tau > t)$. Until τ, the sampled and greedy-reference contexts coincide at every corresponding step, so the hazards along the stored trace are deterministic and precomputable.

Theorem 2 (Exit law)**.** *Let $q_1(T), \ldots, q_H(T)$ be the stepwise hazards along the stored reference trace.* ***(i) Survival identity.***

$$S_T(t) = \prod_{s=1}^{t} \big(1 - q_s(T)\big) \quad \textit{for every stored step } t. \tag{4}$$

This is exact up to the stored horizon. ***(ii) Arrhenius exit time.*** *The* frozen-profile periodic continuation *repeats one recorded lap indefinitely: $q_{kP+j}(T) = q_j(T)$ and $\Delta_{kP+j} = \Delta_j$ for $k \geq 0$ and $1 \leq j \leq P$. Define the within-lap survival factors $A_0(T) := 1$ and $A_r(T) := \prod_{j=1}^{r}(1 - q_j(T))$ for $1 \leq r \leq P$, so $\rho(T) := 1 - A_P(T)$ is the per-lap exit probability. Then*

$$\mathbb{E}_T[\tau] = \frac{\sum_{r=0}^{P-1} A_r(T)}{\rho(T)} = \frac{\sum_{r=0}^{P-1} \prod_{j=1}^{r}(1 - q_j(T))}{\rho(T)}, \tag{5}$$

and consequently $P\,(1 - \rho(T))/\rho(T) \leq \mathbb{E}_T[\tau] \leq P/\rho(T)$; $\lim_{T\to 0^+} T \ln \mathbb{E}_T[\tau] = \Lambda := \min_{j \leq P} \Delta_j$. ***(iii) Exit localization.*** *Let $J \in \{1, \ldots, P\}$ be the phase at which an exit occurs within a lap. If $j^* := \arg\min_j \Delta_j$ is unique, then*

$$\Pr_T(J = j \mid \textit{exit in a lap}) = \frac{q_j(T) \prod_{l<j}(1 - q_l(T))}{\rho(T)}, \qquad \Pr_T(J = j^* \mid \textit{exit}) \to 1 \quad (T \to 0^+). \tag{6}$$

The activation energy of an orbit is its *weakest* gap: escape statistics are controlled by the single most uncertain step of the repeated span. Part (i) gives exact survival and exit-location probabilities from the stored full-logit rows up to the recorded horizon; an expected exit time additionally requires an infinite trace or the frozen-periodic continuation in part (ii). When only the top-2 gap is stored, Equation 3 still bounds the hazard and predicts its scale. Part (iii) predicts which phase breaks first, a microprediction tested in Section 4.2.

3.4 SELF-REINFORCING REPETITION AND METASTABLE PERSISTENCE

Prior work shows that repeated context can reinforce continued repetition (Chiang & Chen, 2021; Xu et al., 2022). Our census exhibits the corresponding finite-window signature: lap-mean gaps rise for four to six traversals and then plateau (Section 4.1). Note this does not imply permanent trapping since positive non-exit probability requires the sum of all future departure hazards to remain finite. The following result converts this condition into indefinitely growing and uniformly bounded gap regimes.

Theorem 3 (Hazard summability and metastable persistence)**.** *Using the same local indexing from t_{ref}, fix $T > 0$ and suppose the greedy-reference trace is defined for all $t \geq 1$. Then*

$$\Pr_T(\tau = \infty) = \prod_{t\geq 1}(1 - q_t(T)), \qquad \Pr_T(\tau = \infty) > 0 \iff \sum_{t\geq 1} q_t(T) < \infty. \tag{7}$$

(i) Permanent deepening leaves non-exit mass. *For any $\Delta_0, b_\Delta > 0$ and any detected period P, the following implication holds:*

$$\begin{aligned} &\Delta_t \geq \Delta_0 + b_\Delta \lfloor (t-1)/P \rfloor, \quad t \geq 1, \\ \implies \quad &0 < \Pr_T(\tau = \infty), \qquad \Pr_T(\tau < \infty) \leq \min\left\{1, (V-1)P \frac{e^{-\Delta_0/T}}{1 - e^{-b_\Delta/T}}\right\}. \end{aligned} \tag{8}$$

(ii) Uniformly bounded gaps give almost-sure but potentially slow departure. *Suppose instead that $0 < \Lambda_{\mathrm{sat}} \leq \Delta_t \leq M < \infty$ for all $t \geq 1$, and define $\varepsilon_T := \frac{1}{2} e^{-M/T}$ and $\bar{q}_T := \sup_{t \geq 1} q_t(T) \leq (V-1) e^{-\Lambda_{\mathrm{sat}}/T}$. Then $S_T(t) \leq (1 - \varepsilon_T)^t$, hence $\Pr_T(\tau < \infty) = 1$, and for every horizon $H \geq 1$,*

$$\begin{aligned} &\frac{e^{\Lambda_{\mathrm{sat}}/T}}{V-1} \leq \frac{1}{\bar{q}_T} \leq \mathbb{E}_T[\tau] \leq \frac{1}{\varepsilon_T} = 2e^{M/T}, \\ &\Pr_T(\tau \leq H) \leq \min\{1, H\bar{q}_T\}. \end{aligned} \tag{9}$$

Part (i) of the theorem is a conditional infinite-horizon result: observing gap growth over only four to six laps is insufficient to establish that its premise holds indefinitely. By contrast, the measured gap plateaus are more consistent with part (ii), although no finite trace can rigorously establish its behavior over an infinite horizon. When the gaps are uniformly bounded, the model eventually departs from the cycle with probability one at every fixed $T > 0$; nevertheless, its mean waiting time remains bracketed between two Arrhenius scales. Moreover, if $H\bar{q}_T \ll 1$, the finite-horizon bound shows that passive decoding is unlikely to leave the cycle within an H-token budget. Thus, the operational quantity of interest is not whether the model eventually departs, but how long it remains in the cycle. By the time the cycle is detected, the preceding transient self-reinforcement may already have pushed this residence time beyond the available token budget. This motivates event-triggered control rather than passively waiting for the model to depart on its own.

3.5 The minimal thermal pulse controller

Proposition 1 (Gap-guided minimal pulse)**.** *Consider a stored k-step window $\mathcal{W}$ and the single-block intervention that raises its temperature to T and then reverts. Define*

$$E_{\mathcal{W}}(T) := 1 - \prod_{j \in \mathcal{W}} (1 - q_j(T)), \qquad T^*_{\mathcal{W}} := \inf\{T > 0 : E_{\mathcal{W}}(T) \geq 1 - \delta\}. \tag{10}$$

(i) *$E_{\mathcal{W}}(T)$ is strictly increasing with $\lim_{T \to \infty} E_{\mathcal{W}}(T) = 1 - V^{-k}$, so $T^*_{\mathcal{W}}$ is finite and computable by bisection whenever $\delta > V^{-k}$; otherwise it is infinite.* ***(ii)*** *With $\widetilde{\Delta}_{\mathcal{W}} := \max_{j \in \mathcal{W}} \Delta_j$, the schedule $\widehat{T}_{\mathcal{W}} = \widetilde{\Delta}_{\mathcal{W}} / \ln\big(k/(2\ln(1/\delta))\big)$ (for $k > 2\ln(1/\delta)$) guarantees $E_{\mathcal{W}}(\widehat{T}_{\mathcal{W}}) \geq 1 - \delta$.* ***(iii)*** *Among the P cyclic placements, minimizing $\widetilde{\Delta}_{\mathcal{W}}$ minimizes this conservative sufficient temperature.*

Monotonicity makes $T^*_{\mathcal{W}}$ computable by binary search (bisection), so the implementation applies this minimal temperature rather than the conservative closed-form schedule (Section 3.7). Note both statements are exact only for the stored frozen profile. If upcoming gaps drift upward, the stored-lap plan can be optimistic. To retain a deterministic guarantee under such circumstance, the bisection must use a certified phase-wise upper envelope of the future gaps.

3.6 The minimum-KL guard controller

This type of controller replaces the pulse's temperature search by the I-projection of base law μ onto a convex exit-mass constraint (Csiszár, 1975). Under the exit-versus-stay coarse-graining $i \mapsto \mathbf{1}\{i \notin \mathcal{C}\}$, the KL chain rule reduces the least intervention information to a Bernoulli mismatch.

Proposition 2 (Information projection and controlled-trajectory certificate)**.** ***(i) I-projection per step.*** *Fix μ, stay set $\mathcal{C}$, $p := \mu(\mathcal{C}^c) \in (0,1)$, target exit rate $r \in (0,1)$, and $\mathcal{F}_r := \{\nu : \nu(\mathcal{C}^c) \geq r\}$. Its I-projection $\nu^* := \arg\min_{\nu \in \mathcal{F}_r} D_{\mathrm{KL}}(\nu \| \mu)$ is the certified no-op μ if $p \geq r$, otherwise*

$$\nu^*_i = \mu_i \frac{1-r}{1-p} \, (i \in \mathcal{C}), \quad \nu^*_i = \mu_i \frac{r}{p} \, (i \notin \mathcal{C}), \quad D_{\mathrm{KL}}(\nu^* \| \mu) = d_2(r \| p), \tag{11}$$

where $d_2(a\|p) := a\ln(a/p) + (1-a)\ln((1-a)/(1-p))$ is the binary KL divergence. For rare base exits this excess binary codelength is $r\Delta_{\mathrm{eff}} - H_2(r) + O(p)$, where H_2 is binary entropy and the

event self-information is $\Delta_{\text{eff}} := -\ln p$. ***(ii) Sequential composition.*** *Let* ℓ_t *be the current longest previously occurring suffix length (overlaps allowed),* $\ell_{\min}$ *the arming threshold,* $\mathcal{C}_t$ *the next-token set keeping* $\ell_{t+1} \geq \ell_{\min}$*, and* ℓ_{traj} *the terminal length of the detected repetition trajectory under control. For trajectory cap* $L > \ell_{\min}$ *and per-trajectory failure budget* $\delta \in (0, 1)$*, set* $\mathsf{D} := \{\ell_{\text{traj}} \geq \ell_{\min}\}$ *and* $r_{L,\delta} := 1 - \delta^{1/(L-\ell_{\min})}$*. Applying (i) at every matcher-live step gives (for* $j \geq 0$*):*

$$\Pr(\ell_{\text{traj}} \geq \ell_{\min} + j \mid \mathsf{D}) \leq (1 - r_{L,\delta})^j, \quad \Pr(\ell_{\text{traj}} \geq L \mid \mathsf{D}) \leq \delta. \tag{12}$$

This projection preserves the conditional token law within both the exit and stay sets, so all cost is binary mismatch. Each live step adds at least $-\ln(1 - r_{L,\delta})$ nats of survival surprisal and $L - \ell_{\min}$ steps reach $\ln(1/\delta)$ hence yeilds the certificate. The per-trajectory certificate re-arms after recapture and incurs the minimum possible KL cost only when the constraint is binding.

3.7 Implementation

Runtime library. We implement the pulse and guard controllers as a shared, backend-agnostic runtime library and expose both controllers through a pip-installable vLLM adapter. An incremental exact trailing-lag matcher detects a qualifying re-traversal and maintains its phase online. The pulse controller places a finite intervention window over the weakest recorded phases, bisects for the minimum temperature $T^*_{\mathcal{W}}$ that attains exit probability $1 - \delta$, then restores the base policy after the first departure. On the other hand, the guard controller applies the minimum-KL binary projection at every live step and re-arms after recapture. Both controllers operate as per-request logits processors and they leave the sampling law unchanged when no qualifying repetition is detected.

Observable requirements and backends. Top-2 logits suffice for the gap law and its bounds; exact temperature-reweighted hazards and survival require the full logit row, while target-string costs require the target token's log-probability. We realize these interfaces through Hugging Face full-logit inference, `llama.cpp` top-k accounting, and the vLLM adapter. Appendix B specifies the software and determinism stack; Appendix I records the model, precision, resource requirements, and preflight details.

4 Experiments and Result Analysis

This section follows repetition from capture to containment. E1 measures greedy-orbit capture and self-reinforcement; E2 tests the parameter-free first-exit law; E3a and E3b evaluate one-shot escape and sustained runtime control; E4 estimates natural-workload incidence and evaluates the guard from matched first-risk prefixes; finally, Section 4.6 tests transfer across model families, scales, and quantization levels.

4.1 E1: Attractor census

Protocol. Input 100 prompts in ten equal groups (news, dialogue, essay, how-to, fiction, review, encyclopedic, technical, general-purpose, academic/research; Appendix C), greedy decoding for 120 new tokens, exact cycle detection with $P \leq 24$; prompts without a short-scope orbit are re-decoded to 200 tokens with $P \leq 48$.

Results. At the primary scope ($P \leq 24$), 64/100 greedy trajectories (64%) are captured within 120 tokens. Extending the remaining cases to 200 tokens and $P \leq 48$ raises descriptive coverage to 85/100 (85%), chiefly through sentence-length orbits ($25 \leq P \leq 48$; Figure 3a,b in Appendix C). Among the primary-scope orbits, all 62 with at least three complete laps show rising lap-mean runner-up gaps (median 0.58 nats/lap; Figure 3c), while the one-period Hellinger distance contracts in every estimable orbit (median $\hat{\lambda} = -0.052$ per step). Both trends plateau after around 4 to 6 laps, indicating transient basin deepening consistent with Theorem 3's bounded-gap rather than permanent-deepening.

4.2 E2: Arrhenius first-exit law

Protocol. Select 4 captured orbits with distinct periods ($P \in \{5, 7, 10, 16\}$) from E1 and choose t_{ref} three complete laps after detected onset and save the prefix. One forward pass over the ensuing greedy reference trace yields the gap profile and exact hazards $q_t(T)$ at every temperature directly

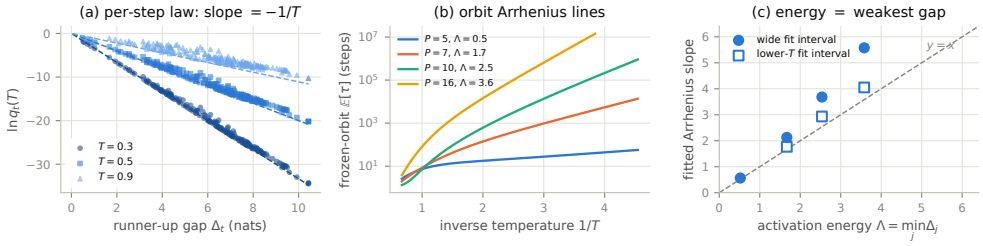


Figure 2: The Arrhenius law on GPT-2, from one greedy trace per orbit. (a) Per-step hazards versus runner-up gap at three temperatures; dashed lines are the parameter-free slopes $-1/T$. (b) Frozen-profile expected exit time on the temperature grid. (c) Fitted slope against activation energy Λ over wide and lower-temperature fit intervals.

from the stored logit rows, while Theorem 1 predicts their scale. Based on these we compute (a) the per-step law across all trace steps; (b) frozen-profile Arrhenius curves $\mathbb{E}_T[\tau]$ using Equation 5, with slopes fitted on a wide interval ($\mathbb{E}_T[\tau] \in [6, 10^5]$) and a lower-temperature interval ($[2 \times 10^2, 10^8]$); (c) exact time-inhomogeneous survival curves $S_T(t)$. Independently, we resampled 48 trajectories per temperature at $T \in \{0.3, 0.5, 0.7, 0.9, 1.2\}$ (horizon 48, then 30 post-exit steps) from the same saved prefix and provide the empirical check. Determinism of the forward pass is asserted during inference, thus pre-exit sampling reduces to the hazard sequence along the trace (Theorem 2i).

Results. Figure 2 connects the stepwise and orbit-level tests. Across ten nats of gap and a 30-nat range of $\ln q$, measured hazards follow the parameter-free lines $\ln q = -\Delta/T$ (panel a); the upward bend at high temperature is the subleading-competitor prefactor allowed by Equation 3.

Aggregating these hazards over each frozen lap gives mean exit times that grow approximately exponentially in $1/T$ (panel b). Fitting $\ln \mathbb{E}_T[\tau]$ against $1/T$, the resulting slopes are 1.09–1.56$\times\Lambda$ on the wide interval and 1.06–1.16$\times\Lambda$ in the lower-temperature fit (panel c). The lower-temperature estimates move toward the predicted $y = x$ line, whereas replacing Λ with the mean gap overpredicts the slopes severalfold. Thus the weakest gap, not the average, sets the orbit's activation energy.

Independent Monte Carlo checks test predictions not shown in Figure 2. At $T \leq 0.7$, 115/126 uncensored exits (91%) occur at the predicted weakest-gap phase, versus a 13% phase-uniform baseline, showing that the gap profile predicts *where* escape occurs as well as its temperature scale. Across 20 conditions with 48 trajectories each, analytic and empirical survival agree with a median sup-deviation of 0.05 (worst: 0.14); at $t = 48$, analytic survival also matches the observed censored fraction. Thus, the product law captures the full finite-horizon waiting-time distribution, while horizon non-exits remain right-censored survival mass rather than evidence of permanent trapping (full diagnostics are shown in Appendix H, Figure 6). Finally, the fitted weakest-gap growth $b_\Delta \in [0.05, 0.57]$ nats/lap over the early laps reduces the Arrhenius hazard scale by $e^{-b_\Delta/T}$ per lap. At $T = 0.5$, the lap-1-to-5 factor ranges from 0.67 at $b_\Delta = 0.05$ to 0.010 at $b_\Delta = 0.57$, a 1.5–95$\times$ reduction that explains the remarkable finite-budget stabilization and, together with the observed plateau after four to six laps, supports the bounded-gap metastable branch of Theorem 3(ii).

4.3 E3a: One-shot escape on the intervention-cost frontier

Protocol and results. We decode ten archived early captures fixed by the E3 protocol for 128 tokens under 11 configurations. *Baselines:* untreated greedy, global temperature T=0.7, repetition penalty θ=1.3 (Keskar et al., 2019), and no-repeat-4-gram. *Ours:* three pulse and four windowed minimum-KL projection variants (Appendix D). Figure 4 and Table 3 compare 4-gram repetition with expected greedy departures, intervention KL, and base-model surprisal. Greedy is severely repetitive (seq-rep-4 0.90). No-repeat-4-gram and repetition penalty drive this metric to zero but change 18.0 and 74.6 tokens (60.8 and 269 nats); global temperature reaches 0.11 at 52.1 expected departures and 2.68 bits. The pulse exits detected orbits with 2.6 departures (median $T^*_{\mathcal{W}}$=1.32), reducing seq-rep-4 to 0.66 at 0.50 bits above greedy. No method wins on both repetition and intervention, but the matched pulse/minimum-KL settings fill a low-intervention regime absent from the repetition-reducing baselines: they require only 3.0/3.1 expected departures, versus at least 18.0. At the same detector and 90% window-exit guarantee, minimum-KL cuts cost from the pulse's 4.86 to 1.59 nats (3.1$\times$; global: 25.8). Therefore, the roles of two proposed controllers are complementary:

pulse provides minimum-temperature one-shot escape, while guard uses minimal-KL steps for sustainable containment, explained in the following section.

4.4 E3B: SUSTAINED CONTAINMENT AND DEPLOYMENT-SCALE CONTROL

Windowed escape has a predictable ceiling: escape–revert–recaptured again by the same or a neighbor basin. The *guard* of Proposition 2 instead acts whenever a period-agnostic re-traversal matcher is live. Over 512 GPT-2 tokens, seq-rep-4 is 0.61 at 26.3 nats—11$\times$ less than an n-gram ban and 3$\times$ less than global temperature (Table 4). The archived repeated-stretch statistics use a different matcher and are descriptive, not a direct certificate test (Appendix E). Operationally, the guard is a pip-installable vLLM logits processor; certificate-preserving integration requires projection and no-op decisions to use the deployed sampling law.

Notably, the guard also has a clear task-compatibility limit: its detector doesn't distinguish degeneration from user-intended repetition. On 60 tasks that *request* long verbatim repetition, the untuned guard costs -1.40 in blind paired task fidelity (95% CI [-2.00,-0.78], 5-point scale), worst on refrains/tables/checklists (≈ -2.4), then code (-0.8), XML unchanged. The certificate is operating as specified—it bounds every detected repetition trajectory beyond $\ell_{\min}$, wanted or not. Therefore, we offer the guard as an opt-in feature and recommend task-aware routing or per-request disabling in real practice (Appendix F).

Table 1: Guard transfer under a fixed raw-prefix protocol. "at-risk" means that the matcher reaches $\ell_{\min}$=8; "sust." means that no detected trajectory reaches L=24 after first risk; "post-g." means $\ell_t \geq 8$ for some $t \in [\min(t_{\text{risk}} + 12, 255), 255]$, where t_{risk} is the zero-based first-risk step. This includes continued first runs, without requiring an exit. Both rates are over "at-risk" trajectories. Arrows mark preferred directions; brackets give guard-arm 95% Wilson intervals; "disc." gives paired improvement/harm counts; p is the exact McNemar test.

model	at-risk	sust. off→guard ↑	post-g. off→guard ↓	KL (nats) ↓	disc.	p
GPT-2 (float32)	10/56	0.50→**0.90** [.60,.98]	0.70→0.70 [.40,.89]	1.10	4/0	0.12
Qwen3-4B	93/140	0.71→**0.91** [.84,.96]	0.74→0.69 [.59,.77]	7.58	19/0	3.8×10^{-6}
Qwen2.5-7B-Instruct	80/140	0.90→**0.99** [.93,1.00]	0.69→0.64 [.53,.73]	5.18	7/0	0.02
Qwen3-14B	64/140	0.84→**1.00** [.94,1.00]	0.53→0.55 [.43,.66]	4.86	10/0	2.0×10^{-3}
R1-Distill-Qwen-14B	74/140	0.76→**0.97** [.91,.99]	0.58→0.55 [.44,.66]	5.90	16/0	3.1×10^{-5}
Mistral-Nemo-12B	42/140	0.83→**1.00** [.92,1.00]	0.55→0.55 [.40,.69]	2.27	7/0	0.02
Phi-3.5-mini	42/140	0.95→**0.98** [.88,1.00]	0.40→0.43 [.29,.58]	1.56	2/1	1.00

Table 1 evaluates guard transfer on six modern checkpoints across three families plus GPT-2, using the same pre-specified protocol across all models. Each modern model uses a fixed 28-prompt evaluation cohort with 5 paired seeds, yielding 140 trajectories per mode (raw prefixes without chat templates, T=0.8, top-p 1.0, fixed 256-token horizon). Runs use the exact matcher and report KL against the tempered law. Without retuning, guard improves the sustained-escape point estimate on every model, reaching 0.91–1.00 at 1.6–7.6 nats on modern checkpoints and 0.90 at 1.10 nats on GPT-2. Qwen3-4B gives the largest paired test: 93/140 at risk, sustained escape 0.71→0.91 ($p = 3.8 \times 10^{-6}$), surprisal 0.60→0.64 nats/token, but post-grace repetition only 0.74→0.69, indicating improved containment not full recovery. Appendix E gives protocol details and the excluded boundary case.

4.5 E4: NATURAL-WORKLOAD INCIDENCE AND MATCHED-PREFIX EVALUATION

To estimate the occurrence of sustained repetition in natural use, we sample 3,000 trajectories from 1,500 unscreened prompts (two seeds; Qwen3-4B, T=0.8, EOS respected, 1,024-token horizon). Of these, 78.3% reach risk at least once ([76.8,79.8]%) and 11.8% reach L=24 ([10.7,13.0]%; paired greedy: 20.8%). Cap-exceedance rates range from 32.2% in code generation to 0.5% in summarization (64.3$\times$), showing that unscreened natural workloads exhibit sustained repetition at strongly task-dependent rates (Appendix G, Table 7). To test whether the guard can contain repetition after risk, enriched Stage B branches from 710 identical first-risk prefixes. The guard lowers cap exceedance from 30.7% to 5.1%, and the 191/9 improvement/harm discordants (exact McNemar p=1.5×10^{-45}) establish a systematic benefit. The observed 5.07% is consistent with the pre-registered 5% budget (Wilson [3.7,6.9]%) at 52.96 nats mean accumulated KL, demonstrating budget-level post-risk containment under matched conditions (Appendix G).

Table 2: Cross-model measurements (100 prompts; 256 new tokens, GPT-2: 120). slope/Λ: median (range), asymptotic target 1; server curves use the top-20 upper bound P/ρ_{20}. Exit loc.: argmin-gap fraction (uniform baseline), from server hazard simulations or GPT-2 model rollouts. AUROC: onset/continuation on 60 real and 60 fabricated referents; V4-Flash scores are censored (Appendix I).

model	params	precision	captured	$b>0$	slope/Λ	exit loc.	AUROC
GPT-2	124M	fp32	64/100	100%	1.13 (1.06–1.16)	91% (13%)	0.96 / 0.77
Gemma-4-E4B-it	4B	bf16	0/100	–	–	–	1.00 / 0.86
Qwen3-4B	4B	bf16	19/100	100%	1.05 (0.98–1.10)	93% (6%)	0.97 / 1.00
Phi-3.5-mini	3.8B	bf16	4/100	100%	1.05 (0.98–1.12)	77% (6%)	1.00 / 0.99
Qwen2.5-7B-Instruct	7B	bf16	0/100	–	–	–	0.99 / 0.80
Mistral-7B-v0.3	7B	bf16	5/100	100%	1.07 (0.99–1.15)	71% (6%)	1.00 / 0.99
Gemma-4-12B-it	12B	bf16	21/100	90%	1.08 (1.06–1.10)	14% (21%)	0.92 / 0.63
Mistral-Nemo	12B	bf16	17/100	100%	1.05 (0.96–1.12)	68% (5%)	0.98 / 0.97
Qwen3-14B	14B	bf16	6/100	100%	1.02 (1.02–1.06)	84% (6%)	0.96 / 0.98
R1-Distill-Qwen	14B	bf16	2/100	100%	1.05 (1.05–1.05)	95% (5%)	0.94 / 0.44
V4-Flash (IQ4_XS)	284B-A13B	$\sim$4.2 bpw	17/100	100%	1.09 (0.99–1.23)	69% (7%)	0.96 / 0.99
V4-Flash (IQ3_XXS)	284B-A13B	$\sim$3.1 bpw	21/100	100%	1.06 (1.00–1.10)	83% (8%)	0.94 / 0.98

4.6 E5: Cross-model replication

Protocol. All eleven server configurations use the same 100 openers and entity benchmark. Their orbit statistics use top-20 renormalization, P/ρ_{20} frozen-mean upper bounds and hazard simulations (Appendix I).

Results (Table 2). Modern-model capture spans 0%–21% versus GPT-2's 64%, with no monotonic size trend common across configurations: Qwen3 falls 19%$\rightarrow$6% from 4B to 14B, whereas Gemma-4 rises 0%$\rightarrow$21% from 4B to 12B, and the 284B model still captures 17%. The 2 zero-capture configurations remain orbit-free after extending the 12 archived early-onset prompts to 512 tokens. Gap growth is positive in every $\geq$3-lap orbit on 8/9 server configurations with captures (90% on Gemma-4-12B), and all 37 orbits fit slopes within 0.96–1.23$\times$ their minimum gap across 0.9–6.1 nats and six families. Conditional on capture, the results support gap growth and weakest-gap scaling under the top-20 approximation. For G3-passing configurations, 68%–95% of simulated low-T exits localize at the weakest-gap phase versus 5%–8% baselines. Deep Gemma-4-12B exits fall outside that regime, marking an applicability boundary rather than a counterexample.

Quantization. To test whether weight compression changes the reported law, we compare the same checkpoint at $\sim$4.2 and $\sim$3.1 bits per weight. Although 77/100 paired generations diverge within ten tokens, slope/Λ remains 1.09$\rightarrow$1.06, gap growth stays positive, onset AUROC is 0.96$\rightarrow$0.94, and capture is 17/100$\rightarrow$21/100. Thus, the paired quantization-level comparison suggests that compression rewrites individual trajectories without materially changing the aggregate Arrhenius, capture, or onset statistics.

As a scoped extension beyond repetition, we find that the same logit-energy observables provide a diagnostic lens on fabricated-entity onset across models: onset AUROCs remain 0.92–1.00, whereas post-onset continuation signals are model-dependent (0.44–1.00). We collect the full spike-and-capture case study in Appendix J as evidence and a starting point for future work on hallucination diagnostics.

5 Limitations and future work

Scope. Finite traces cannot certify either infinite-horizon branch of Theorem 3: neither prolonged observed survival nor rising gaps establishes permanent trapping. This work limits detected verbatim repetition, but semantic recovery, hallucinations, and paraphrastic, topical, or agentic loops are left for future work. Because detection is intent-agnostic, requests that legitimately require repetition (e.g., a poem with a refrain) must currently bypass the controllers (Appendix F). **Future work.** We will identify the causal mechanisms of first capture and develop task-aware routing that distinguishes harmful degeneration from intended repetition.

6 Conclusion

Full logits make detected repetition quantitatively tractable: the weakest gap sets its escape scale. Inverting this law yields a theory of repetition control and two derived controllers—a provably minimal-temperature pulse and a minimum-KL guard—that contain loops across modern transformer-based language models.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic data sets (100 input prompts), prototyping and developing scripts and analysis programs required for conducting experiments, as well as rephrasing and condensing this manuscript. We have not used generative AI tools for theoretical development or the reported results: all methodological contributions, mathematical derivations, experimental design, result verification, and scientific interpretation were conducted independently by the authors. We have reviewed all AI-assisted work: All LLM-generated prompts and code were verified, revised where necessary, and tested for correctness by two authors. All AI-assisted rephrasing of the manuscript was subsequently revised and reviewed by all authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The work analyzes and mitigates failure modes (degeneration, hallucination onset) of generative models, and we do not foresee direct negative applications beyond the dual-use character of any improved decoding control. No human-subjects research was conducted. The workload census draws prompts from publicly released corpora, including user-generated conversations (WildChat-1M, collected with user consent and PII-filtered by its curators); we use these texts only as decoding inputs, do not redistribute raw texts (the release ships a deterministic regeneration script with per-prompt content hashes), and entity lists contain public historical figures and synthetic names.

REPRODUCIBILITY STATEMENT

All experiments are checkpointed and re-entrant and the complete artifact is available from the anonymous repository. It contains the exact prompts where redistribution is permitted and, for restricted natural-workload corpora, dataset identifiers, content hashes, and a deterministic regeneration script. It also contains the seeds, the official GPT-2 model identifier and pinned Transformers/PyTorch environment, per-unit JSON outputs for every figure and table, and scripts that regenerate every quoted number and table (`paper/scripts/`). The artifact includes the HF and llama.cpp reproduction paths, per-configuration model and provenance records, and llama.cpp preflight reports (detailed in Appendix I).

REFERENCES

Mark Braverman, Xinyi Chen, Sham Kakade, Karthik Narasimhan, Cyril Zhang, and Yi Zhang. Calibration, entropy rates, and memory in language models. In *International Conference on Machine Learning*, pp. 1089–1099. PMLR, 2020.

Ting-Rui Chiang and Yun-Nung Chen. Relating neural text degeneration to exposure bias. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 228–239, 2021.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The annals of probability*, pp. 146–158, 1975.

Shehzaad Dhuliawala, Ilia Kulikov, Ping Yu, Asli Celikyilmaz, Jason Weston, Sainbayar Sukhbaatar, and Jack Lanchantin. Adaptive decoding via latent preference optimization. *arXiv preprint arXiv:2411.09661*, 2024.

Jianshuo Dong, Ziyuan Zhang, Qingjie Zhang, Tianwei Zhang, Hao Wang, Hewu Li, Qi Li, Chao Zhang, Ke Xu, and Han Qiu. An engorgio prompt makes large language model babble on. In *International Conference on Learning Representations*, volume 2025, pp. 67280–67307, 2025.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, 2018.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Mark I. Freidlin and Alexander D. Wentzell. *Random Perturbations of Dynamical Systems*. Springer, 3rd edition, 2012.

Zihao Fu, Wai Lam, Anthony Man-Cho So, and Bei Shi. A theoretical analysis of the repetition problem in text generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 12848–12856, 2021.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. *Bulletin (New Series) of the American Mathematical Society*, 62(3):427, 2025.

Ghaith Hammouri, Kemal Derya, and Berk Sunar. Non-halting queries: Exploiting fixed points in llms. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 1–22. IEEE, 2025.

Peter Hänggi, Peter Talkner, and Michal Borkovec. Reaction-rate theory: fifty years after kramers. *Reviews of modern physics*, 62(2):251, 1990.

John Hewitt, Christopher D Manning, and Percy Liang. Truncation sampling as language model desmoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 3414–3427, 2022.

Tatsuya Hiraoka and Kentaro Inui. Repetition neurons: How do language models produce repetitions? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 483–495, 2025.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

Donghao Huang, Thanh-Son Nguyen, Fiona Liausvia, and Zhaoxia Wang. Rap: A metric for balancing repetition and performance in open-source large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1479–1496, 2025.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*, 2019.

Hendrik Anthony Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *physica*, 7(4):284–304, 1940.

Huayang Li, Tian Lan, Zihao Fu, Deng Cai, Lemao Liu, Nigel Collier, Taro Watanabe, and Yixuan Su. Repetition in repetition out: Towards understanding neural text degeneration from the data perspective. *Advances in Neural Information Processing Systems*, 36:72888–72903, 2023.

Potsawee Manakul, Adian Liusie, and Mark 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*, pp. 9004–9017, 2023.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. Locally typical sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121, 2023.

Sidharth Mudgal, Jong Lee, Harish Ganapathy, YaGuang Li, Tao Wang, Yanping Huang, Zhifeng Chen, Heng-Tze Cheng, Michael Collins, Trevor Strohman, et al. Controlled decoding from language models. *arXiv preprint arXiv:2310.17022*, 2023.

Adam S Shai, Sarah E Marzen, Lucas Teixeira, Alexander G Oldenziel, and Paul M Riechers. Transformers represent belief state geometry in their residual stream. *Advances in Neural Information Processing Systems*, 37:75012–75034, 2024.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pp. 35151–35174. PMLR, 2023.

Zhichao Wang, Dongyang Ma, Xinting Huang, Deng Cai, Tian Lan, Jiahao Xu, Haitao Mi, Xiaoying Tang, and Yan Wang. The end of manual decoding: Towards truly end-to-end language models. In *International Conference on Learning Representations*, volume 2026, pp. 17625–17643, 2026.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*, 2019.

Dongjie Xu, Hao Wu, Weijie Shi, Yue Cui, Yuanjun Liu, Jiawei Li, Haolun Ma, An Liu, Jia Zhu, and Jiajie Xu. Loopguard: Breaking self-reinforcing attention loops via dynamic kv cache intervention. *arXiv preprint arXiv:2604.10044*, 2026.

Jin Xu, Xiaojiang Liu, Jianhao Yan, Deng Cai, Huayang Li, and Jian Li. Learning to break the loop: Analyzing and mitigating repetitions for neural text generation. *Advances in Neural Information Processing Systems*, 35:3082–3095, 2022.

A PROOFS

Throughout the proofs, fixed logits $z \in \mathbb{R}^V$ have a unique greedy index $g := \arg\max_i z_i$. For each token i, let $u_i := z_g - z_i \geq 0$ be its logit deficit and let $\Delta := \min_{i \neq g} u_i > 0$ be the runner-up gap. The tempered law is $\pi_i^{(T)} = e^{z_i/T}/Z(T)$ with $Z(T) = \sum_j e^{z_j/T}$, and the scalar departure hazard is $q(T) = 1 - \pi_g^{(T)}$. The failure budget is denoted by δ, and controlled probability laws by ν.

A.1 PROOF OF LEMMA 1

Fix $T > 0$. Because the greedy index is unique, $z_g - z_i > 0$ for every $i \neq g$. Differentiating the log probability of the greedy token gives

$$\ln \pi_g^{(T)} = \frac{z_g}{T} - \ln Z(T), \qquad \frac{d}{dT} \ln Z(T) = -\frac{1}{T^2} \sum_i \pi_i^{(T)} z_i,$$

and hence

$$\frac{d}{dT} \ln \pi_g^{(T)} = -\frac{z_g}{T^2} + \frac{1}{T^2} \mathbb{E}_{\pi^{(T)}}[z] = \frac{\mathbb{E}_{\pi^{(T)}}[z] - z_g}{T^2}.$$

The numerator is strictly negative,

$$z_g - \mathbb{E}_{\pi^{(T)}}[z] = \sum_{i \neq g} \pi_i^{(T)} (z_g - z_i) > 0,$$

because softmax assigns positive probability to every token. Therefore

$$\frac{d}{dT} \pi_g^{(T)} = \pi_g^{(T)} \frac{d}{dT} \ln \pi_g^{(T)} < 0, \qquad q'(T) = -\frac{d}{dT} \pi_g^{(T)} > 0.$$

For completeness, factor $e^{z_g/T}$ out of the softmax denominator:

$$\pi_g^{(T)} = \frac{1}{1 + \sum_{i \neq g} e^{-(z_g - z_i)/T}}.$$

As $T \to 0^+$, every term in the sum tends to zero, so $\pi_g^{(T)} \to 1$ and $q(T) \to 0$. As $T \to \infty$, each $e^{z_i/T} \to 1$, so $\pi_g^{(T)} \to 1/V$ and $q(T) \to 1 - 1/V$. This proves both strict monotonicity and the two stated limits. □

A.2 PROOF OF THEOREM 1

Factoring the greedy logit out of the partition function yields

$$\pi_g^{(T)} = \frac{1}{1 + s(T)}, \qquad s(T) := \sum_{i \neq g} e^{-u_i/T}, \qquad q(T) = \frac{s(T)}{1 + s(T)}.$$

By definition, every non-greedy deficit satisfies $u_i \geq \Delta$. There are $V-1$ such terms, and at least one runner-up token has deficit exactly Δ. Consequently,

$$e^{-\Delta/T} \leq s(T) \leq (V-1)e^{-\Delta/T}.$$

Since $s/(1+s) \leq s$, the upper estimate immediately gives $q(T) \leq (V-1)e^{-\Delta/T}$. For the lower estimate, the map $s \mapsto s/(1+s)$ is increasing, so retaining only a runner-up term gives

$$q(T) \geq \frac{e^{-\Delta/T}}{1+e^{-\Delta/T}} \geq \tfrac{1}{2}e^{-\Delta/T},$$

where the last inequality uses $0 < e^{-\Delta/T} \leq 1$. We have therefore proved the two-sided estimate in Equation 3.

Because every quantity in that estimate is positive, taking logarithms and multiplying by T gives

$$-\Delta - T\ln 2 \leq T\ln q(T) \leq -\Delta + T\ln(V-1).$$

The lower and upper bounds both converge to $-\Delta$ as $T \to 0^+$. The squeeze theorem then gives $\lim_{T\to 0^+} T\ln q(T) = -\Delta$. □

A.3 Proof of Theorem 2

(i) Survival identity. For $1 \leq s \leq H$, define the agreement event $G_s := \{W_s = g_s\}$. By the definition of the first-exit time,

$$\{\tau > t\} = \bigcap_{s=1}^{t} G_s.$$

Conditioned on $G_1 \cap \cdots \cap G_{s-1}$, the sampled tokens up to step $s-1$ equal the stored greedy-reference tokens. The two decoding contexts are therefore identical at step s, and the saved logit row gives

$$\Pr_T(G_s \mid G_1 \cap \cdots \cap G_{s-1}) = \Pr_T(W_s = g_s \mid \tau > s-1) = 1 - q_s(T).$$

Applying the probability chain rule to the intersection gives

$$S_T(t) = \Pr_T(\tau > t) = \prod_{s=1}^{t} \Pr_T(G_s \mid G_1 \cap \cdots \cap G_{s-1}) = \prod_{s=1}^{t}(1-q_s(T)).$$

Every conditional probability here is computed from a stored logit row, so this identity is exact for $t \leq H$ and makes no independence assumption.

(ii) Arrhenius exit time. Under the frozen-profile idealization, the hazards are P-periodic. Write

$$A_r(T) := \prod_{j=1}^{r}(1-q_j(T)), \qquad A_0(T) := 1, \qquad \rho(T) := 1 - A_P(T).$$

Thus $A_r(T)$ is the probability of surviving the first r phases of a lap, and $\rho(T)$ is the probability of exiting somewhere in that lap. If $t = kP + r$ with $k \geq 0$ and $0 \leq r < P$, periodicity and part (i) give

$$S_T(kP+r) = A_P(T)^k A_r(T) = (1-\rho(T))^k A_r(T).$$

For the positive integer-valued variable τ, the tail-sum identity is $\mathbb{E}_T[\tau] = \sum_{t=0}^{\infty} \Pr_T(\tau > t)$. Substituting the preceding decomposition and summing the nonnegative terms by laps gives

$$\mathbb{E}_T[\tau] = \sum_{k=0}^{\infty}\sum_{r=0}^{P-1}(1-\rho(T))^k A_r(T) = \frac{\sum_{r=0}^{P-1} A_r(T)}{\rho(T)},$$

where the last equality is the geometric-series formula. This is Equation 5. Each factor $1-q_j(T)$ lies in $(0,1)$, and hence $A_P(T) \leq A_r(T) \leq 1$ for $0 \leq r < P$. Therefore

$$P(1-\rho(T)) = PA_P(T) \leq \sum_{r=0}^{P-1} A_r(T) \leq P.$$

Dividing by $\rho(T) > 0$ proves

$$P\frac{1-\rho(T)}{\rho(T)} \le \mathbb{E}_T[\tau] \le \frac{P}{\rho(T)}.$$

It remains to identify the low-temperature exponential scale. Let $\Lambda := \min_{1\le j\le P} \Delta_j$. Because $\rho(T) = 1-\prod_{j=1}^{P}(1-q_j(T))$, we have $\max_j q_j(T) \le \rho(T) \le \sum_{j=1}^{P} q_j(T)$: the lower bound follows from $\prod_l(1-q_l(T)) \le 1-q_j(T)$ for every j, and the upper bound is the standard product–sum inequality. Choosing a phase whose gap equals Λ and applying Theorem 1 then yields

$$\tfrac{1}{2}e^{-\Lambda/T} \le \max_j q_j(T) \le \rho(T) \le \sum_{j=1}^{P} q_j(T) \le P(V-1)e^{-\Lambda/T}.$$

After taking logarithms and multiplying by T, the two outer terms imply

$$\lim_{T\to 0^+} T\ln\rho(T) = -\Lambda.$$

Moreover, every $q_j(T) \to 0$, so every $A_r(T) \to 1$ and $\sum_{r=0}^{P-1} A_r(T) \to P$. Taking the logarithm of Equation 5 now gives

$$T\ln\mathbb{E}_T[\tau] = T\ln\left(\sum_{r=0}^{P-1} A_r(T)\right) - T\ln\rho(T) \longrightarrow \Lambda.$$

(iii) Exit localization. Within a lap, exiting first at phase j requires survival through phases $1,\dots,j-1$ followed by departure at phase j. Its probability is therefore

$$\Pr_T(J = j,\ \text{exit in the lap}) = q_j(T)A_{j-1}(T).$$

These mutually exclusive probabilities sum to $\rho(T)$, so division by $\rho(T)$ proves the first equality in Equation 6.

Now suppose j^* is the unique minimizer, so $\Delta_{j^*} = \Lambda$, and define $\Lambda_2 := \min_{j\ne j^*} \Delta_j > \Lambda$. By Theorem 1, for every $j \ne j^*$,

$$\frac{q_j(T)}{q_{j^*}(T)} \le 2(V-1)e^{-(\Delta_j-\Lambda)/T} \le 2(V-1)e^{-(\Lambda_2-\Lambda)/T}.$$

The survival factor preceding phase j^* tends to one because all $q_l(T) \to 0$. Hence, for all sufficiently small T, $A_{j^*-1}(T) \ge 1/2$. Using the j^* exit event as a lower bound on $\rho(T)$ gives

$$\begin{aligned}\Pr_T(J \ne j^* \mid \text{exit in the lap}) &= \frac{\sum_{j\ne j^*} q_j(T)A_{j-1}(T)}{\rho(T)} \\ &\le \frac{\sum_{j\ne j^*} q_j(T)}{q_{j^*}(T)A_{j^*-1}(T)} \\ &\le 4(V-1)(P-1)e^{-(\Lambda_2-\Lambda)/T} \longrightarrow 0.\end{aligned}$$

Thus the conditional probability of exit at j^* tends to one. □

A.4 Proof of Theorem 3

The events $\{\tau > t\}$ decrease to $\{\tau = \infty\}$. Continuity of probability from above and Equation 4 therefore give

$$\Pr_T(\tau = \infty) = \lim_{t\to\infty} S_T(t) = \prod_{t=1}^{\infty}(1-q_t(T)).$$

We next recall why this infinite product is positive exactly when the hazards are summable. If $\sum_t q_t(T) = \infty$, the elementary inequality $\ln(1-x) \le -x$ for $0 \le x < 1$ gives

$$\sum_{t=1}^{n}\ln(1-q_t(T)) \le -\sum_{t=1}^{n} q_t(T) \longrightarrow -\infty.$$

Exponentiating shows that the partial products converge to zero. Conversely, if $\sum_t q_t(T) < \infty$, then $q_t(T) \to 0$. Choose t_0 so that $q_t(T) \le 1/2$ for $t \ge t_0$. Since $\ln(1-x) \ge -2x$ on $[0, 1/2]$,

$$\sum_{t=t_0}^{\infty} \ln(1 - q_t(T)) \ge -2 \sum_{t=t_0}^{\infty} q_t(T) > -\infty.$$

Thus the tail product is positive. Its finite prefix is also positive because finite-logit softmax gives $0 < q_t(T) < 1$. This proves Equation 7.

(i) Permanent deepening leaves non-exit mass. For a step t in lap $m = \lfloor (t-1)/P \rfloor$, the growth premise of Equation 8 and Theorem 1 give

$$q_t(T) \le (V-1)e^{-\Delta_t/T} \le (V-1)e^{-(\Delta_0 + b_\Delta m)/T}.$$

There are exactly P phases in each lap. Grouping the hazards by laps and summing the resulting geometric series gives

$$\sum_{t \ge 1} q_t(T) \le (V-1)P \sum_{m \ge 0} e^{-(\Delta_0 + b_\Delta m)/T} = (V-1)P \frac{e^{-\Delta_0/T}}{1 - e^{-b_\Delta/T}}.$$

Because $b_\Delta > 0$, the denominator is positive and the series is finite. Equation 7 then implies $\Pr_T(\tau = \infty) > 0$. In addition, the probability of first exit at step t is $S_T(t-1)q_t(T) \le q_t(T)$, so

$$\Pr_T(\tau < \infty) = \sum_{t \ge 1} S_T(t-1) q_t(T) \le \sum_{t \ge 1} q_t(T).$$

Combining this inequality with the trivial probability bound $\Pr_T(\tau < \infty) \le 1$ yields the stated minimum of one and the geometric-series expression.

(ii) Uniformly bounded gaps. If $\Delta_t \le M$, the lower estimate in Theorem 1 provides the uniform hazard floor

$$q_t(T) \ge \tfrac{1}{2} e^{-M/T} =: \varepsilon_T > 0.$$

Consequently,

$$S_T(t) = \prod_{s=1}^{t} (1 - q_s(T)) \le (1 - \varepsilon_T)^t \longrightarrow 0.$$

Thus $\Pr_T(\tau = \infty) = 0$. Applying the tail-sum identity once more,

$$\mathbb{E}_T[\tau] = \sum_{t=0}^{\infty} S_T(t) \le \sum_{t=0}^{\infty} (1 - \varepsilon_T)^t = \varepsilon_T^{-1}.$$

For the complementary lower estimates, assume $\Delta_t \ge \Lambda_{\mathrm{sat}} > 0$ and set $\bar{q}_T := \sup_t q_t(T)$. Let $C_T := (V-1)e^{-\Lambda_{\mathrm{sat}}/T}$ and write $\sigma_t(T)$ for the step-t deficit sum of Theorem 1, so that $q_t(T) = \sigma_t(T)/(1 + \sigma_t(T))$, and $\sigma_t(T) \le C_T$ because every deficit at step t is at least $\Delta_t \ge \Lambda_{\mathrm{sat}}$. Because $x \mapsto x/(1+x)$ is increasing, $q_t(T) \le C_T/(1 + C_T) < 1$ at every step. Hence $\bar{q}_T < 1$, and the looser bound stated in the theorem is

$$\bar{q}_T \le C_T = (V-1)e^{-\Lambda_{\mathrm{sat}}/T}.$$

Since $q_t(T) \le \bar{q}_T$ at every step,

$$S_T(t) \ge (1 - \bar{q}_T)^t, \qquad \mathbb{E}_T[\tau] \ge \sum_{t=0}^{\infty} (1 - \bar{q}_T)^t = \bar{q}_T^{-1} \ge \frac{e^{\Lambda_{\mathrm{sat}}/T}}{V-1}.$$

Finally, summing the mutually exclusive first-exit events over the first H steps gives

$$\Pr_T(\tau \le H) = \sum_{t=1}^{H} S_T(t-1) q_t(T) \le \sum_{t=1}^{H} q_t(T) \le H \bar{q}_T.$$

Taking the minimum with one completes the finite-horizon bound. □

A.5 Proof of Proposition 1

(i) Exact calibration. For every stored phase j, Lemma 1 shows that $q_j(T)$ is continuous and strictly increasing. Hence each survival factor $1 - q_j(T)$ is continuous and strictly decreasing. For a nonempty window $\mathcal{W}$ of size k, their product is strictly decreasing, so

$$E_{\mathcal{W}}(T) = 1 - \prod_{j \in \mathcal{W}} (1 - q_j(T))$$

is continuous and strictly increasing. The low-temperature limit is zero. In the high-temperature limit, every next-token law becomes uniform, $q_j(T) \to 1 - 1/V$, and therefore

$$E_{\mathcal{W}}(T) \longrightarrow 1 - \prod_{j \in \mathcal{W}} \frac{1}{V} = 1 - V^{-k}.$$

If $\delta > V^{-k}$, the target $1 - \delta$ lies strictly between these two limits. The intermediate value theorem gives a unique finite solution of $E_{\mathcal{W}}(T) = 1 - \delta$, and strict monotonicity makes bisection applicable. If $\delta \le V^{-k}$, no finite solution exists. Indeed, at every finite temperature each deficit u_i is positive, so the denominator $1 + \sum_{i \neq g} e^{-u_i/T}$ of the greedy probability is strictly below V; every survival factor therefore satisfies $1 - q_j(T) > 1/V$, whence $E_{\mathcal{W}}(T) < 1 - V^{-k} \le 1 - \delta$. This proves the stated existence condition for $T^*_{\mathcal{W}}$.

(ii) Conservative closed form. Let $\widetilde{\Delta}_{\mathcal{W}} := \max_{j \in \mathcal{W}} \Delta_j$. For every $j \in \mathcal{W}$, we have $\Delta_j \le \widetilde{\Delta}_{\mathcal{W}}$, so the lower estimate in Theorem 1 gives

$$q_j(T) \ge \tfrac{1}{2} e^{-\Delta_j/T} \ge \tfrac{1}{2} e^{-\widetilde{\Delta}_{\mathcal{W}}/T}.$$

It follows that

$$1 - E_{\mathcal{W}}(T) \le \big(1 - \tfrac{1}{2} e^{-\widetilde{\Delta}_{\mathcal{W}}/T}\big)^k \le \exp\!\big(-\tfrac{k}{2} e^{-\widetilde{\Delta}_{\mathcal{W}}/T}\big),$$

where the second inequality uses $1 - x \le e^{-x}$. To make the final quantity at most δ, it is sufficient that

$$\frac{k}{2} e^{-\widetilde{\Delta}_{\mathcal{W}}/T} \ge \ln(1/\delta).$$

The assumption $k > 2\ln(1/\delta)$ makes $\ln\!\big(k/[2\ln(1/\delta)]\big)$ positive. Solving the preceding inequality for T shows that every

$$T \ge \widehat{T}_{\mathcal{W}} := \frac{\widetilde{\Delta}_{\mathcal{W}}}{\ln\!\big(k/[2\ln(1/\delta)]\big)}$$

suffices. At equality, $e^{-\widetilde{\Delta}_{\mathcal{W}}/\widehat{T}_{\mathcal{W}}} = 2\ln(1/\delta)/k$, so the exponential upper bound is exactly δ.

(iii) Cyclic placement. For fixed k and δ, the denominator in $\widehat{T}_{\mathcal{W}}$ is a positive constant. Thus the conservative sufficient temperature is strictly increasing in $\widetilde{\Delta}_{\mathcal{W}}$. Comparing all cyclic windows of length k therefore reduces to minimizing their maximum phase gap, which is precisely the placement rule in the proposition. □

A.6 Proof of Proposition 2

(i) One-step information projection. Let $\chi(i) := \mathbf{1}\{i \notin \mathcal{C}\}$ be the exit indicator, $p := \mu(\chi = 1)$, and $a := \nu(\chi = 1)$. Write μ_b and ν_b for the conditional token laws given $\chi = b$. The relative-entropy chain rule gives

$$D_{\mathrm{KL}}(\nu\|\mu) = d_2(a\|p) + (1-a) D_{\mathrm{KL}}(\nu_0\|\mu_0) + a D_{\mathrm{KL}}(\nu_1\|\mu_1).$$

Thus the data-processing lower bound $D_{\mathrm{KL}}(\nu\|\mu) \ge d_2(a\|p)$ is, for $a \in (0,1)$, tight exactly when the coarse-graining loses no paid information, i.e., $\nu_b = \mu_b$ for both groups. For such a the unique finite-cost minimizer therefore satisfies

$$\nu_i = \mu_i \frac{1-a}{1-p} \quad (i \in \mathcal{C}), \qquad \nu_i = \mu_i \frac{a}{p} \quad (i \notin \mathcal{C}),$$

and its cost is $d_2(a\|p)$. Moreover,

$$\frac{\partial}{\partial a} d_2(a\|p) = \ln \frac{a(1-p)}{p(1-a)},$$

which is positive for $a > p$. If $p \geq r$, the unconstrained minimizer $a = p$ is feasible, giving the certified no-op $\nu = \mu$ at zero cost. If $p < r$, the binary divergence is increasing throughout the feasible interval $a \geq r$, so the I-projection lies on the information-geometric boundary $a = r$ and the preceding rescaling is exactly Equation 11.

It remains to interpret the binding-case cost $d_2(r\|p)$ and derive its rare-exit form. From an information-theoretic viewpoint, relative entropy is the excess codelength given by cross-entropy minus entropy. Let the controlled exit/stay decision follow $\mathrm{Bernoulli}(r)$ and its base-law counterpart $\mathrm{Bernoulli}(p)$. The intrinsic uncertainty of the controlled decision is the binary entropy

$$H_2(r) := -r \ln r - (1-r)\ln(1-r),$$

whereas the expected codelength for outcomes drawn from the controlled decision but encoded with the base law is the binary cross-entropy

$$H_\times(r,p) := -r \ln p - (1-r)\ln(1-p).$$

Expanding the binary relative entropy and grouping these two quantities gives

$$\begin{aligned} d_2(r\|p) &= r \ln \frac{r}{p} + (1-r)\ln\frac{1-r}{1-p} \\ &= \big[-r\ln p - (1-r)\ln(1-p)\big] - \big[-r\ln r - (1-r)\ln(1-r)\big] \\ &= H_\times(r,p) - H_2(r). \end{aligned}$$

Now define the effective exit barrier as the self-information of a natural exit under the base law, $\Delta_{\mathrm{eff}} := -\ln p$. Substitution into the preceding identity yields the exact decomposition

$$d_2(r\|p) = r\Delta_{\mathrm{eff}} - H_2(r) - (1-r)\ln(1-p).$$

Thus $r\Delta_{\mathrm{eff}}$ is the exit contribution to the base-coded description length, $H_2(r)$ removes the irreducible uncertainty of the controlled binary decision, and the final term is the stay-event correction. Finally, for fixed $r \in (0,1)$, the Taylor expansion $-\ln(1-p) = p + O(p^2)$ gives

$$-(1-r)\ln(1-p) = (1-r)p + O(p^2) = O(p),$$

and hence

$$d_2(r\|p) = r\Delta_{\mathrm{eff}} - H_2(r) + O(p),$$

which proves the stated rare-exit asymptotic and completes part (i).

(ii) Sequential survival information. Work under D, and let L_j be the event that the armed trajectory is still live j tokens after arming; L_0 is certain given D. On every live step m, part (i) makes the controlled exit mass at least $r_{L,\delta}$: the constraint binds when the current base exit mass is smaller, and otherwise the base law already complies and is a certified no-op. Leaving $\mathcal{C}_m$ drops the matcher below $\ell_{\min}$ and ends the trajectory, so

$$\Pr(\mathsf{L}_{m+1} \mid \mathsf{L}_m, \mathsf{D}) \leq 1 - r_{L,\delta}.$$

The events are nested, so the probability chain rule gives, without an independence assumption,

$$\Pr(\mathsf{L}_j \mid \mathsf{D}) = \prod_{m=0}^{j-1} \Pr(\mathsf{L}_{m+1} \mid \mathsf{L}_m, \mathsf{D}) \leq (1 - r_{L,\delta})^j,$$

equivalently $-\ln \Pr(\mathsf{L}_j \mid \mathsf{D}) \geq j[-\ln(1 - r_{L,\delta})]$: conditional surprisal is additive. A matched suffix lengthens by at most one token per step, so $\ell_{\mathrm{traj}} \geq \ell_{\min} + j$ costs at least j live steps, i.e. $\{\ell_{\mathrm{traj}} \geq \ell_{\min} + j\} \subseteq \mathsf{L}_j$, which turns the display into the first inequality of Equation 12. By definition of $r_{L,\delta}$,

$$(1 - r_{L,\delta})^{L-\ell_{\min}} = \delta, \qquad (L - \ell_{\min})[-\ln(1 - r_{L,\delta})] = \ln(1/\delta),$$

so $j = L - \ell_{\min}$ gives the second. The argument conditions only on the current detected repetition trajectory, so it is period-agnostic and restarts whenever the matcher re-arms after a departure. □

B Implementation details

Software Stack. GPT-2 experiments load the official `gpt2` checkpoint (revision `607a30d`) and its 50,257-type byte-level BPE tokenizer through Hugging Face Transformers 5.15 and PyTorch 2.13.

Model inference uses float32, the analysis layer promotes every stored logit row to float64 before softmax, and the sequential decoder uses the native key–value cache. The same backend exposes full predictive rows. The submitted checkpoints include all 100 primary and 36 eligible extended E1 records. The separate matched-baseline cohort below retains its pre-registered bfloat16 inference protocol.

Determinism and Precision. The survival identity requires that context-identical steps produce identical logits. Deterministic PyTorch algorithms and fixed seeds are enabled, model loading asserts bitwise agreement between repeated identical forwards, and every E2 trajectory independently recomputes one mid-window row and checks agreement with the stored full-logit trace to 2×10^{-3} absolute (covering harmless fp32 accumulation-order changes across sequence lengths).

Cycle detection. Exact trailing-lag matching: the largest k such that $w_t = w_{t-P}$ for the last k positions, accepted when $k \geq 2P$; the reported period is the $P \leq 24$ (or ≤ 48) maximizing coverage. Exact equality makes every accepted tail periodic under this event definition.

Pulse controller. Default configuration is $\delta = 0.1$, $k = 6$, refractory 4. The first detection requires a stable period for three consecutive steps and later detections for one. We use $T \in [0.2, 3]$ as the bisection bracket. The E3a planner stores the last ≤ 20 float64 logit rows ($\sim$8 MB at $V = 50{,}257$, and the streaming port buffers float32 rows, $\sim$4 MB). It selects among P cyclic placements using Pk gap evaluations, then performs 42 bisection iterations on the chosen window ($42k$ softmax evaluations), reducing the initial bracket width by 2^{-42}. The fast-re-arm variant sets refractory 2 and a one-step stability requirement throughout; the cruise variant samples at $T_c = 0.25$ between pulses.

Algorithm 1 Gap-guided minimal thermal pulse (δ=0.1, k=6, refractory 4)

1: decode greedily; keep the last ≤ 20 logit rows and a trailing token window
2: **on** stable period P detected for 3 steps: ▷ MONITOR → ARMED
3: choose $\mathcal{W}^* = \arg\min_{\mathcal{W}} \max_{j \in \mathcal{W}} \Delta_j$ over the P cyclic k-step windows
4: bisect $T^*_{\mathcal{W}^*} = \min\{T : E_{\mathcal{W}^*}(T) \geq 1 - \delta\}$ using Equation 10
5: wait until $\mathcal{W}^*$; **then** sample at $T^*_{\mathcal{W}^*}$ for $\leq k$ steps ▷ PULSING
6: **on** first departure or window end: revert to greedy, refractory 4 steps

C PROMPTS (E1)

The census uses 100 openers in ten equal groups: news/reportage, dialogue/first-person, essay/opinion, how-to/instructional, fiction/narrative, review/commerce, QA/encyclopedic, technical/software, general-purpose, and academic/research. The fixed list ships in `arrhenius/src/arrhenius/prompts.py`, with ten indexed openers per group. Representative examples include "The city council met on Tuesday to discuss" (news), "He walked into the room and" (dialogue), "To make the perfect cup of coffee," (how-to), "The function returns an error when" (technical), and "Recent studies suggest that the" (academic/research).

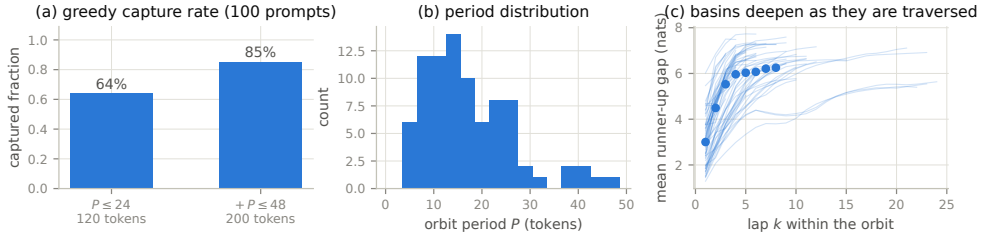


Figure 3: The attractor census (100 prompts, greedy). (a) Fraction of trajectories captured by a periodic orbit at two detection scopes. (b) Orbit periods. (c) Per-lap mean runner-up gap for every captured orbit with ≥ 3 laps (thin, 62 in total) and the cross-orbit median (dots, drawn only if at least 5 trajectories have an observed value for that lap): traversal deepens the basin.

Flagship orbit (E1/E2). Prompt "The most important thing to remember is" is captured at full-sequence index $t = 24$ (including the prompt tokens) by the period-12 orbit "*You have to say "I'm going to do this."* ", which it then traverses verbatim; per-lap mean next-token probability rises 0.56, 0.70, 0.88, 0.94, 0.96, 0.97 over six laps.

D ONE-SHOT ESCAPE AND MINIMUM-KL ABLATIONS (E3A)

Metrics and accounting. For a token sequence y, let $N_n(y)$ and $U_n(y)$ denote its numbers of total and unique contiguous n-grams. We report seq-rep-4$(y) = 1 - U_4(y)/N_4(y)$ on the trailing 96 continuation tokens and distinct-2$(y) = U_2(y)/N_2(y)$ on all 128 tokens. Model-relative continuation surprisal is $128^{-1}\sum_{t=1}^{128} -\log_2 \mu_t(w_t)$ bits/token, where μ_t is the unmodified $T{=}1$ law at the realized context. Over intervened steps $\mathcal{I}$, $\sum_{t\in\mathcal{I}} \mathrm{TV}(\nu_t, \mathbf{e}_{g_t}) = \sum_{t\in\mathcal{I}}[1-\nu_t(g_t)]$ is the expected number of non-greedy selections, while $\sum_{t\in\mathcal{I}} D_{\mathrm{KL}}(\nu_t\|\mu_t)$, with $D_{\mathrm{KL}}(\nu\|\mu) = \sum_i \nu_i \ln(\nu_i/\mu_i)$, measures full-distribution distortion in nats. Departures are the realized count $\sum_{t=1}^{128} \mathbf{1}\{w_t \neq g_t\}$. These quantities separate end-to-end degeneration and diversity from model-relative plausibility and intervention size.

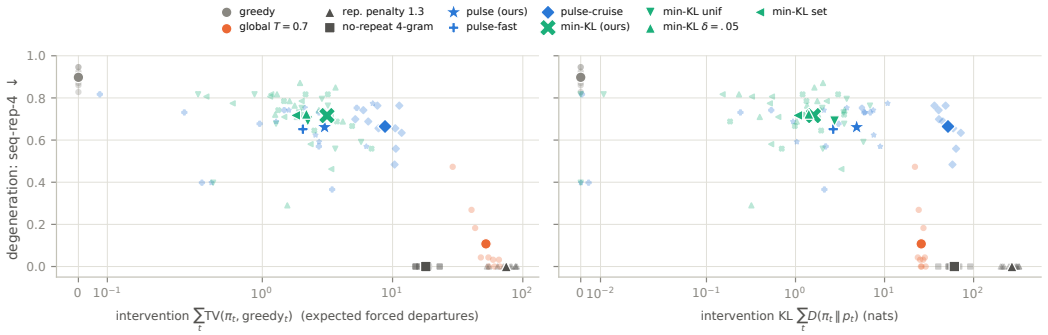


Figure 4: Degeneration against intervention cost on GPT-2 (10 prompts per policy; faint marks: single trajectories; solid marks: policy means), in the TV metric (left) and in intervention KL (right). No policy dominates the left panel, where the pulse and min-KL controllers occupy the low-intervention end; in distribution space (right) the minimum-KL controller occupies a low-KL region empirically. Proposition 2 guarantees single-step KL optimality at a fixed exit target, not end-to-end Pareto optimality.

Table 3: Escape policies on 10 degeneration-prone prompts (means over prompts; 128 new tokens each; seq-rep-4 is evaluated on the trailing 96). Intervention and KL follow the accounting above; departures are realized changed tokens.

policy	seq-rep-4 ↓	distinct-2 ↑	surprisal (bits)	intervention ↓	KL (nats) ↓	departures
greedy	0.90	0.09	0.38	0	0	0
global $T{=}0.7$	0.11	0.76	2.68	52.1	25.8	52.4
repetition penalty 1.3	0.00	1.00	3.84	74.6	269	74.6
no-repeat 4-gram	0.00	0.67	1.98	18.0	60.8	18.0
pulse (ours)	0.66	0.21	0.88	3.0	4.86	2.6
pulse, fast re-arm	0.65	0.22	0.82	2.0	2.65	1.8
pulse + cruise $T_c{=}0.25$	0.66	0.26	0.88	8.8	51.6	7.9
min-KL (ours)	0.72	0.17	0.79	3.1	1.59	3.2
min-KL, uniform-rate	0.69	0.19	0.76	2.2	2.74	2.3
min-KL, $\delta{=}0.05$	0.72	0.18	0.70	2.2	1.41	2.4
min-KL, cycle-set stay	0.72	0.18	0.74	1.8	1.05	2.3

Windowed minimum-KL ablation. The pulse intervenes in temperature space. Keeping its detector, stored-lap planning, and refractory machinery fixed yields the distribution-space allocation studied here. The public API exposes two complementary per-request modes: `pulse` uses the Arrhenius gap law for a minimum-temperature finite-window escape and then restores the base policy, whereas `guard` repeatedly applies the minimum-KL projection for sustained containment. The windowed minimum-KL schedule evaluated here is an E3a ablation that isolates the projection within the pulse's window, not a separately shipped controller.

Equation 11 is the distribution-space counterpart of the thermal law: the effective barrier Δ_{eff} is paid linearly at exchange rate r by the cheapest direct intervention. In TV terms the same step moves only $r - p \leq r$ of mass when the constraint binds.

The window program. Over an H_{win}-step stored window, choose per-step exit masses $r_j \geq p_j$, where p_j is the base exit mass at phase j. Writing $s_j := -\ln(1 - r_j)$, the frozen-profile constraint is $\sum_j s_j \geq \ln(1/\delta)$ and the objective is the survival-discounted sum $\sum_j \prod_{m<j}(1 - r_m)\, d_2(r_j \| p_j)$. A numerical dynamic program approximates this allocation on a 241-point spend grid with linear interpolation and local golden-section refinement (`arrhenius.core.minkl_window_plan`); no continuous global-optimality error bound is claimed. Each phase contributes its natural spend $-\ln(1 - p_j)$ for free, so naturally uncertain phases become certified no-ops. A uniform-rate ablation uses $r_j = \max\{p_j, 1 - \delta^{1/H_{\text{win}}}\}$ to isolate the value of the allocation.

At the pulse's own operating point (δ=0.1, H_{win}=6, identical detector and refractory), the minimum-KL controller escapes with 1.59 nats of intervention KL per 128-token trajectory against the pulse's 4.86 (3.1× lower), the uniform-rate ablation's 2.74 (allocation alone is worth 1.7×), and global temperature's 25.8 (16×). It fires 3.5 windows per trajectory and realizes 91% in-window escape, close to the frozen-profile $1 - \delta$ design target; certified no-ops are a minority on these deep basins (9% of intervened steps). On the quality axes the controller's repetition ceiling sits slightly above the pulse's (seq-rep-4 0.72 vs. 0.66) at lower continuation surprisal (0.79 vs. 0.88 bits). These are empirical trade-offs, not a guarantee of end-to-end minimum distortion or a causal identification of recapture. Table 3 also reports the δ=0.05 and cycle-set stay ablations. Both windowed controllers re-detect exact repetition only at $P \leq 16$; longer orbits, finite intervention windows, refractory delays, and post-exit reversion can all contribute to residual repetition. Appendix E evaluates sustained control with a period-agnostic matcher.

E E3b SUPPLEMENT: SUSTAINED CONTAINMENT AND THE vLLM PLUGIN

The windowed controllers of E3a share a ceiling that is easy to observe: they escape the orbit they detect, revert, and the reverted flow is recaptured—by the same basin or a neighboring one, sometimes with a period beyond the windowed detector's $P \leq 16$ scope, where the controller is blind. The initially rising gaps measured in E1 explain why re-entered basins can rapidly become metastable, but no infinite-horizon branch of Theorem 3 is asserted for these trajectories. No single-window scheme fixes the ceiling: escape is an event, but the attractor is a standing property of the context. The remedy is to make the guarantee a standing property too.

For a fixed forbidden-token set $\mathcal{C}$, the $r \to 1$ I-projection is hard masking followed by conditional sampling, $\mu(\cdot \mid \mathcal{C}^c)$, with KL cost $-\ln \mu(\mathcal{C}^c)$. In E3a and the 512-token E3b experiment, the hard n-gram baseline selects the highest-logit allowed token w and records $-\ln \mu(w)$ only when this changes the greedy choice. Table 5 instead samples from the masked tempered law. Both variants exclude the forbidden continuation, including when repetition is intended. The guard pays $d_2(r_{L,\delta} \,\|\, p_t)$ per binding step, leaves repetition shorter than $\ell_{\min}$ untouched, and degrades to no-ops when the model complies; however, an *intended* detected repetition trajectory beyond $\ell_{\min}$ is governed by the same certificate and is interrupted before L with probability at least $1 - \delta$.

The plugin. The controllers ship in the pip-installable `arrhenius-control` distribution (Python package `arrhenius`); `arrhenius/src/arrhenius/adapter.py` supplies one engine-level vLLM logits processor. It is configured per request, with `pulse` and `guard` selectable by name and `off` as the pass-through default. On an intervened step the controller forces a token it sampled internally, so subsequent temperature/top-p/top-k operations do not change that sampled law. The sequential certificate additionally requires the projection and no-op decision to use the actual unmodified sampling law; forcing binding steps alone does not ensure this. Idle steps touch nothing, and the guard keeps no logits history—its detector is an incremental $\ell_{\min}$-gram-anchored matcher with $O(1)$ amortized cost per token, so idle requests incur no accelerator–host logits traffic.

Live-engine integration check. On vLLM 0.27.1 with Qwen3-4B in bfloat16, the engine-registered processor evaluated the fixed 100-prompt cohort for 256 new tokens with off and guard modes selected per request. Prompts with seq-rep-4 ≥ 0.5 fall from 26/100 to 6/100, continuations with a longest exact substring rematch > 24 from 37/100 to 7/100, and mean seq-rep-4 from 0.31 to 0.20. The archived run used greedy decoding outside matcher-live steps and the controller's T=1 base law at live steps, so it is a descriptive end-to-end integration check rather than validation of the controlled-trajectory certificate. The artifact includes all per-prompt records and a documented matched-law follow-up runner source (`experiments/Appendix/vllm_integration.py`).

Protocol. The ten E3a prompts are decoded for 512 new tokens through the plugin's own streaming controllers: greedy; the windowed pulse and minimum-KL controllers (the E3a designs, now streaming); the guard at $(\ell_{\min}, L, \delta) = (8, 24, 0.05)$, i.e. $r_{24,0.05} = 0.171$, and a tighter L=16 point; no-repeat-4-gram; global T=0.7. Guard decoding is greedy outside matcher-live steps. At live steps its base is $\mathrm{softmax}(z)$ (T=1): no-ops sample from this base, binding steps from its projection. Guard KL is measured against this live-step base. Metrics are as in E3a, augmented with post-hoc repeated-stretch diagnostics: a four-gram matcher supplies suffix-rematch lengths, stretches are consecutive positions with length ≥ 8, and each stretch is summarized by its maximum rematch length. These differ from the online $\ell_{\min} = 8$ matcher-live trajectories in Equation 12. The archive retains decoded text, exact continuation token IDs, and stretch summaries; an offline stored-token audit (`paper/scripts/audit_e3c_text_replay.py`) reproduces every matcher-derived aggregate diagnostic for the ten guard trajectories. These post-hoc statistics still cannot be relabeled as measurements of the certificate event because the per-step logits, base laws, and online episode boundaries were not archived.

Results. Table 4 and Figure 5. Greedy has median longest post-hoc stretch 497 of 512 tokens, with 10/10 terminal stretches censored at the horizon. The windowed controllers reduce repetition (seq-rep-4 0.85/0.87, median longest post-hoc stretch 220/228). For the guard, 11 of 340 uncensored post-hoc stretches exceed 24 tokens (3.2%); 14 reach 24 tokens (4.1%), and one terminal stretch is censored. Neither fraction directly estimates the failure probability in Equation 12. The post-hoc tail slope -0.19 is compared descriptively with $\ln(1 - r_{L,\delta}) = -0.19$ in Figure 5b; the geometric curve is a reference, not a validated bound for this different diagnostic. The median longest post-hoc stretch is 28. Aggregate repetition remains higher than under the global baselines (seq-rep-4 0.61 vs. 0.00/0.31; final-96: 0.44 vs. 0.00/0.29). The n-gram ban forbids repeated 4-grams but spends 11$\times$ more KL; global temperature leaves a maximum post-hoc stretch of 333 tokens with 4/10 censored, versus 46 and 1/10 for the guard. The guard pays 26.3 nats per 512 tokens and lower surprisal than either baseline (1.29 vs. 1.85/1.97 bits). The L=16 setting yields seq-rep-4 0.56 at 42.1 nats. Direct certificate validation requires a new run recording the per-step logits, base laws, and online episode boundaries.

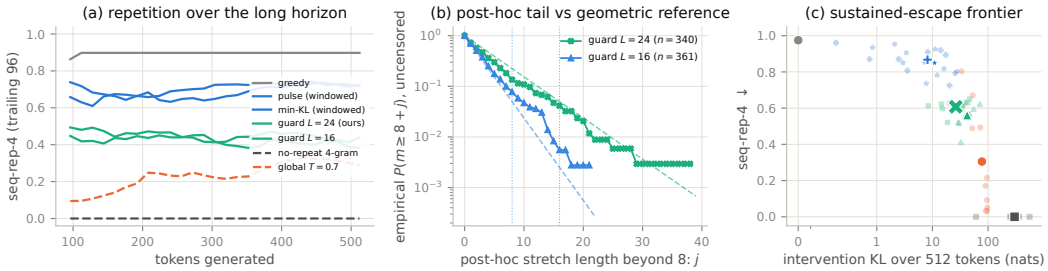


Figure 5: Sustained escape at 512 tokens (10 prompts per policy). (a) Repetition over the horizon. (b) Uncensored post-hoc repeated-stretch lengths m, not the online certificate event: empirical $\Pr(m \geq 8 + j)$ versus the geometric reference $(1 - r_{L,\delta})^j$ (dashed); dotted vertical lines mark $j = L - 8$. This offline rematch diagnostic is not a direct test of Equation 12. (c) Repetition against intervention KL.

Table 4: E3b: sustained containment over 512 new tokens (means over 10 prompts). Post-hoc stretch: median (max) of each continuation's longest post-hoc suffix-rematch diagnostic; cens.: terminal stretches censored at the horizon. These lengths are not the online controlled-trajectory variable in Equation 12.

policy	seq-rep-4 ↓	final-96	surprisal (bits)	KL (nats) ↓	post-hoc stretch	cens.
greedy	0.98	0.90	0.11	0	497 (507)	10/10
pulse, windowed	0.85	0.72	0.52	10.9	220 (454)	7/10
min-KL, windowed	0.87	0.72	0.47	8.17	228 (454)	9/10
guard L=24 (ours)	0.61	0.44	1.29	26.3	28 (46)	1/10
guard L=16	0.56	0.38	1.36	42.1	21 (29)	5/10
no-repeat 4-gram	0.00	0.00	1.85	300	0 (0)	0/10
global T=0.7	0.31	0.29	1.97	77.6	19 (333)	4/10

Transfer protocol and boundary (Table 1). The six modern configurations use raw census prefixes without chat templates. The horizon is fixed: turn-end tokens are not stop conditions, so trajectories continue past a completed answer. Gemma-4-12B-it is excluded as a boundary case: its answers typically end within $\sim$60 tokens, so this horizon is dominated by forced continuation, on which it degenerates on 140/140 trajectories (guard 0.56 at 43 nats).

Precision boundary for the two GPT-2 cohorts. The GPT-2 row in Table 1 and the GPT-2 census group in Table 5 are independent decodes of the same 28 prompts and two seeds with matched temperature, horizon, and matcher; the former uses float32 and the latter bfloat16. Changing inference precision produces different sampled paths in these archived runs: 20 of the 56 untreated prompt–seed pairs change at-risk classification (five enter and fifteen leave the at-risk set from bfloat16 to float32). Their risk-conditioned success rates and KL totals are therefore separate estimates and should not be compared directly.

Global baselines under identical accounting, and a natural-prompt cohort. Table 5 adds two comparisons on a self-contained single-platform cohort: the same harness drives the official Transformers GPT-2 checkpoint under its pre-registered bfloat16 configuration. Against the standard global heuristics at the deployment temperature, the ordering of E3a returns at controlled-trajectory scope: the guard delivers sustained success 1.00 over the at-risk trajectories at 1.27 nats per 256 tokens and 2.38 vs 2.32 nats/token of base-model surprisal, while the no-repeat-4-gram ban spends 11.27 nats and the repetition penalty 317.38 nats at roughly $+2$ nats/token. Both have no observed at-risk trajectories in this cohort, but only the hard ban excludes repeated 4-grams structurally; the finite repetition penalty retains support and its zero count is empirical. Their conditional sustained-success and post-grace repetition rates are consequently unestimable in this sample. The natural cohort retires the census-selection worry: on twelve-word prefixes of randomly drawn Wikipedia summaries (the list ships with the code), 28 of 80 untreated trajectories still enter a qualifying re-traversal at $T{=}0.8$. Sustained success is 0.86 without control and 0.96 with the guard at 2.08 nats (discordant 3/0, exact McNemar $p = 0.25$); this descriptive difference is not statistically resolved. Post-grace repetition (which does not require a prior departure) remains 0.57 $\rightarrow$ 0.57 (discordant 3/3, $p = 1.00$); the certificate controls each detected trajectory's length rather than promising that the evolving context will never be captured again. The cohort is platform-consistent by construction and is not pooled with the modern-checkpoint cohorts.

Table 5: Common baselines under identical KL accounting (GPT-2, bfloat16, $T{=}0.8$, 256 new tokens, 2 seeds). These policies are not tuned to a matched KL or risk budget. Left: the fixed census evaluation split; right: 40 twelve-word prefixes of randomly drawn Wikipedia summaries, each decoded with two seeds (80 trajectories per policy). Risk: trajectories whose matcher reaches $\ell_{\min}{=}8$. The repetition penalty $\theta{=}1.3$ (Keskar et al., 2019) has zero observed risk here, whereas the hard 4-gram ban excludes repeated 4-grams structurally; their conditional cells are unestimable because no at-risk trajectories were observed. KL: summed per-step divergence against $\mathrm{softmax}(z/0.8)$, in nats per trajectory for all policies. NLL: base-model surprisal at $T{=}1$, in nats/token. Post-g.: a qualifying rematch from $\min(t_{\text{risk}} + 12, 255)$ onward, without requiring an intervening departure. Brackets: 95% Wilson intervals.

	census evaluation split					natural prefixes (Wikipedia)				
policy	risk	sust.	post-g.	KL	NLL	risk	sust.	post-g.	KL	NLL
off	20	0.65 [.43,.82]	0.60 [.39,.78]	0.00	2.32	28	0.86 [.69,.94]	0.57 [.39,.73]	0.00	2.38
rep. penalty	0	–	–	317.38	4.36	0	–	–	294.90	4.44
no-repeat-4	0	–	–	11.27	2.50	0	–	–	16.10	2.54
guard (ours)	20	1.00 [.84,1.00]	0.50 [.30,.70]	1.27	2.38	28	0.96 [.82,.99]	0.57 [.39,.73]	2.08	2.39

F COMPATIBILITY BOUNDARY: INTENDED REPETITION

Protocol. We pair off and guard decoding on 60 instructions that explicitly require long verbatim repetition: ten each for refrains, Python code, checklists, Markdown tables, XML, and PlantUML. Qwen3-4B is sampled at $T{=}0.8$ with the frozen $(\ell_{\min}, L, \delta) = (8, 24, 0.05)$ guard and common per-task seeds. A committed task fingerprint (`65369515a3fa...`) fixes the suite. Arm labels are independently randomized A/B per task; Qwen3-14B scores task fidelity and readability from 1 to 5 without seeing the map, after which scores are unblinded and paired. The blind packet, map, raw judge records, and task fingerprint are archived.

Table 6: Blinded 5-point task-fidelity scores for intended-repetition tasks. Δ is guard minus off; ten paired tasks per class.

Requested artifact	off	guard	Δ
Refrain	4.6	2.2	-2.40
Markdown table	4.2	1.8	-2.40
Checklist	5.0	2.7	-2.30
Python code	3.8	3.0	-0.80
PlantUML	3.6	3.1	-0.50
XML	3.8	3.8	+0.00

Results. The guard-minus-off paired difference is -1.40 (bootstrap 95% CI [-2.00,-0.78]) for task fidelity and -1.32 ([-1.88,-0.72]) for readability. Table 6 localizes the boundary: refrains, tables, and checklists lose about 2.4 points, code loses 0.8, and XML is unchanged on average. The matcher triggers on 53/60 tasks; the guard is active for 738 generated steps (3.6%) and accumulates 36.3 nats mean KL per guarded task.

Why the automatic score is not evidence of compatibility. The strict exact-format checker reports 0/60 successes for *both* arms, hence zero measured "compatibility harm." This is a floor effect: the base model's reasoning wrappers, truncation, or formatting deviations make the checker fail even when the requested repeated body is visibly correct. Therefore `compatibility_harm=0` cannot be used to claim compatibility; the blinded paired ratings are the informative endpoint.

Paired examples.

- `refrain-00`: off emits eight exact copies of "We return by starlight."; guard mutates later copies to `We.return.by.starlight.`, `We$return by starlight.`, and "starlit" (fidelity 5 vs. 2).
- `checklist-00`: off emits eight exact `- [ ] verify checksum` lines; guard introduces `- [ } verify checksum`, punctuation, and a missing space (5 vs. 3).
- `table-00`: off supplies all eight identical data rows; guard introduces spacing mutations and truncates the eighth row (5 vs. 3).

These are not controller malfunctions: they are the observable cost of enforcing a content-agnostic controlled-trajectory certificate on a task whose intent requires violating that certificate.

G Natural-workload incidence and matched-prefix evaluation

Table 7 reports an unscreened, server-scale measurement of how often the frozen matcher reaches risk and cap thresholds under natural workloads. Incidence counts exact online re-traversal events under ordinary sampling.

Table 7: Natural-workload repetition incidence on Qwen3-4B. The fixed, unscreened cohort contains 1,500 prompts from five strata, each decoded with two seeds (3,000 prompt–trajectory pairs per arm), chat template, top-p 1.0, and an EOS-respecting 1,024-token horizon. Risk is defined as the exact matcher first reaching $\ell_{\min}=8$; cap exceedance is a trajectory reaching $L=24$. EOS and horizon columns are counts. Cap rates are percentages with 95% Wilson intervals; the greedy arm is a paired descriptive control, not a population-wide estimate.

		sampled ($T=0.8$)				greedy
stratum	n	risk n	cap % [95% CI]	EOS n	horizon n	cap % [95% CI]
Chat	600	478	10.7 [8.4,13.4]	122	478	22.0 [18.9,25.5]
Long-form	600	349	3.0 [1.9,4.7]	9	591	8.0 [6.1,10.4]
Summarization	600	431	0.5 [0.2,1.5]	600	0	0.7 [0.3,1.7]
QA	600	495	12.8 [10.4,15.7]	524	76	28.3 [24.9,32.1]
Code	600	597	32.2 [28.6,36.0]	43	557	45.0 [41.1,49.0]
All	3,000	2,350	11.8 [10.7,13.0]	1,298	1,702	20.8 [19.4,22.3]

Paired Stage B. To test conditional treatment power without prevalence dilution, we branched new continuations from the identical first-risk prefix for every 355 source exceedance and 355 preselected stratum-matched at-risk control; the cohort labels select prefixes, not replay outcomes. Across 710 pairs, cap exceedance falls from 218/710 (30.7%, Wilson 95% CI [27.4,34.2]%) off to 36/710 (5.1%, [3.7,6.9]%) guarded (improvement/harm discordants 191/9; two-sided exact McNemar

p=1.5×10^{-45}). This deliberately enriched conditional contrast does not revise the incidence estimate. It measures intervention efficacy at 52.96 nats mean accumulated KL per selected branch and 6.4% guard-active generated steps.

H E2 SURVIVAL AND EXIT-LOCATION DIAGNOSTICS

E2 Monte Carlo diagnostics. At $T \leq 0.7$, 115/126 uncensored exits (91%) occur at the predicted weakest-gap phase, versus a mean phase-uniform baseline of 13%. Across the 20 (orbit, temperature) conditions, the median sup-deviation between analytic and empirical survival is 0.05 (worst 0.14). The finite-horizon residual mass $S_T(48)$ matches the censored fraction at every temperature; at T=0.3, the exact/empirical pairs across the four orbits are 0.61/0.65; 0.99/1.00; 1.00/1.00; 1.00/1.00.

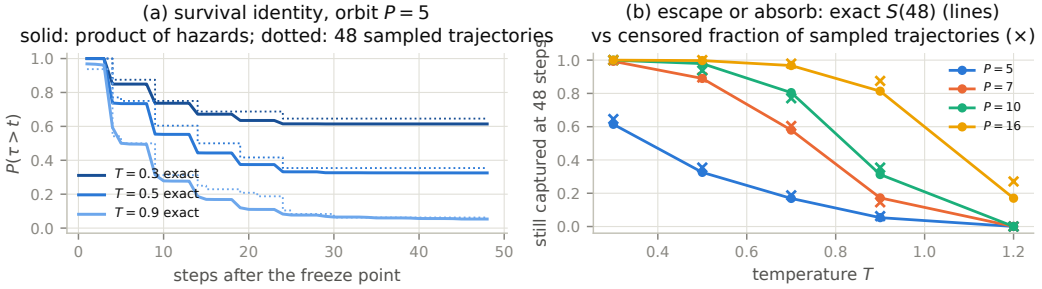


Figure 6: (a) Exact survival curves $S_T(t)$ (lines; Equation 4) against empirical survival over 48 sampled trajectories (dotted steps) for the P=5 orbit at three temperatures. (b) Finite-horizon residual mass $S_T(48)$ (lines) against the censored fraction of sampled trajectories (crosses).

I SCALE REPLICATION PROTOCOL

Hardware and backends. The nine dense-model configurations use HF `transformers` in bfloat16, exposing full logits during inference but archiving only the top 20 per orbit step, KV-cached greedy decoding (`do_sample=False`), and no cross-prompt batching. DeepSeek-V4-Flash-0731 (284B parameters, 43 MoE layers, 256 experts with 6 routed per token) is evaluated through `llama.cpp`; the runtime configuration pins the KV cache to f16 and disables flash attention. Community GGUF artifacts at ∼4.2 and ∼3.1 bits per weight provide the two quantization levels. The archived client records preserve the endpoint and build metadata but not immutable weight hashes; the release workflow therefore requires matching local GGUF files and host-side verification as specified in `experiments/E5/E5.json`.

Prompting and horizons. Raw text continuation with the same 100 openers for every model, including instruct-tuned ones (no chat template); census horizon 256 tokens with $P \leq 64$; reference windows re-decoded to onset + $3P$ + 48, with t_{ref} at onset + $3P$; models with zero captures are re-decoded at 512 tokens on the twelve archived early-onset GPT-2 prompts fixed in `experiments/E5/E5.json` before a zero is accepted.

Recorded checks. (G1) Regressing $\ln \widehat{q}$ on Δ at T=0.5 over census rows with $0.5 < \Delta < 8$ meets the 10% relative-error gate for all configurations; the largest observed slope error relative to $-1/T$ is 1.28%. This is an empirical diagnostic, not an exact regression identity: the hazard prefactor varies across logit rows. (G2) Capture is measured at 256 tokens; zero-capture models receive the 512-token follow-up. (G3) The recorded orbit criteria require at least 60% positive gap-growth slopes, median slope/Λ in $[0.8, 1.6]$, and mean localization at least 3× the phase-uniform baseline. (G4) Onset AUROC is compared with continuation AUROC. These are not data-inclusion requirements: all configurations are reported. G3 is inapplicable without captured orbits and fails localization on Gemma-4-12B; G4 fails on both Qwen3 configurations and both V4-Flash quantization levels. Client preflight checks duplicate-trajectory determinism, string/token-id prompt agreement, and pre-sampling top-logprobs. The host operator separately verifies weight files, the llama.cpp build, f16 KV, disabled flash attention, and absence of swap traffic, as specified in `experiments/E5/E5.json`.

Orbit accounting. All eleven server configurations archive top-20 logprobs $u_{t,(i)}$ per orbit step. The top two entries determine the exact runner-up gap and hence Equation 3. The analysis renormalizes

the retained entries at each temperature:

$$\widehat{q}_t(T) = 1 - \frac{e^{u_{t,(1)}/T}}{\sum_{i=1}^{20} e^{u_{t,(i)}/T}}, \qquad \rho_{20}(T) = 1 - \prod_{j=1}^{P}(1 - \widehat{q}_j(T)).$$

Omitting the tail gives $\widehat{q}_t(T) \le q_t(T)$. Server slopes fit P/ρ_{20}, an upper bound on the corresponding frozen mean, not the exact mean in Equation 5. Localization uses 400 hazard simulations from the stored window at the first grid temperature with at least 40 expected exits; these are not independent model rollouts. These retained-row calculations check low-temperature scaling, whose activation gap is preserved, but do not constitute full-vocabulary finite-temperature validation. GPT-2 E2 instead uses full rows and independent model sampling.

Entity accounting. HF entity scores use full-softmax target-token probabilities. The MoE endpoint supplies top-20 logprobs; a target outside the top 20 receives the 20th entry as an upper bound on its logprob (marked censored), so its string cost is a lower bound. At the IQ4 quantization level, the first name token is censored for every referent, real or fabricated. Beyond it, 1.2% of real-name tokens and 85.2% of fabricated-name tokens are censored. Overall censoring is 42.1% vs. 87.8% when pooling tokens, or 46.2% vs. 88.5% when averaging per-name fractions (real vs. fabricated). AUROCs computed on these censored values are estimates: censoring compresses the fabricated tail, while ties involving censored real-name tokens can push comparisons the other way.

Paired quantization-level design. Both quantization levels undergo the identical suite on the identical prompts; the paired comparison in Section 4.6 uses within-prompt trajectory agreement plus the per-configuration observables, with unquantized anchors provided by the nine bfloat16 configurations and the fp32 GPT-2 stack. Every experimental unit is checkpointed and re-entrant, and `paper/scripts/cross_model_table.py` regenerates Table 2 from the checkpoints.

J Case study: spike and capture at hallucination onset

Entity benchmark. Real referents are 60 historical and contemporary public figures with pre-2019 fame (the GPT-2 training cutoff), spanning science, arts, politics, and sport. Fabricated referents are 60 two-part names sampled from a syllable grammar with seed 99 (examples: *Kryvdri Halbrec*, *Tvielgu Trevik*). In the archived GPT-2 prompt frame, fabricated names occupy 4–8 BPE tokens and real names 1–6; the groups are not token-length matched. Lists and the generator ship in `arrhenius/src/arrhenius/utils.py`. The prompt frame conditions every name token on a preceding context ("According to the encyclopedia,"), so the string cost $C_{\text{name}} := -\sum_{t \in \text{name}} \log_2 \pi_t(w_t)$ is well defined for the full name and is reported in bits.

Protocol. Each name is placed in the frame "According to the encyclopedia, {name} is best known for", followed by a greedy 30-token continuation. Onset signals: the name's per-token string cost and its maximum token surprisal. Continuation signals: mean surprisal and mean entropy over the entire 30-token suffix, not conditioned on detected verbatim capture. The archived `capture_steps` field is a surprisal-threshold proxy, not an exact-cycle detection time, and is not used for the comparisons below. AUROC separates fabricated from real referents over all 60×60 positive–negative item pairs.

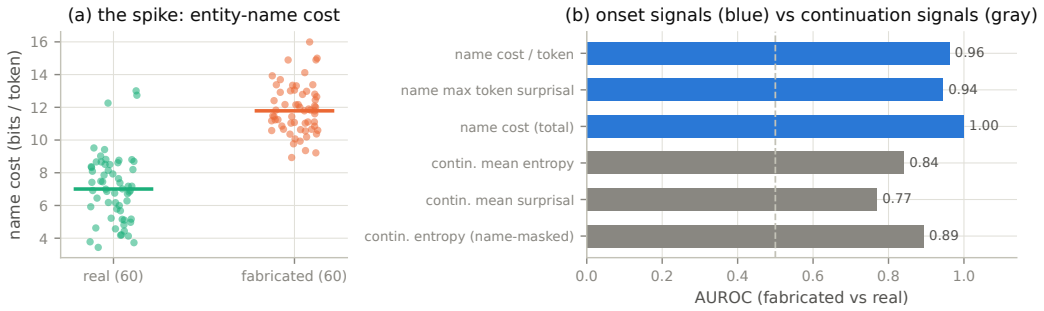


Figure 7: Entity diagnostics on GPT-2 (60 real, 60 fabricated referents). (a) Entity-name cost per token. (b) AUROC of onset signals (blue) versus whole-continuation signals (gray); the dashed line is chance. Continuation statistics are not conditioned on detected verbatim capture.

Results. Fabricated names cost 11.9 bits per token against 7.1 for real names (70 vs 17 bits for the whole string), and onset AUROCs reach 0.96 (per-token cost; 100% recall at 10% false positives) and 0.94 (max token surprisal; 88% recall). The total-cost AUROC of 1.00 partly reflects fabricated names' longer tokenizations, so the per-token statistic is primary. Continuation signals are consistently weaker on GPT-2: surprisal 0.77 and entropy 0.84, which at the same 10%-false-positive operating point recover only 55% and 70% of fabricated referents (Figure 7); fabricated and real continuations have surprisal 2.40 and 2.05 bits/token, respectively. Fabricated continuations re-mention their entity in 8.9% of tokens (real: 2.8%); masking these low-surprisal re-mentions *raises* the continuation AUROCs to 0.85/0.89, so the observed downstream discrimination is not solely due to these copies. Generic biography templates recur in the qualitative examples below; a basin interpretation is a hypothesis, not a demonstrated mechanism or an instance of the verbatim-capture certificate. Section 4.6 retests the asymmetry on eleven further configurations: the onset signal survives everywhere; the downstream deficit does not.

Fabricated referents. High-cost names followed by native-confidence continuations in a generic biography channel: "*Tvielgu Trevik* is best known for his role as the 'father of the Tvielgu' in the original Star Trek series"; "*Kryvdri Halbrec* is best known for his role as the 'father of the modern Russian mafia' "; "*Lompgu Durnov* is best known for his role as the 'father of the Soviet Union's first nuclear weapons program' ". The 'father of X' template recurs across referents; this semantic regularity is not evidence of an exact verbatim-repetition trajectory. Full entity generations ship in `results/`.