Speculative Decoding Is Per-Token, Not Path

Songyan Dai [1,2]
[1]School of Physics, Zhejiang University, Hangzhou, China
[2]Chu Kochen Honors College, Zhejiang University, Hangzhou, China
soyonagasaki698@gmail.com

Abstract

Standard TokenVerify accepts draft tokens through sequential clipped-ratio tests. We study how its per-token rate $\sigma_L(P,Q)$ and path-level overlap $\Omega_L(P,Q) = \sum_{t=1}^{L} \mathrm{ov}(P_{\le t}, Q_{\le t})$ govern target-verification-round complexity. We show $\sigma_L \le \Omega_L$ always, with strict inequality already on i.i.d. products, and construct binary product laws with fixed-precision conditionals for which $\sigma_L \le 2$ while $\Omega_L \to \infty$. An explicit binary trajectory-exact rejection-style protocol retaining the LKM residual on $L = 2$ achieves $\mathbb{E}[\tau] = 19/16 > 9/8 = \sigma_L$, so σ_L is not a universal upper bound. Inside the admissible Leviathan-style clipped-ratio class, σ_L is tight, and the speculative decoding complexity satisfies $\mathrm{SDC}^{\mathcal{Q}}_{B,L}(M;u,T) = \Theta(T/(\sigma_{L,B}(P_M^{u,T}, \mathcal{Q}) + 1))$ under exact proposal optimization, a positive reachable-rate floor, bounded reachable-rate heterogeneity, and $T \ge L$. On the path side, $\sum_t \sqrt{\eta_t/2} = o(L)$ for the prefix-KL profile forces $\Omega_L = L - o(L)$, whereas geometrically bounded overlap $o_t \le \rho^t$, $\rho < 1$, forces $\Omega_L \le 1/(1-\rho)$, with no unconditional KL converse. With precision $b = \max\{6, \lceil \log_2(2L) \rceil\}$, two explicit Transformer families of $O(\log L)$-bit description length realize a $\Theta(L)$ separation in verification rounds under the same three-element proposal family: a constant-readout family has $\mathrm{SDC} = O(T/L)$, while an alternating-parity family has $\mathrm{SDC} = \Omega(T)$, for $L2^{-b} \le \varepsilon \le 1/2$ and $T \ge 2L$. These results distinguish sequential acceptance from path coupling and show how the target/proposal pair controls the achievable rate.

1 Introduction

What controls the verification-round speedup of exact speculative decoding? We compare the per-token rate with the path-level overlap and analyze their implications for protocol-dependent decoding complexity. Speculative decoding [Leviathan et al., 2023, Chen et al., 2023] drafts L proposal tokens with a cheap model and verifies them against the target in a single forward pass. In token verification, exact correctness follows from a per-token clipped-ratio test and a residual sample at the first rejection [Leviathan et al., 2023, Alg. 1, App. A.1]. The quantity studied here is the speculative decoding complexity (SDC): the expected number of sequential target-verification rounds required to produce T exact tokens. We ask which rate governs SDC for this protocol class and how it differs from the path-level coupling bound.

Two candidates. The first is the per-token rate (the Leviathan rate)

$$\sigma_L(P,Q) = \sum_{t=1}^{L} \mathbb{E}_{\tilde{y}\sim Q}\Big[\prod_{s=1}^{t} \min\Big(1, \tfrac{P_{s|\tilde{y}_{<s}}(\tilde{y}_s)}{Q_{s|\tilde{y}_{<s}}(\tilde{y}_s)}\Big)\Big], \tag{1}$$

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

which equals the expected accepted prefix length of the standard one-step clipped-ratio protocol [Leviathan et al., 2023, Alg. 1]. The second is the path-level overlap (a coupling ceiling)

$$\Omega_L(P,Q) = \sum_{t=1}^{L} \mathrm{ov}(P_{\le t}, Q_{\le t}) = \sum_{t=1}^{L} \Big(1 - \mathrm{TV}(P_{\le t}, Q_{\le t})\Big), \tag{2}$$

which sums the overlaps of the prefix marginals over depths $1, \dots, L$ [Lindvall, 2002, Thm. 5.2] and upper-bounds the expected common-prefix length of a coupling. It also upper-bounds the expected accepted length of any trajectory-exact rejection-style round in our setting.

Per-token rate and path-level overlap differ. We show that $\sigma_L \le \Omega_L$ always, with strict inequality in general (Propositions 1 and 2). Even for i.i.d. products, the two quantities differ. An explicit binary trajectory-exact rejection-style protocol retaining the LKM residual achieves $\mathbb{E}[\tau] = 19/16 > 9/8 = \sigma_L(P,Q)$ on $L = 2$ (Proposition 3). Thus $\mathbb{E}[\tau] \le \sigma_L$ does not follow from trajectory exactness alone. It holds inside the admissible Leviathan-style clipped-ratio class of Definition 2, where TokenVerify attains equality. Outside that class, Ω_L remains an upper bound.

How big is the gap? In Proposition 2, $p = (0.9, 0.1)$ and $q = (0.1, 0.9)$ give $\sigma_2 = 0.24$ and $\Omega_2 = 0.40$, establishing strict inequality. In the separate fixed-precision instance of Proposition 3, the non-admissible protocol attains $\mathbb{E}[\tau] = \Omega_2 = 19/16 > 9/8 = \sigma_2$. To establish an unbounded ratio of the complete sums, Proposition 4 in section A constructs staged binary product laws with fixed-precision conditional probabilities and lengths $L_J = 2^{J+1} + 2J - 2$, for which $\sigma_{L_J} \le 2$ while $\Omega_{L_J} \ge (3/2)^J/(2\sqrt{J}) \to \infty$.

Tight SDC characterization. Within the admissible class, the SDC bounds depend on the per-token rate and its variation across reachable states. Writing $\sigma_{L,B}(P,\mathcal{Q})$ for the branched per-token rate (supremum over B-branch proposal mixtures), we prove

$$\mathrm{SDC}^{\mathcal{Q}}_{B,L}(M;u,T) \;=\; \Theta\Big(\tfrac{T}{\sigma_{L,B}(P_M^{u,T},\mathcal{Q})+1}\Big) \tag{3}$$

under a positive uniform per-token-rate floor, bounded reachable-rate heterogeneity $(\sigma^* + 1)/(\sigma_* + 1) = O(1)$, exact proposal optimization, and $T \ge L$ (Theorem 3). The lower bound is Wald applied to the per-round emitted length $T'_r = \tau_r + \mathbf{1}\{\tau_r < L\}$; the upper bound is achieved by the iterated Leviathan decoder. The additive 1 bounds the residual-token contribution to the expected progress of a round.

KL phase transition for Ω_L. We give sufficient conditions for linear and bounded path-level overlap. Write $\eta_t = D_{\mathrm{KL}}(P_{\le t} \| Q_{\le t})$ for the prefix-KL profile. If $\sum_{t=1}^{L} \sqrt{\eta_t/2} = o(L)$, then $\Omega_L(P,Q) = L - o(L)$ and the effective speculation length $\Omega_L/L \to 1$ (Theorem 4). If the prefix overlap is geometrically bounded $o_t \le \rho^t$ for some $0 \le \rho < 1$, then $\Omega_L(P,Q) \le 1/(1-\rho)$ uniformly in L (Theorem 5). We further prove that no unconditional KL-only converse exists: a fixed-precision counterexample with infinite per-prefix KL but linear Ω_L rules out a one-sided KL "if and only if" (Theorem 6).

Architecture-dependent $\Theta(L)$ separation. We give two explicit fixed-precision Transformer workload families realizing the extreme regimes under the same proposal family $\mathcal{Q} = \{Q^{\mathrm{flat}}, Q_0^{\mathrm{const}}, Q_1^{\mathrm{const}}\}$ (uniform plus two deterministic constants). The positive family $\mathcal{F}^{+}_{a,L}$ has constant-readout Transformers emitting a with probability $1 - 2^{-b}$ at every step; under Q_a^{const} its per-token rate is at least $(1-\varepsilon)L$, giving $\mathrm{SDC} = O(T/L)$. The negative family $\mathcal{F}^{-}_{w,L}$ has alternating-parity Transformers of constant hidden dimension realizing $Y_n = w \oplus (n \bmod 2)$; under any $Q \in \mathcal{Q}$ its per-token rate is at most 1, giving $\mathrm{SDC} \ge T/2$. The speedup ratio is $\Theta(L)$ versus $O(1)$, a $\Theta(L)$ separation under the same proposal family for $L2^{-b} \le \varepsilon \le 1/2$ and $T \ge 2L$ (Theorem 7).

Why this matters. A rate estimate must reflect the verification rule. Inside the admissible class, σ_L controls the accepted-prefix length and yields the SDC characterization under the stated reachable-state conditions. The larger Ω_L is a coupling ceiling and can exceed σ_L by an unbounded factor. The architecture separation shows that two small Transformers can differ by $\Theta(L)$ in verification rounds under the same proposal family.

Comparison with representation succinctness. Bergsträßer et al. [2025] study the descriptional size of Transformers; section 7 contrasts this with the verification cost studied here.

Outline. section 2 fixes notation. section 3 proves the per-token vs. path-level dichotomy. section 4 proves the tight SDC characterization. section 5 proves the KL phase transition. section 6 proves the architecture separation. section 7 discusses implications. Complete proofs are in the appendix.

2 Preliminaries

Fixed precision. The probabilistic analysis applies to arbitrary finite distributions. Fixed precision specifies the explicit probability constructions and Transformer realizations. Fix a finite alphabet Σ, integer precision $b \geq 6$, and a constant dynamic-range bound $R \geq 2$. The model arithmetic uses $D_b = \{m2^{-b} : m \in \mathbb{Z},\ |m| \leq R2^b\}$, with deterministic rounding ρ_b after each operation. The range accommodates the coefficient -2 and intermediate value 2 in our constructions. Fixed precision constrains the explicit model parameters and next-token probability tables; path probabilities, expectations, and auxiliary verifier randomization are interpreted exactly. The atom probabilities in Proposition 3 are representable already in D_4, and hence in every permitted D_b. A decoder-only Transformer M [Vaswani et al., 2017] with self-attention layers and softmax attention defines a deterministic next-token distribution $p_M(\cdot \mid u) \in \Delta(\Sigma)$ for any prefix $u \in \Sigma^*$.

Path laws and prefix marginals. For a context u and horizon T, the target path law is $P_M^{u,T}(y_{1:T}) = \prod_{t=1}^{T} p_M(y_t \mid u, y_{<t})$. For a path law P on Σ^L, write $P_{\leq t}$ for its prefix marginal on Σ^t and $P_{s|z_{<s}}(z_s)$ for the conditional next-token mass at position s given prefix $z_{<s}$. For a path law with horizon $T \geq L$, σ_L acts on its first L coordinates.

Proposal family and branching. A proposal family $\mathcal{Q}$ is a finite set of distributions on Σ^L. A B-branch mixture $Q^{\mathrm{mix}} = \sum_{i=1}^{B} \alpha_i Q_i$ with $Q_i \in \mathcal{Q}$ and $\alpha \in \Delta_{B-1}$ extends the family. We write $\mathcal{Q}_B$ for the set of B-branch mixtures.

Rejection-style rounds. Fix a target $P \in \Delta(\Sigma^L)$ and a mixture $Q^{\mathrm{mix}} \in \mathcal{Q}_B$. A *rejection-style round* samples a proposal trajectory $\widetilde{y}_{1:L} \sim Q^{\mathrm{mix}}$, queries the target conditionals along the proposal path with one verification call, chooses an accepted-prefix length $\tau \in \{0, \ldots, L\}$ via measurable acceptance decisions, sets $Y_{1:\tau} = \widetilde{y}_{1:\tau}$, and if $\tau < L$ samples a residual token at position $\tau + 1$ from the LKM-residual law $\mathrm{norm}(\max(0, P_{\tau+1|\widetilde{y}_{<\tau+1}} - Q^{\mathrm{mix}}_{\tau+1|\widetilde{y}_{<\tau+1}}))$. The emitted length is $T' = \tau + \mathbf{1}\{\tau < L\} \in \{1, \ldots, L\}$. We use capped rounds: when all draft tokens are accepted, the round emits L tokens and omits the additional bonus token in the standard algorithm [Leviathan et al., 2023, Alg. 1]. We use TokenVerify for the capped version of that algorithm, with clipped-ratio acceptance and the LKM residual at the first rejection. The iterated Leviathan decoder applies TokenVerify in each round.

Trajectory exactness. A round is *trajectory-exact* if the joint law of $(T', Y_{1:T'})$ matches that of $(\mathrm{stop}, Z_{1:\mathrm{stop}})$ for some $Z \sim P$ and some stopping rule stop. Equivalently, the round implements L token samples from P truncated at stop tokens.

Two complexity measures. The per-token rate $\sigma_L(P, Q^{\mathrm{mix}})$ is defined in (1); the branched rate is $\sigma_{L,B}(P, \mathcal{Q}) = \sup_{Q^{\mathrm{mix}} \in \mathcal{Q}_B} \sigma_L(P, Q^{\mathrm{mix}})$. The path-level overlap $\Omega_L(P, Q^{\mathrm{mix}})$ is defined in (2) via $\mathrm{ov}(P_{\leq t}, Q^{\mathrm{mix}}_{\leq t}) = \sum_{z \in \Sigma^t} \min(P_{\leq t}(z), Q^{\mathrm{mix}}_{\leq t}(z)) = 1 - \mathrm{TV}(P_{\leq t}, Q^{\mathrm{mix}}_{\leq t})$, with branched version $\Omega_{L,B}(P, \mathcal{Q}) = \sup_{Q^{\mathrm{mix}} \in \mathcal{Q}_B} \Omega_L(P, Q^{\mathrm{mix}})$.

Multi-round decoder and SDC. A multi-round speculative decoder $\Phi = (\Pi_1, \Pi_2, \ldots)$ runs an adapted sequence of rejection-style rounds on the local target $P_M^{uY^{<r},L}$ where $Y^{<r}$ is the prefix emitted before round r. Let T'_r be the emitted length of round r, $S_n = \sum_{r=1}^{n} T'_r$, and $N = \inf\{n : S_n \geq T\}$. The decoder halts at round N and outputs the first T tokens of the untruncated emitted prefix. Φ is multi-round trajectory-exact if its length-T output has law $P_M^{u,T}$.

Definition 1 (SDC). $\mathrm{SDC}^{\mathcal{Q}}_{B,L}(M;u,T) = \inf_{\Phi} \mathbb{E}_{\Phi}[N]$, where the infimum ranges over multi-round trajectory-exact decoders using $\mathcal{Q}_B$ proposals whose every round is admissible in the sense of Definition 2, conditionally on the preceding rounds.

3 Path-Level Overcounts: Why $\sigma_L < \Omega_L$

We begin by comparing the two quantities and identifying the protocol restriction under which the per-token rate is tight.

Proposition 1 (Path-Level Dominates Per-Token). *For every target $P \in \Delta(\Sigma^L)$ and proposal $Q \in \Delta(\Sigma^L)$, $\sigma_L(P,Q) \le \Omega_L(P,Q)$.*

Proof sketch. Multiplying the per-step clipped ratio $\min(1, P_s/Q_s)$ by $Q_{\le t}$ and summing over $z \in \Sigma^t$ gives a quantity $\le \sum_z \min(P_{\le t}(z), Q_{\le t}(z)) = \mathrm{ov}(P_{\le t}, Q_{\le t})$; summing over t gives $\sigma_L \le \Omega_L$. Full proof in section A. □

Proposition 2 (Strict gap, even on i.i.d. products). *There exist factorized $P = p^{\otimes L}$ and $Q = q^{\otimes L}$ with $p, q \in \Delta(\Sigma)$ such that $\sigma_L(P,Q) < \Omega_L(P,Q)$. For $\Sigma = \{0,1\}$, $p = (0.9, 0.1)$, $q = (0.1, 0.9)$, and $L = 2$, the $t = 2$ summand of σ_L is $\mathrm{ov}(p,q)^2 = 0.04$, whereas $\mathrm{ov}(p^{\otimes 2}, q^{\otimes 2}) = 0.20$. Consequently $\sigma_2(P,Q) = 0.24 < 0.40 = \Omega_2(P,Q)$.*

Proposition 2 shows that the two quantities can differ even for i.i.d. product laws. For example, the string 01 has probability 0.09 under both product laws and contributes 0.09 to the two-token path overlap. Its contribution to TokenVerify's two-token survival mass is $0.09 \times 1 \times (1/9) = 0.01$. Proposition 4 in section A gives $\sigma_{L_J} \le 2$ and $\Omega_{L_J} \ge (3/2)^J/(2\sqrt{J})$, so the ratio of the complete sums is unbounded. Proposition 3 constructs a trajectory-exact rejection-style protocol with expected accepted length exceeding σ_L. The admissible class defined below captures the clipped-ratio survival bound attained by TokenVerify.

Definition 2 (Admissible Leviathan-style clipped-ratio class). A rejection-style round on local target P and mixture Q^{mix} is *admissible* if for every $t \le L$ and every proposal prefix $z_{1:t} \in \Sigma^t$ with positive $Q^{\mathrm{mix}}_{\le t}$-mass,

$$\Pr(\tau \ge t \mid \widetilde{y}_{1:t} = z_{1:t}) \;\le\; \prod_{s=1}^{t} \min\Big(1, \tfrac{P_{s|z_{<s}}(z_s)}{Q^{\mathrm{mix}}_{s|z_{<s}}(z_s)}\Big). \tag{4}$$

The standard Leviathan protocol [Leviathan et al., 2023] is admissible with equality in (4).

For a non-anticipating verifier, the acceptance probability at position s is conditioned on the current proposal prefix and acceptance of all preceding tokens. With auxiliary randomness independent of drafting, the clipped-ratio bounds on these probabilities imply (4) by the chain rule. TokenVerify attains equality. The condition defines admissibility through the resulting conditional survival probabilities.

Proposition 3 (σ_L Is Not Universal). *For every integer precision $b \ge 6$, there exists an explicit binary instance with $L = 2$, one proposal branch, and a trajectory-exact (but not admissible in the sense of Definition 2) rejection-style protocol whose accepted prefix length τ satisfies*

$$\mathbb{E}[\tau] \;=\; \tfrac{19}{16} \;>\; \tfrac{9}{8} \;=\; \sigma_L(P,Q).$$

The construction is in section A: it uses correlated cross-position acceptance decisions on the proposal $\widetilde{y} \in \{00, 01, 10, 11\}$ that preserve trajectory exactness via a randomized stopping rule. The decision to accept the first token depends on the second draft token. On the proposal 00, the probability of accepting both tokens is 1, exceeding the clipped-ratio product $2/3$ in Definition 2.

Outside the admissible class, the path-level bound still holds: since $\tau \ge t$ forces $T' \ge t$ and $Y_{1:t} = \widetilde{y}_{1:t}$, trajectory exactness gives

$$\Pr(\tau \ge t, \widetilde{y}_{1:t} = z) \le \min\{Q_{\le t}(z), P_{\le t}(z)\}.$$

Summing over z gives $\Pr(\tau \ge t) \le \mathrm{ov}(P_{\le t}, Q_{\le t})$, and summing over t gives $\mathbb{E}[\tau] \le \Omega_L(P,Q)$.

Remark 1 (Class scope). Proposition 3 explains why the tight SDC characterization in section 4 is stated for the admissible class of Definition 2. Within that class, σ_L is the tight per-token rate. Across the larger class of full-lookahead trajectory-exact protocols, Ω_L is an upper bound, which can be strictly larger by Proposition 2.

4 Per-Token Is Tight: SDC $= \Theta(T/(\sigma_L + 1))$

We now restrict to the admissible class of Definition 2 and prove the tight characterization (3).

Lemma 1 (Per-round Leviathan upper bound)**.** *For every admissible rejection-style round on local target P and mixture $Q^{\mathrm{mix}} \in \mathcal{Q}_B$,*

$$\mathbb{E}[\tau] \le \sigma_L(P, Q^{\mathrm{mix}}), \qquad \mathbb{E}[T'] = \mathbb{E}[\tau] + \Pr(\tau < L) \le \sigma_{L,B}(P, \mathcal{Q}) + 1.$$

The standard Leviathan protocol attains equality in the first bound for a fixed mixture. If the selected mixture has rate within $\varepsilon_b \ge 0$ of the supremum at every reachable state, the iterated decoder attains that rate up to this uniform proposal-optimization tolerance. No quantitative dependence of ε_b on b is assumed.

Proof sketch. Take expectations in (4) over $\widetilde{y} \sim Q^{\mathrm{mix}}$ and sum over $t \le L$ to get $\mathbb{E}[\tau] \le \sigma_L(P, Q^{\mathrm{mix}})$. Optimizing over mixtures gives $\sigma_{L,B}$. The $+1$ comes from $T' = \tau + \mathbf{1}\{\tau < L\}$ and is the residual sample at the first rejection; it is necessary for exactness when $\tau < L$. For Leviathan, the inequality in (4) is equality, so all averaging steps are equalities. Full proof in section B. □

Lemma 2 (Wald identity for stopped sums)**.** *Let $(\mathcal{F}_r)$ be the natural round filtration, X_r nonnegative bounded $\mathcal{F}_r$-measurable, and N an integrable stopping time. If $a \le \mathbb{E}[X_r \mid \mathcal{F}_{r-1}] \le b$ on $\{N \ge r\}$, then $a\,\mathbb{E}[N] \le \mathbb{E}[\sum_{r=1}^{N} X_r] \le b\,\mathbb{E}[N]$ [Williams, 1991, §10.10].*

Definition 3 (Uniform per-token-rate floor)**.** Let $\mathcal{R}(u, T)$ be the set of target-reachable prefixes of length less than T. Define $\sigma_* = \inf_{v \in \mathcal{R}(u,T)} \sigma_{L,B}(P_M^{uv,L}, \mathcal{Q})$ and $\sigma^* = \sup_{v \in \mathcal{R}(u,T)} \sigma_{L,B}(P_M^{uv,L}, \mathcal{Q})$. We assume $\sigma_* > 0$ throughout this section. The hypothesis holds when $\mathcal{Q}$ contains a baseline proposal with uniformly positive overlap against every reachable local target.

Theorem 1 (SDC lower bound)**.** *For every multi-round trajectory-exact admissible decoder Φ, $\mathbb{E}_\Phi[N] \ge T/(\sigma^* + 1)$. Hence $\mathrm{SDC}_{B,L}^{\mathcal{Q}}(M; u, T) \ge T/(\sigma^* + 1)$.*

Proof sketch. Lemma 1 gives $\mathbb{E}[T'_r \mid \mathcal{F}_{r-1}] \le \sigma^* + 1$ on $\{N \ge r\}$ (the local rate is at most σ^* on every reachable prefix). Lemma 2 with $X_r = T'_r$ and $b = \sigma^* + 1$ then gives $T \le \mathbb{E}[S_N] \le (\sigma^* + 1)\mathbb{E}[N]$. Full proof in section B. □

Theorem 2 (SDC upper bound, iterated Leviathan)**.** *Suppose $\sigma_* > 0$ and the uniform proposal-optimization tolerance satisfies $0 \le \varepsilon_b < \sigma_*$. The iterated Leviathan decoder satisfies*

$$\mathrm{SDC}_{B,L}^{\mathcal{Q}}(M; u, T) \;\le\; \frac{T + L - 1}{\max\{1, \sigma_* - \varepsilon_b\}}.$$

For fixed L, exact optimization, and $T \to \infty$, $\mathrm{SDC}_{B,L}^{\mathcal{Q}}(M; u, T) \le T/\sigma_*\,(1 + o(1))$.

Here $\bar{\sigma}$ is the rate at the initial context, and σ_* and σ^* are its extrema over reachable prefixes. Under the hypotheses below, the bounds match up to constant factors when $(\sigma^* + 1)/(\sigma_* + 1)$ is uniformly bounded.

Theorem 3 (Tight Per-Token SDC)**.** *Under Definition 3 with $T \ge L$ and exact optimization,*

$$\frac{\sigma_* + 1}{\sigma^* + 1}\,\frac{T}{\bar{\sigma} + 1} \;\le\; \mathrm{SDC}_{B,L}^{\mathcal{Q}}(M; u, T) \;\le\; 4\,\frac{\sigma^* + 1}{\sigma_* + 1}\,\frac{T}{\bar{\sigma} + 1},$$

where $\bar{\sigma} = \sigma_{L,B}(P_M^{u,T}, \mathcal{Q})$. If $(\sigma^* + 1)/(\sigma_* + 1) = O(1)$, these bounds yield the uniform characterization $\mathrm{SDC}_{B,L}^{\mathcal{Q}}(M; u, T) = \Theta\big(T/(\sigma_{L,B}(P_M^{u,T}, \mathcal{Q}) + 1)\big)$ with constants independent of L, T when the reachable-rate ratio is uniformly bounded.

The "+1" in the denominator is the residual-token contribution to $\mathbb{E}[T']$, present whenever $\tau < L$ and absent when $\tau = L$. The explicit bounds hold without bounded heterogeneity; a uniform Θ statement additionally requires $(\sigma^* + 1)/(\sigma_* + 1) = O(1)$.

5 When the Path-Level Overlap Is Linear: A KL Phase Transition

The path-level overlap Ω_L is the universal upper bound across all trajectory-exact rejection-style protocols (section 3) and an upper bound on the per-token rate σ_L (Proposition 1). The following

conditions describe linear and bounded regimes of the path-level overlap. A uniform bound on Ω_L over reachable local targets and available proposals gives an $\Omega(T)$ lower bound on target-verification rounds throughout the trajectory-exact rejection-style class with one proposal trajectory per round. Proposition 4 shows that Ω_L can grow while the per-token rate remains bounded. Corollary 1 gives a uniform local-overlap condition yielding $\sigma_L = \Theta(L)$ and $O(T/L)$ verification rounds.

Definition 4 (Prefix KL and overlap profiles)**.** For $1 \le t \le L$, the prefix-KL profile is $\eta_t = D_{\mathrm{KL}}(P_{\le t} \| Q_{\le t})$ and the prefix-overlap profile is $o_t = \mathrm{ov}(P_{\le t}, Q_{\le t}) = 1 - \mathrm{TV}(P_{\le t}, Q_{\le t})$. The effective speculation length is $L_{\mathrm{eff}}(P, Q) = \Omega_L(P, Q)/L \in [0, 1]$.

Lemma 3 (Pinsker on prefix marginals)**.** *With KL divergence computed using natural logarithms,* $\mathrm{TV}(P_{\le t}, Q_{\le t}) \le \sqrt{\eta_t/2}$*, and consequently* $o_t \ge 1 - \sqrt{\eta_t/2}$ *[Cover, 1999, Thm. 11.6.1].*

Theorem 4 (Positive Pinsker Phase)**.** *If* $\sum_{t=1}^{L} \sqrt{\eta_t/2} \le L/2$*, then* $\Omega_L(P, Q) \ge L/2$*. More generally, if* $\sum_{t=1}^{L} \sqrt{\eta_t/2} = o(L)$ *along a sequence of instances, then* $\Omega_L(P, Q) = L - o(L)$ *and* $L_{\mathrm{eff}}(P, Q) \to 1$*. Under the linear-accumulation profile* $\eta_t \le t\delta$*, the hypothesis holds whenever* $L \le 1/(2\delta)$ *for* $\delta > 0$*.*

Theorem 5 (Negative Geometric Phase)**.** *If* $o_t \le \rho^t$ *for some* $0 \le \rho < 1$ *and all* $1 \le t \le L$*, then* $\Omega_L(P, Q) \le 1/(1 - \rho)$ *uniformly in* L*, and hence* $L_{\mathrm{eff}}(P, Q) \to 0$ *as* $L \to \infty$*.*

Theorem 6 (No unconditional KL converse)**.** *There is no unconditional implication from linear* Ω_L *to small one-sided prefix KL. Concretely, for every* $\varepsilon \in (0, 1)$ *representable in* D_b *and every* L*, the path laws* $P(0^L) = 1 - \varepsilon$*,* $P(1^L) = \varepsilon$ *on* $\Sigma = \{0, 1\}$*, paired with* $Q(0^L) = 1$*, satisfy* $\Omega_L(P, Q) = (1 - \varepsilon)L$ *but* $D_{\mathrm{KL}}(P_{\le t} \| Q_{\le t}) = +\infty$ *for every* t*.*

The example in Theorem 6 has linear path-level overlap and infinite prefix KL: P assigns positive mass to prefixes with zero Q-probability. Thus linear Ω_L is compatible with infinite one-sided prefix KL. By Proposition 1, a uniform bound on Ω_L also bounds the per-token rate σ_L.

Corollary 1 (SDC consequence under singleton proposal)**.** *Let* $\mathcal{Q} = \{Q\}$*,* $B = 1$*, and* $P_v = P_M^{uv,L}$ *for* $v \in \mathcal{R}(u, T)$*. Then* $\Omega_{L,1}(P_v, \mathcal{Q}) = \Omega_L(P_v, Q)$ *and* $\sigma_{L,1}(P_v, \mathcal{Q}) \le \Omega_L(P_v, Q)$*. Suppose that, uniformly over these reachable states, every history* $h \in \Sigma^{s-1}$ *with positive sequential meet mass*

$$m_{v,s-1}(h) := \prod_{i=1}^{s-1} \min\{P_{v,i}(h_i \mid h_{<i}), Q_i(h_i \mid h_{<i})\} > 0$$

satisfies, for $1 \le s \le L$*,*

$$\mathrm{ov}\big(P_{v,s}(\cdot \mid h), Q_s(\cdot \mid h)\big) \ge 1 - c/L,$$

where $c \ge 0$ *is independent of* L, T, v *and the empty product is* 1*. For sufficiently large* L*,* $\sigma_L(P_v, Q) \ge \sum_{t=1}^{L} (1 - c/L)^t = \Theta(L)$ *uniformly, so for* $T \ge L$ *the iterated Leviathan decoder uses* $O(T/L)$ *verification rounds. In the negative direction, if* $\sup_{v \in \mathcal{R}(u,T)} \Omega_L(P_v, Q) \le C$*, then* $\mathrm{SDC}_{1,L}^{\mathcal{Q}}(M; u, T) \ge T/(C + 1)$*. A uniform geometric-overlap bound gives such a constant* C *by Theorem 5.*

6 Same Proposal, Two Architectures, $\Theta(L)$ Gap

We construct two fixed-precision Transformer families with different verification-round complexity under a common proposal family consisting of a uniform draft and two deterministic constant drafts.

Definition 5 (Proposal family $\mathcal{Q}$)**.** $\mathcal{Q} = \{Q^{\mathrm{flat}}, Q_0^{\mathrm{const}}, Q_1^{\mathrm{const}}\}$, where Q^{flat} is uniform on $\{0, 1\}^L$ and Q_a^{const} is the point mass on a^L.

Definition 6 (Positive family $\mathcal{F}_{a,L}^{+}$)**.** For $a \in \{0, 1\}$ and $\delta = 2^{-b}$, the constant-readout Transformer $M_{a,L}^{+}$ has next-token law $p_a(a) = 1 - \delta$, $p_a(1 - a) = \delta$ at every step and prefix. The path law is $P_{a,L}^{+} = \bigotimes_{s=1}^{L} p_a$. An explicit hidden-dimension-2 realization has token embeddings $e(0) = e(1) = (1, 0)$, zero attention and feed-forward matrices, and fixed-precision stochastic readout table R_a with $R_a(1, 0) = (1 - \delta, \delta)$ if $a = 0$ and $(\delta, 1 - \delta)$ if $a = 1$.

Definition 7 (Negative family $\mathcal{F}_{w,L}^{-}$)**.** For $w \in \{0, 1\}$, the alternating-parity Transformer $M_{w,L}^{-}$ emits $Y_n = w \oplus (n \bmod 2)$ at absolute decoded position n. The local path law after a prefix of length r is the point mass on the deterministic alternating string $x_s^{(w,r)} = w \oplus ((r + s) \bmod 2)$,

$1 \le s \le L$. An explicit realization has hidden coordinates $(1, w, \pi_n)$ with $\pi_n = n \bmod 2$ supplied by the binary positional encoding, zero attention, and a two-neuron ReLU feed-forward block $z = \mathrm{ReLU}(w + \pi_n) - 2\,\mathrm{ReLU}(w + \pi_n - 1) = w \oplus \pi_n$ followed by a ρ_b-rounded one-hot readout. The displayed construction has hidden dimension $O(1)$ [Merrill et al., 2022, §3] and all weights in $\{-2, -1, 0, 1, 2\}$.

Lemma 4 (Positive realizability and per-token rate)**.** *$M^+_{a,L}$ realizes $P^+_{a,L}$ exactly under the construction of Definition 6. Under Q^{const}_a,*

$$\sigma_L(P^+_{a,L}, Q^{\mathrm{const}}_a) \;=\; \sum_{t=1}^{L} (1 - 2^{-b})^t \;\ge\; L(1 - L\,2^{-b}).$$

If $L\,2^{-b} \le \varepsilon$, then $\sigma_L(P^+_{a,L}, Q^{\mathrm{const}}_a) \ge (1 - \varepsilon)L$.

Lemma 5 (Negative realizability and anti-correlation)**.** *$M^-_{w,L}$ realizes the alternating local law of Definition 7 exactly. For every decoded prefix length r, every $w \in \{0, 1\}$, every $Q \in \mathcal{Q}$, and every $1 \le t \le L$, $\mathbb{E}_{\widetilde{y} \sim Q}\Big[\prod_{s=1}^{t} \min(1, P^-_{w,r,s}(\widetilde{y}_s \mid \widetilde{y}_{<s}) / Q_s(\widetilde{y}_s \mid \widetilde{y}_{<s}))\Big] \le 2 \cdot 2^{-t}$. Consequently $\sigma_L(P^-_{w,r,L}, Q) \le 1$ uniformly ($\sigma_L < 1$ for Q^{flat}).*

Proof sketch. For a deterministic target $x_{1:L}$, the per-token rate and path-level overlap both equal $\sum_{t=1}^{L} Q_{\le t}(x_{1:t})$: the clipped-ratio product is the indicator of a matching prefix. This sum is $1 - 2^{-L}$ for Q^{flat} and at most 1 for either constant proposal, since a constant prefix can match the alternating target only at depth 1. Hence both quantities are uniformly bounded by 1. The per-depth bound $2 \cdot 2^{-t}$ covers both cases. Full proof in section D. □

Theorem 7 ($\Theta(L)$ Architecture Separation)**.** *Under proposal family $\mathcal{Q}$ of Definition 5, for every ε with $L\,2^{-b} \le \varepsilon \le 1/2$ and every $T \ge 2L$,*

$$\mathrm{SDC}^{\mathcal{Q}}_{1,L}(M^+_{a,L}; u, T) \le \frac{T + L - 1}{(1 - \varepsilon)L}, \qquad \mathrm{SDC}^{\mathcal{Q}}_{1,L}(M^-_{w,L}; u, T) \ge \frac{T}{2}.$$

The speedup ratio is $\Theta(L)$ for $\mathcal{F}^+$ and $O(1)$ for $\mathcal{F}^-$, a $\Theta(L)$ separation under the same proposal family. (The negative-side $T/2$ bound is the admissible-class lower bound; an Ω_L-based argument extends it to trajectory-exact rejection-style decoders using one proposal trajectory per round, since for the deterministic alternating target every $Q \in \mathcal{Q}$ has $\Omega_L(P^-_{w,r,L}, Q) \le 1$, giving $\mathbb{E}_\Phi[N] \ge T/2$ for every such decoder.)

For the $O(\log L)$-bit description-length claim, choose $\varepsilon = 1/2$ and $b = \max\{6, \lceil \log_2(2L) \rceil\}$, which ensures $L2^{-b} \le 1/2$. Each construction uses a constant number of b-bit parameters and readout entries and a fixed positional-parity rule, so its description length is $O(b) = O(\log L)$. The SDC bounds hold throughout the stated parameter range.

Proof sketch. For $\mathcal{F}^+$, use Q^{const}_a in every round. By Lemma 4, $\mathbb{E}[T'_r \mid \mathcal{F}_{r-1}] \ge \mathbb{E}[\tau_r \mid \mathcal{F}_{r-1}] \ge (1 - \varepsilon)L$. Wald's bound and $S_N \le T + L - 1$ give $\mathrm{SDC} \le (T + L - 1)/((1 - \varepsilon)L) = O(T/L)$, without proposal optimization. For $\mathcal{F}^-$, Lemma 5 gives $\sigma^* \le 1$ uniformly across reachable local targets; Theorem 1 yields the lower bound. For the larger trajectory-exact rejection-style class, the same lower bound follows from $\Omega_L \le 1$: the prefix-overlap sum is $1 - 2^{-L}$ for Q^{flat} and at most 1 for either constant proposal, so the expected emitted length per round is at most 2. Full proof in section D. □

Numerical illustration. Take $L = 64$, $b = 12$, $\varepsilon = L2^{-b} = 1/64$, and $T = 10^4$. The positive family has $\mathrm{SDC} \le 10063/63 \approx 159.73$, while the negative family has $\mathrm{SDC} \ge 5000$. The certified ratio exceeds 31. The asymptotic separation holds along parameter sequences satisfying $L2^{-b} \le \varepsilon \le 1/2$ and $T \ge 2L$.

7 Discussion

Per-token rate and path-level overlap. TokenVerify has expected accepted-prefix length σ_L, while Ω_L bounds the expected accepted length of every trajectory-exact rejection-style round. Proposition 4 shows that Ω_L/σ_L is unbounded, and Theorem 3 relates σ_L to multi-round verification complexity. Optimal-transport formulations of speculative decoding are studied by Sun et al. [2023].

Relation to optimal verification. Yin et al. [2024] characterize expected rejection counts and establish optimality within the sequential framework of their Algorithm 2. Their Appendix C takes each acceptance probability to depend on the output prefix and current draft token. Proposition 3 allows acceptance of an earlier token to depend on a later draft token and gives a binary example retaining the LKM residual.

Sun et al. [2025] establish block-verification optimality for the expected number of generated tokens after a fixed number of iterations in their valid-verifier model. Their residual includes a prefix-dependent weight on the target conditional. Proposition 3 illustrates nonlocal gains in accepted-prefix length with the LKM residual, and Theorems 1 to 3 relate per-token rates to finite-lookahead target-call complexity. The open question in our Limitations concerns tightness of Ω_L within the rejection-style class that retains the LKM residual.

Comparison with representation succinctness. Bergsträßer et al. [2025] show that UHATs can be exponentially more succinct than LTL and doubly exponentially more succinct than finite automata. Their object is the descriptional size of M on a regular task; ours is the achievable speedup of exact decoding from M on an autoregressive task. With the stated precision choice $b = \max\{6, \lceil \log_2(2L) \rceil\}$, Theorem 7 gives two Transformers with $O(\log L)$-bit descriptions whose target-verification cost differs by $\Theta(L)$ under the same proposal family.

KV-cache and compression analyses. KV-cache lower bounds [Haris and Onak, 2025] concern memory requirements. Our SDC bounds measure sequential target-verification rounds under a fixed proposal family. Theorem 7 establishes a $\Theta(L)$ separation in this resource between targets of comparable description size.

Implications for proposal choice. The clipped-ratio acceptance rule of Leviathan et al. [2023] yields expected accepted-prefix length σ_L. Under the hypotheses of Theorem 3, the iterated Leviathan decoder achieves the optimal verification-round complexity up to constants. Theorem 7 illustrates the dependence on the proposal family: the alternating target has bounded local acceptance rate under each of the three proposals, giving $\mathrm{SDC} = \Theta(T)$. If the proposal family contains both alternating length-L strings, selecting the one matching the current target phase gives $\sigma_L = L$ at every reachable state.

Trace-based estimation. For a fixed local target/proposal pair, define the following statistics on a draft trace $\widetilde{y} \sim Q$:

$$U(\widetilde{y}) = \sum_{t=1}^{L} \prod_{s=1}^{t} \min\left(1, \frac{P_s(\widetilde{y}_s \mid \widetilde{y}_{<s})}{Q_s(\widetilde{y}_s \mid \widetilde{y}_{<s})}\right),$$

$$V(\widetilde{y}) = \sum_{t=1}^{L} \min\left(1, \frac{P_{\le t}(\widetilde{y}_{\le t})}{Q_{\le t}(\widetilde{y}_{\le t})}\right).$$

They satisfy $\mathbb{E}_Q[U] = \sigma_L$ and $\mathbb{E}_Q[V] = \Omega_L$; under TokenVerify, $U = \mathbb{E}[\tau \mid \widetilde{y}]$. The identity $\mathbb{E}_Q[V] = \Omega_L$ follows by multiplying each summand by $Q_{\le t}$ and summing over prefixes. Averages over draft traces can support offline drafter and lookahead calibration and compare sequential acceptance with the coupling ceiling. Wall-clock speedup additionally depends on drafting and verification costs and on the states visited during decoding.

Limitations. The SDC analysis applies to admissible clipped-ratio verification. Its uniform Θ characterization assumes exact proposal optimization, $T \ge L$, a positive reachable-rate floor, and bounded reachable-rate heterogeneity. Corollary 1 gives the local conditional-overlap condition for linear sequential speedup.

For the larger rejection-style class retaining the fixed LKM residual, tightness of Ω_L for every target/proposal pair remains open. The architecture separation uses the three-element proposal family of Definition 5; extensions to richer draft families and empirical evaluation on production model pairs are directions for further study. Tree-based and multi-head verification [Cai et al., 2024, Miao et al., 2023] provide additional protocol classes for analysis.

Acknowledgments

The author thanks the anonymous reviewers and the area chair for their constructive comments, which improved the presentation of this paper.

The author used Claude (Anthropic) and ChatGPT (OpenAI) for language polishing and for assistance with theoretical work, including discussing proof strategies and checking derivations. All results, proofs, and claims were verified by the author, who takes full responsibility for the content of this paper.

This work was supported by the Deep Research Training Program (General Program) of Chu Kochen Honors College, Zhejiang University.

The author declares no competing interests.

References

Pascal Bergsträßer, Ryan Cotterell, and Anthony W Lin. Transformers are inherently succinct. *arXiv preprint arXiv:2510.19315*, 2025.

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D Lee, Deming Chen, and Tri Dao. Medusa: Simple llm inference acceleration framework with multiple decoding heads. *arXiv preprint arXiv:2401.10774*, 2024.

Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318*, 2023.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Themistoklis Haris and Krzysztof Onak. Compression barriers for autoregressive transformers. *arXiv preprint arXiv:2502.15955*, 2025.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning*, pages 19274–19286. PMLR, 2023.

Torgny Lindvall. *Lectures on the coupling method*. Courier Corporation, 2002.

William Merrill, Ashish Sabharwal, and Noah A Smith. Saturated transformers are constant-depth threshold circuits. *Transactions of the Association for Computational Linguistics*, 10:843–856, 2022.

Xupeng Miao, Gabriele Oliaro, Zhihao Zhang, Xinhao Cheng, Zeyu Wang, Zhengxin Zhang, Rae Ying Yee Wong, Alan Zhu, Lijie Yang, Xiaoxiang Shi, et al. Specinfer: Accelerating generative large language model serving with tree-based speculative inference and verification. *arXiv preprint arXiv:2305.09781*, 2023.

Ziteng Sun, Ananda Theertha Suresh, Jae Hun Ro, Ahmad Beirami, Himanshu Jain, and Felix Yu. SpecTr: Fast Speculative Decoding via Optimal Transport. In *Neural Information Processing Systems*, 2023. URL `https://mlanthology.org/neurips/2023/sun2023neurips-spectr/`.

Ziteng Sun, Uri Mendlovic, Yaniv Leviathan, Asaf Aharoni, Jae Hun Ro, Ahmad Beirami, and Ananda Theertha Suresh. Block Verification Accelerates Speculative Decoding. In *International Conference on Learning Representations*, 2025. URL `https://mlanthology.org/iclr/2025/sun2025iclr-block/`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

David Williams. *Probability with martingales*. Cambridge university press, 1991.

Ming Yin, Minshuo Chen, Kaixuan Huang, and Mengdi Wang. A Theoretical Perspective for Speculative Decoding Algorithm. In *Neural Information Processing Systems*, 2024. doi: 10.52202/079017-4067. URL `https://mlanthology.org/neurips/2024/yin2024neurips-theoretical/`.

A Proofs for section 3: per-token vs. path-level

Proof of Proposition 1. For a fixed prefix $z_{1:t} \in \Sigma^t$ with positive $Q_{\le t}$-mass,

$$\begin{aligned} Q_{\le t}(z_{1:t}) \prod_{s=1}^{t} \min\Big(1, \tfrac{P_{s|z_{<s}}(z_s)}{Q_{s|z_{<s}}(z_s)}\Big) &= \prod_{s=1}^{t} Q_{s|z_{<s}}(z_s) \min\Big(1, \tfrac{P_{s|z_{<s}}(z_s)}{Q_{s|z_{<s}}(z_s)}\Big) \\ &= \prod_{s=1}^{t} \min\big(P_{s|z_{<s}}(z_s), Q_{s|z_{<s}}(z_s)\big) \\ &\le \min\Big(\prod_{s=1}^{t} P_{s|z_{<s}}(z_s), \prod_{s=1}^{t} Q_{s|z_{<s}}(z_s)\Big) \\ &= \min\big(P_{\le t}(z_{1:t}), Q_{\le t}(z_{1:t})\big). \end{aligned}$$

The inequality applies $\min(\prod a_s, \prod b_s) \ge \prod \min(a_s, b_s)$, valid because each right-hand factor is dominated by each a_s and each b_s. Summing over $z_{1:t} \in \Sigma^t$ with positive $Q_{\le t}$-mass—zero-mass prefixes contribute nothing to σ_L by definition—gives

$$\mathbb{E}_{\widetilde{y}\sim Q}\Big[\prod_{s=1}^{t} \min\big(1, P_{s|\widetilde{y}_{<s}}(\widetilde{y}_s)/Q_{s|\widetilde{y}_{<s}}(\widetilde{y}_s)\big)\Big] \le \sum_{z_{1:t}} \min\big(P_{\le t}(z_{1:t}), Q_{\le t}(z_{1:t})\big) = \mathrm{ov}(P_{\le t}, Q_{\le t}).$$

Summing over $t = 1, \dots, L$ completes the proof. □

Proof of Proposition 2. Set $\Sigma = \{0,1\}$, $p = (0.9, 0.1)$, $q = (0.1, 0.9)$, $L = 2$. The coordinate overlap is $\mathrm{ov}(p,q) = \min(0.9, 0.1) + \min(0.1, 0.9) = 0.2$. For $P = p^{\otimes 2}$ and $Q = q^{\otimes 2}$, the $t = 2$ summand of $\sigma_L(P, Q)$ factors as a product over independent positions:

$$\begin{aligned} &\sum_{(y_1,y_2)\in\{0,1\}^2} q(y_1)q(y_2) \min\Big(1, \tfrac{p(y_1)}{q(y_1)}\Big) \min\Big(1, \tfrac{p(y_2)}{q(y_2)}\Big) \\ &\quad = \Big(\sum_{y\in\{0,1\}} \min(p(y), q(y))\Big)^2 = 0.2^2 = 0.04. \end{aligned}$$

The $t = 2$ summand of $\Omega_L(P, Q)$, by contrast, sums pairwise minima over the four joint atoms:

$$\begin{aligned} \mathrm{ov}(p^{\otimes 2}, q^{\otimes 2}) &= \min(0.81, 0.01) + \min(0.09, 0.09) \\ &\quad + \min(0.09, 0.09) + \min(0.01, 0.81) \\ &= 0.01 + 0.09 + 0.09 + 0.01 = 0.20. \end{aligned}$$

The two $t = 2$ summands differ by a factor of five, confirming $\sigma_L(P, Q) < \Omega_L(P, Q)$ on this product instance. □

Proof of Proposition 3. Fix any integer precision $b \ge 6$. We work with $L = 2$, $\Sigma = \{0, 1\}$, and a single proposal branch. Define

$$\begin{aligned} &P(00) = \tfrac{1}{2}, && P(01) = 0, && P(10) = \tfrac{1}{4}, && P(11) = \tfrac{1}{4}; \\ &Q(00) = \tfrac{3}{16}, && Q(01) = \tfrac{9}{16}, && Q(10) = \tfrac{1}{8}, && Q(11) = \tfrac{1}{8}. \end{aligned}$$

The displayed atom probabilities are representable already in D_4, and hence in every permitted D_b. The induced conditionals are

$$P_1(0) = \tfrac{1}{2},\ Q_1(0) = \tfrac{3}{4}; \quad P(0 \mid 0) = 1,\ Q(0 \mid 0) = \tfrac{1}{4}; \quad P(\cdot \mid 1) = Q(\cdot \mid 1) = (\tfrac{1}{2}, \tfrac{1}{2}).$$

Computing σ_L. Write $a_t = \Pr_{\mathrm{TokenVerify}}(\tau \ge t)$ for the t-prefix survival probability. Then $\sigma_L(P, Q) = a_1 + a_2$ for this instance. The $t = 1$ contribution is

$$\begin{aligned} a_1 &= \mathbb{E}_{\widetilde{y}_1\sim Q_1} \min\big(1, P_1(\widetilde{y}_1)/Q_1(\widetilde{y}_1)\big) \\ &= \sum_{y\in\{0,1\}} \min(P_1(y), Q_1(y)) = \min(\tfrac{1}{2}, \tfrac{3}{4}) + \min(\tfrac{1}{2}, \tfrac{1}{4}) = \tfrac{3}{4}. \end{aligned}$$

The $t = 2$ contribution

$$a_2 = \mathbb{E}_{\widetilde{y}\sim Q} \min\big(1, \tfrac{P_1(\widetilde{y}_1)}{Q_1(\widetilde{y}_1)}\big) \min\big(1, \tfrac{P(\widetilde{y}_2|\widetilde{y}_1)}{Q(\widetilde{y}_2|\widetilde{y}_1)}\big)$$

is evaluated by enumerating the four proposal paths and recording $Q(\widetilde{y}) \cdot [\text{first clip}] \cdot [\text{second clip}]$:

- $\widetilde{y} = 00$: $Q = 3/16$, factors $\min(1, (1/2)/(3/4)) \cdot \min(1, 1/(1/4)) = \frac{2}{3} \cdot 1 = \frac{2}{3}$, so contribution $\frac{3}{16} \cdot \frac{2}{3} = \frac{1}{8}$.
- $\widetilde{y} = 01$: $Q = 9/16$, factors $\frac{2}{3} \cdot \min(1, 0/(3/4)) = \frac{2}{3} \cdot 0 = 0$, contribution 0.
- $\widetilde{y} = 10$: $Q = 1/8$, factors $\min(1, (1/2)/(1/4)) \cdot \min(1, (1/2)/(1/2)) = 1 \cdot 1 = 1$, contribution $\frac{1}{8}$.
- $\widetilde{y} = 11$: $Q = 1/8$, factors $1 \cdot \min(1, (1/2)/(1/2)) = 1$, contribution $\frac{1}{8}$.

Summing, $a_2 = \frac{1}{8} + 0 + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}$, and hence $\sigma_L(P, Q) = a_1 + a_2 = \frac{3}{4} + \frac{3}{8} = \frac{9}{8}$.

Constructing the trajectory-exact protocol with $\mathbb{E}[\tau] > \sigma_L$. We define an acceptance rule that depends on the full proposal $\widetilde{y}_{1:2}$:

- For $\widetilde{y} \in \{00, 10, 11\}$: accept both tokens, $\tau = 2$.
- For $\widetilde{y} = 01$: with probability $\frac{5}{9}$, accept the first token and reject the second ($\tau = 1$); with the remaining probability $\frac{4}{9}$, reject the first token ($\tau = 0$).
- Residuals: when the protocol rejects the second token after proposal 01, the residual at position 2 is the token 0. When it rejects the first token after proposal 01, the residual at position 1 is the token 1.

One checks that the output masses on Σ^2 are

$$\Pr(Y_{1:2} = 00) = Q(00) + Q(01) \cdot \tfrac{5}{9} = \tfrac{3}{16} + \tfrac{9}{16} \cdot \tfrac{5}{9} = \tfrac{3}{16} + \tfrac{5}{16} = \tfrac{1}{2},$$
$$\Pr(Y_{1:2} = 10) = Q(10) = \tfrac{1}{8}, \quad \Pr(Y_{1:2} = 11) = Q(11) = \tfrac{1}{8}, \quad \Pr(Y_{1:2} = 01) = 0,$$

and the length-one output mass at token 1 is $Q(01) \cdot \frac{4}{9} = \frac{9}{16} \cdot \frac{4}{9} = \frac{1}{4}$. These agree with the stopped-target masses for $Z \sim P$ stopped by the rule that, when $Z_1 = 1$, emits one token with probability $\frac{1}{2}$ using an independent coin, and otherwise emits two tokens: the length-1 output 1 carries mass $P_1(1) \cdot \frac{1}{2} = \frac{1}{4}$, and each length-2 output is distributed as P conditioned on Z_1 proceeding past length 1. Hence the protocol is trajectory-exact.

Computing $\mathbb{E}[\tau]$. Applying the layer-cake formula $\mathbb{E}[\tau] = \Pr(\tau \geq 1) + \Pr(\tau \geq 2)$:

$$\begin{aligned}\Pr(\tau \geq 1) &= \Pr(\widetilde{y} \in \{00, 10, 11\}) + \tfrac{5}{9} \Pr(\widetilde{y} = 01) \\ &= \tfrac{3}{16} + \tfrac{1}{8} + \tfrac{1}{8} + \tfrac{5}{9} \cdot \tfrac{9}{16} = \tfrac{3}{16} + \tfrac{2}{16} + \tfrac{2}{16} + \tfrac{5}{16} = \tfrac{12}{16} = \tfrac{3}{4}, \\ \Pr(\tau \geq 2) &= \Pr(\widetilde{y} \in \{00, 10, 11\}) = Q(00) + Q(10) + Q(11) = \tfrac{3}{16} + \tfrac{1}{8} + \tfrac{1}{8} = \tfrac{7}{16}.\end{aligned}$$

Thus $\mathbb{E}[\tau] = \frac{12}{16} + \frac{7}{16} = \frac{19}{16}$, which exceeds $\sigma_L(P, Q) = \frac{9}{8} = \frac{18}{16}$ by $\frac{1}{16}$. Here $\mathrm{ov}(P_1, Q_1) = \frac{3}{4}$ and $\mathrm{ov}(P_{\leq 2}, Q_{\leq 2}) = \frac{3}{16} + 0 + \frac{1}{8} + \frac{1}{8} = \frac{7}{16}$, so $\mathbb{E}[\tau] = \Omega_2(P, Q) = \frac{19}{16}$: the protocol attains the path-level overlap.

The protocol is trajectory-exact and rejection-style, yet it fails the admissibility condition of Definition 2: along proposal path $\widetilde{y}_{1:2} = 00$, the survival probability $\Pr(\tau \geq 2 \mid \widetilde{y} = 00) = 1$ exceeds the per-token clipped-ratio product $\frac{2}{3} \cdot 1 = \frac{2}{3}$. Acceptance decisions are correlated across positions in a manner that violates (4), and it is precisely this inter-position correlation that allows $\mathbb{E}[\tau]$ to exceed σ_L. □

Proposition 4 (Unbounded gap between the complete sums)**.** *There is a sequence of binary product-law pairs $(P^{(J)}, Q^{(J)})$ with fixed-precision conditional probabilities and horizons $L_J = 2^{J+1} + 2J - 2$ such that*

$$\sigma_{L_J}(P^{(J)}, Q^{(J)}) = 2 - 4^{-J} - 2^{-J} \leq 2, \qquad \Omega_{L_J}(P^{(J)}, Q^{(J)}) \geq \frac{(3/2)^J}{2\sqrt{J}} \longrightarrow \infty.$$

In particular, $\Omega_{L_J} / \sigma_{L_J} \to \infty$.

Proof. Set $p = (3/4, 1/4)$ and $q = (1/4, 3/4)$. For $j = 1, \ldots, J$, stage j consists of two active coordinates with laws p under $P^{(J)}$ and q under $Q^{(J)}$, followed by 2^j filler coordinates that are deterministically zero under both laws. All coordinates are independent. Each conditional probability lies in $\{0, 1/4, 3/4, 1\}$, a fixed precision grid independent of J. The total length is $\sum_{j=1}^{J}(2 + 2^j) = L_J$.

Each active coordinate multiplies the sequential survival mass by $\mathrm{ov}(p, q) = 1/2$, and each filler leaves it unchanged. The two active coordinates in stage j therefore contribute $2^{-(2j-1)} + 2^{-2j} = 3 \cdot 4^{-j}$ to σ_{L_J}, while its filler block contributes $2^j 4^{-j} = 2^{-j}$. Thus

$$\sigma_{L_J} = \sum_{j=1}^{J}(3 \cdot 4^{-j} + 2^{-j}) = 2 - 4^{-J} - 2^{-J} \leq 2.$$

At every prefix in the last filler block, the path overlap equals $\mathrm{ov}(p^{\otimes 2J}, q^{\otimes 2J})$. Restricting the overlap sum to active strings with exactly J zeros and J ones gives

$$\mathrm{ov}(p^{\otimes 2J}, q^{\otimes 2J}) \geq \binom{2J}{J}(3/16)^J \geq \frac{(3/4)^J}{2\sqrt{J}}.$$

Here $\binom{2J}{J} \geq 4^J/(2\sqrt{J})$ for $J \geq 1$: equality holds at $J = 1$, and the induction step follows from $(2J+1)/(2J+2) \geq \sqrt{J/(J+1)}$. There are 2^J prefixes in the last filler block, so their contribution alone gives

$$\Omega_{L_J} \geq 2^J \frac{(3/4)^J}{2\sqrt{J}} = \frac{(3/2)^J}{2\sqrt{J}} \longrightarrow \infty.$$

□

B Proofs for section 4: tight SDC characterization

Proof of Lemma 1. Write $\mathbb{E}[\tau] = \sum_{t=1}^{L} \Pr(\tau \geq t)$ for any nonnegative integer-valued $\tau \leq L$. The admissibility condition of Definition 2 then gives

$$\begin{aligned}
\Pr(\tau \geq t) &= \sum_{z_{1:t} \in \Sigma^t} Q^{\mathrm{mix}}_{\leq t}(z_{1:t}) \Pr(\tau \geq t \mid \widetilde{y}_{1:t} = z_{1:t}) \\
&\leq \sum_{z_{1:t}} Q^{\mathrm{mix}}_{\leq t}(z_{1:t}) \prod_{s=1}^{t} \min\Big(1, \tfrac{P_{s|z_{<s}}(z_s)}{Q^{\mathrm{mix}}_{s|z_{<s}}(z_s)}\Big) \\
&= \mathbb{E}_{\widetilde{y} \sim Q^{\mathrm{mix}}}\Big[\prod_{s=1}^{t} \min\big(1, P_{s|\widetilde{y}_{<s}}(\widetilde{y}_s)/Q^{\mathrm{mix}}_{s|\widetilde{y}_{<s}}(\widetilde{y}_s)\big)\Big],
\end{aligned}$$

which is exactly the t-th summand of $\sigma_L(P, Q^{\mathrm{mix}})$. Summing over t yields $\mathbb{E}[\tau] \leq \sigma_L(P, Q^{\mathrm{mix}})$, and optimizing over $Q^{\mathrm{mix}} \in \mathcal{Q}_B$ gives $\mathbb{E}[\tau] \leq \sigma_{L,B}(P, \mathcal{Q})$. In the standard Leviathan protocol [Leviathan et al., 2023, Alg. 1], position s is accepted with conditional probability exactly $\min(1, P_{s|\widetilde{y}_{<s}}(\widetilde{y}_s)/Q^{\mathrm{mix}}_{s|\widetilde{y}_{<s}}(\widetilde{y}_s))$ given prior acceptances, and these decisions are sequentially independent on the proposal path; consequently $\Pr(\tau \geq t \mid \widetilde{y}_{1:t}) = \prod_{s=1}^{t} \min(1, P_s/Q^{\mathrm{mix}}_s)$ holds with equality, and all the foregoing inequalities become equalities, giving $\mathbb{E}[\tau] = \sigma_L(P, Q^{\mathrm{mix}})$ for every fixed mixture and $\mathbb{E}[\tau] \geq \sigma_{L,B}(P, \mathcal{Q}) - \varepsilon_b$ whenever the selected mixture has rate within the uniform proposal-optimization tolerance ε_b of the supremum. Here ε_b denotes the assumed uniform tolerance in proposal selection.

For $T' = \tau + \mathbf{1}\{\tau < L\}$,

$$\mathbb{E}[T'] = \mathbb{E}[\tau] + \Pr(\tau < L) \leq \mathbb{E}[\tau] + 1 \leq \sigma_{L,B}(P, \mathcal{Q}) + 1.$$

The extra $+1$ accounts for the residual sample drawn at the first rejection: without it, a round with $\tau < L$ would produce only τ tokens, leaving per-step exactness incomplete. When $\tau = L$, no residual is needed and $T' = L$. □

Proof of Lemma 2. The argument is the standard Wald-style stopped-sum identity [Williams, 1991, §10.10]. Because X_r is bounded and N is integrable,

$$\mathbb{E}\Big[\sum_{r=1}^{N} X_r\Big] = \sum_{r\geq 1} \mathbb{E}[\mathbf{1}_{\{N\geq r\}} X_r].$$

Since N is a stopping time, $\{N \geq r\} = \{N > r-1\} \in \mathcal{F}_{r-1}$, so $\mathbb{E}[\mathbf{1}_{\{N\geq r\}} X_r] = \mathbb{E}[\mathbf{1}_{\{N\geq r\}} \mathbb{E}[X_r \mid \mathcal{F}_{r-1}]]$. Bounding $\mathbb{E}[X_r \mid \mathcal{F}_{r-1}]$ between a and b on $\{N \geq r\}$ and summing via $\sum_{r\geq 1} \Pr(N \geq r) = \mathbb{E}[N]$ yields the stated two-sided inequality. □

Lemma 6 (Reachable prefixes are target-reachable)**.** *For any multi-round trajectory-exact decoder, every emitted prefix of length less than T that occurs with positive probability has positive prefix mass under $P_M^{u,T}$.*

Proof. Suppose prefix v with $|v| < T$ is emitted with positive probability, yet $P_M^{u,T}$ assigns zero probability to every length-T extension of v. Every subsequent round only appends tokens to v, so the final output begins with v with positive probability. This contradicts multi-round trajectory exactness, which demands that the final length-T output has law $P_M^{u,T}$. □

Proof of Theorem 1. Apply Lemma 2 with $X_r = T'_r$. On the event $\{N \geq r\}$, the round-r local target $P_M^{uY^{<r},L}$ has a reachable prefix $Y^{<r}$ by Lemma 6, hence its local rate is at most σ^*. Lemma 1 then gives $\mathbb{E}[T'_r \mid \mathcal{F}_{r-1}] \leq \sigma^* + 1$ on $\{N \geq r\}$. Taking $b = \sigma^* + 1$ in Lemma 2,

$$\mathbb{E}[S_N] \leq (\sigma^* + 1)\mathbb{E}[N].$$

At termination $S_N \geq T$, so $T \leq (\sigma^* + 1)\mathbb{E}[N]$, giving $\mathbb{E}[N] \geq T/(\sigma^* + 1)$. The SDC lower bound follows by taking the infimum over admissible decoders. □

Lemma 7 (Composition exactness)**.** *The iterated Leviathan decoder, with exact auxiliary randomization for the fixed-precision target tables, is multi-round trajectory-exact.*

Proof. TokenVerify admits an equivalent sequential execution in which draft tokens are revealed as they are tested: each acceptance decision uses the current proposal prefix and auxiliary randomness independent of drafting. Conditional on the emitted prefix and the past exposed decisions, the next untested draft token has its proposal conditional law q, while the target conditional is $p = p_M(\cdot \mid u, Y_{<t})$. For a token x, the accepted mass is $\min\{p(x), q(x)\}$. The total rejection probability is $r = 1 - \sum_x \min\{p(x), q(x)\}$; if $r > 0$, the LKM residual assigns mass $(p(x) - q(x))_+/r$ to x. Thus the next emitted token has conditional mass

$$\min\{p(x), q(x)\} + r\,\frac{(p(x) - q(x))_+}{r} = p(x),$$

with the residual term omitted when $r = 0$ [Leviathan et al., 2023, Alg. 1, App. A.1]. At a round boundary, including after L acceptances in a capped round, a fresh draft is sampled and the same identity applies for any adaptively selected proposal.

Averaging over the exposed decisions gives $\Pr(Y_t = x \mid Y_{<t} = y_{<t}) = p_M(x \mid u, y_{<t})$ for every positive-probability output prefix. Starting with the empty prefix and inducting on each deterministic $t \leq T$ therefore yields

$$\Pr(Y_{1:t} = y_{1:t}) = \Pr(Y_{<t} = y_{<t})\, p_M(y_t \mid u, y_{<t}) = P_M^{u,t}(y_{1:t}).$$

Zero-probability prefixes satisfy the same identity trivially. Since each round emits at least one token, the decoder reaches T tokens; truncation at this deterministic length gives law $P_M^{u,T}$. □

Proof of Theorem 2. Run the iterated Leviathan decoder. By Lemma 7 it is admissible and exact. At every reachable prefix $\sigma_r \geq \sigma_*$, so Lemma 1 gives

$$\mathbb{E}[T'_r \mid \mathcal{F}_{r-1}] \geq \max\{1, \sigma_* - \varepsilon_b\} =: a.$$

Lemma 2 gives $a\,\mathbb{E}[N] \leq \mathbb{E}[S_N] \leq T + L - 1$, where the overshoot $S_N - T$ is at most $L - 1$. Hence $\mathbb{E}[N] \leq (T + L - 1)/a$. With $\varepsilon_b = 0$ and fixed L, the asymptotic upper bound follows from $\max\{1, \sigma_*\} \geq \sigma_*$ and $(T + L - 1)/T \to 1$. □

Proof of Theorem 3. Combine Theorems 1 and 2. Set $\varepsilon_b = 0$ (exact optimization) and write $\bar{\sigma} = \sigma_{L,B}(P_M^{u,T}, \mathcal{Q})$. Reachability of the initial prefix gives $\sigma_* \le \bar{\sigma} \le \sigma^*$. The two bounds then yield, for $T \ge L$,

$$\mathrm{SDC}_{B,L}^{\mathcal{Q}}(M; u, T) \ge \frac{T}{\sigma^* + 1} = \frac{\bar{\sigma} + 1}{\sigma^* + 1} \cdot \frac{T}{\bar{\sigma} + 1} \ge \frac{\sigma_* + 1}{\sigma^* + 1} \cdot \frac{T}{\bar{\sigma} + 1},$$

and

$$\begin{aligned} \mathrm{SDC}_{B,L}^{\mathcal{Q}}(M; u, T) &\le \frac{T + L - 1}{\max\{1, \sigma_*\}} \le \frac{2T}{\max\{1, \sigma_*\}} \le \frac{4T}{\sigma_* + 1} \\ &= 4\,\frac{\bar{\sigma} + 1}{\sigma_* + 1} \cdot \frac{T}{\bar{\sigma} + 1} \le 4\,\frac{\sigma^* + 1}{\sigma_* + 1} \cdot \frac{T}{\bar{\sigma} + 1}. \end{aligned}$$

These explicit bounds hold for all reachable-rate ratios under the theorem's hypotheses. If $(\sigma^* + 1)/(\sigma_* + 1) = O(1)$ uniformly, they imply $\mathrm{SDC}_{B,L}^{\mathcal{Q}}(M; u, T) = \Theta(T/(\bar{\sigma} + 1))$ with uniform constants. □

C Proofs for section 5: KL phase transition

Proof of Lemma 3. Apply Pinsker's inequality [Cover, 1999, Thm. 11.6.1] to the finite laws $P_{\le t}$ and $Q_{\le t}$ on Σ^t: $\mathrm{TV}(P_{\le t}, Q_{\le t}) \le \sqrt{D_{\mathrm{KL}}(P_{\le t} \| Q_{\le t})/2} = \sqrt{\eta_t/2}$. Subtracting from 1 and substituting $o_t = 1 - \mathrm{TV}(P_{\le t}, Q_{\le t})$ from Definition 4 gives $o_t \ge 1 - \sqrt{\eta_t/2}$. □

Proof of Theorem 4. By Lemma 3,

$$\Omega_L(P, Q) = \sum_{t=1}^{L} o_t \ge \sum_{t=1}^{L} \bigl(1 - \sqrt{\eta_t/2}\bigr) = L - \sum_{t=1}^{L} \sqrt{\eta_t/2}.$$

When $\sum_{t=1}^{L} \sqrt{\eta_t/2} \le L/2$, this is at least $L/2$. If the left-hand sum is $o(L)$, then $\Omega_L = L - o(L)$ and $L_{\mathrm{eff}} \to 1$.

For the linear-accumulation case $\eta_t \le t\delta$, one has $\sum_{t=1}^{L} \sqrt{\eta_t/2} \le \sum_{t=1}^{L} \sqrt{t\delta/2} \le L\sqrt{L\delta/2}$. The condition $L\sqrt{L\delta/2} \le L/2$ rearranges to $L\delta \le 1/2$, i.e. $L \le 1/(2\delta)$ when $\delta > 0$; for $\delta = 0$ all $\eta_t = 0$. □

Proof of Theorem 5. From the geometric overlap bound, $\Omega_L(P, Q) = \sum_{t=1}^{L} o_t \le \sum_{t=1}^{L} \rho^t = \rho(1 - \rho^L)/(1 - \rho) \le 1/(1 - \rho)$. Since this upper bound is uniform in L, we get $L_{\mathrm{eff}} = \Omega_L/L \le 1/(L(1 - \rho)) \to 0$ as $L \to \infty$. □

Proof of Theorem 6. Set $\Sigma = \{0, 1\}$ and choose $\varepsilon \in (0, 1)$ representable in D_b. Define $P(0^L) = 1 - \varepsilon$, $P(1^L) = \varepsilon$, and $Q(0^L) = 1$, with all remaining strings of mass 0 under both measures. The prefix marginals are $P_{\le t}(0^t) = 1 - \varepsilon$, $P_{\le t}(1^t) = \varepsilon$, and $Q_{\le t}(0^t) = 1$, $Q_{\le t}(1^t) = 0$. Hence $\mathrm{TV}(P_{\le t}, Q_{\le t}) = \varepsilon$ and $o_t = 1 - \varepsilon$, so $\Omega_L(P, Q) = (1 - \varepsilon)L$ is linear in L. Yet $P_{\le t}(1^t) = \varepsilon > 0$ while $Q_{\le t}(1^t) = 0$, forcing $D_{\mathrm{KL}}(P_{\le t} \| Q_{\le t}) = +\infty$ for every $t \ge 1$. Linear Ω_L therefore does not imply any finite one-sided prefix KL. □

Proof of Corollary 1. For a singleton family, the only mixture is Q and Proposition 1 gives $\sigma_L(P_v, Q) \le \Omega_L(P_v, Q)$. Fix a reachable state v, and let

$$A_t = \sum_{z \in \Sigma^t} \prod_{s=1}^{t} \min\{P_{v,s}(z_s \mid z_{<s}), Q_s(z_s \mid z_{<s})\}, \qquad A_0 = 1.$$

Grouping by histories $h \in \Sigma^{t-1}$ and using the local overlap assumption on those with positive meet mass gives

$$A_t = \sum_{h \in \Sigma^{t-1}} m_{v,t-1}(h)\, \mathrm{ov}\bigl(P_{v,t}(\cdot \mid h), Q_t(\cdot \mid h)\bigr) \ge (1 - c/L) A_{t-1}.$$

Thus $A_t \geq (1 - c/L)^t$ and $\sigma_L(P_v, Q) = \sum_{t=1}^{L} A_t \geq \sum_{t=1}^{L} (1 - c/L)^t$. For $L \geq \max\{1, 2c\}$, the last sum lies between Le^{-2c} and L, since $\log(1 - c/L) \geq -2c/L$. The rate floor is therefore $\Omega(L)$ uniformly over reachable states. There is no proposal optimization for a singleton family, so Theorem 2 with $\varepsilon_b = 0$ and $T \geq L$ gives $O(T/L)$ rounds.

For the negative direction, the uniform bound on $\Omega_L(P_v, Q)$ implies $\sigma^* \leq C$ by Proposition 1. Applying the per-round bound and Wald's argument from Theorem 1 gives $\mathrm{SDC} \geq T/(C + 1)$; this lower-bound argument does not require a positive rate floor. □

D Proofs for section 6: $\Theta(L)$ separation

Proof of Lemma 4. The Transformer construction in Definition 6 has constant token embedding $(1, 0)$, zero attention and feed-forward matrices, and readout R_a. The residual hidden vector is therefore $(1, 0)$ at every position, and the readout yields $\Pr[Y = a \mid \text{any prefix}] = 1 - \delta$ and $\Pr[Y = 1 - a \mid \text{any prefix}] = \delta$, with all values in D_b. Hence $M^+_{a,L}$ realizes $P^+_{a,L} = \bigotimes_{s=1}^{L} p_a$.

Under Q_a^{const}, the only sampled proposal path is $\widetilde{y} = a^L$, so for every s,

$$\min\Big(1, \tfrac{P^+_s(a|a^{s-1})}{Q^{\mathrm{const}}_{a,s}(a|a^{s-1})}\Big) = \min(1, 1 - \delta) = 1 - \delta.$$

The t-prefix product is $(1 - \delta)^t$, giving

$$\sigma_L(P^+_{a,L}, Q_a^{\mathrm{const}}) = \sum_{t=1}^{L} (1 - \delta)^t.$$

Bernoulli's inequality gives $(1 - \delta)^t \geq 1 - t\delta$, so $\sum_{t=1}^{L} (1 - \delta)^t \geq L - \delta\, L(L + 1)/2 \geq L - L^2\delta = L(1 - L\delta)$. With $\delta = 2^{-b}$ and $L\delta \leq \varepsilon$, this is $\geq (1 - \varepsilon)L$. □

Proof of Lemma 5. The construction in Definition 7 uses hidden coordinates $(1, w, \pi_n)$ with $\pi_n = n \bmod 2$ from the binary positional encoding, zero attention, and a two-neuron ReLU feed-forward block computing

$$z = \mathrm{ReLU}(w + \pi_n) - 2\,\mathrm{ReLU}(w + \pi_n - 1).$$

On bit inputs $w, \pi_n \in \{0, 1\}$, the three possible values of $w + \pi_n$ map as

$$0 \mapsto 0, \quad 1 \mapsto 1, \quad 2 \mapsto 2 - 2 \cdot 1 = 0.$$

Hence $z = w \oplus \pi_n$. The ρ_b-rounded one-hot readout emits token z with probability 1, so the local path law is the point mass on $x_s^{(w,r)} = w \oplus ((r + s) \bmod 2)$ as required. The hidden dimension is $O(1)$ and all weights lie in $\{-2, -1, 0, 1, 2\}$. The explicit ReLU computation and readout therefore establish the claimed realization directly.

For the anti-correlation bound, the local target is the point mass on the alternating string $x_{1:L} = x_{1:L}^{(w,r)}$. The clipped-ratio product

$$\prod_{s=1}^{t} \min\Big(1, \tfrac{P^-_{w,r,s}(\widetilde{y}_s|\widetilde{y}_{<s})}{Q_s(\widetilde{y}_s|\widetilde{y}_{<s})}\Big)$$

is zero unless $\widetilde{y}_{1:t} = x_{1:t}$, since the target conditional vanishes on every other token. On the matching prefix the target conditional equals 1 and the proposal conditional is at most 1, so each factor is exactly 1. Thus the per-token rate and path-level overlap coincide:

$$\sigma_L(P^-_{w,r,L}, Q) = \Omega_L(P^-_{w,r,L}, Q) = \sum_{t=1}^{L} Q_{\leq t}(x_{1:t}).$$

Under Q^{flat}, the probability that the proposal prefix matches the fixed string $x_{1:t}$ is exactly 2^{-t}, so the t-prefix survival expectation equals 2^{-t}, and $\sigma_L(P^-_{w,r,L}, Q^{\mathrm{flat}}) = \sum_{t=1}^{L} 2^{-t} = 1 - 2^{-L} < 1$.

Under Q_a^{const}, the only proposed prefix is a^t. Since $x_s^{(w,r)}$ alternates with s, the prefix a^t matches $x_{1:t}$ only at $t = 1$, and only when $a = x_1^{(w,r)}$. Hence the $t = 1$ survival expectation is at most 1 and all $t \geq 2$ terms vanish, giving $\sigma_L(P^-_{w,r,L}, Q_a^{\mathrm{const}}) \leq 1$.

For $t = 1$ the bound $1 \le 2 \cdot 2^{-1}$ holds; for $t \ge 2$, the constant-proposal survival is $0 \le 2 \cdot 2^{-t}$ and the flat-proposal survival is $2^{-t} \le 2 \cdot 2^{-t}$. Thus the t-prefix survival expectation is at most $2 \cdot 2^{-t}$ for every $Q \in \mathcal{Q}$. Combining the two cases gives $\sigma_L(P^-_{w,r,L}, Q) = \Omega_L(P^-_{w,r,L}, Q) \le 1$ for every $Q \in \mathcal{Q}$. □

Proof of Theorem 7. For the positive family $\mathcal{F}^+_{a,L}$, every reachable local target is $P^+_{a,L}$. Use the fixed proposal Q^{const}_a in each round. By Lemma 4 and exact Leviathan acceptance,

$$\mathbb{E}[T'_r \mid \mathcal{F}_{r-1}] \ge \mathbb{E}[\tau_r \mid \mathcal{F}_{r-1}] = \sigma_L(P^+_{a,L}, Q^{\mathrm{const}}_a) \ge (1-\varepsilon)L.$$

Wald's bound and $S_N \le T + L - 1$ yield

$$\mathrm{SDC}^{\mathcal{Q}}_{1,L}(M^+_{a,L}; u, T) \le \frac{T + L - 1}{(1-\varepsilon)L}.$$

The decoder uses Q^{const}_a in every round.

For the negative family $\mathcal{F}^-_{w,L}$, every reachable local target is a point mass on an alternating string. Lemma 5 gives $\sigma^* \le 1$, so Theorem 1 yields

$$\mathrm{SDC}^{\mathcal{Q}}_{1,L}(M^-_{w,L}; u, T) \ge \frac{T}{2}.$$

For any trajectory-exact rejection-style round with one proposal trajectory, the path-level bound gives $\mathbb{E}[\tau_r \mid \mathcal{F}_{r-1}] \le \Omega_L(P^-_{w,r,L}, Q) \le 1$, so $\mathbb{E}[T'_r \mid \mathcal{F}_{r-1}] \le 2$. Applying Lemma 2 therefore gives $\mathbb{E}_\Phi[N] \ge T/2$ for this larger class as well.

For $L2^{-b} \le \varepsilon \le 1/2$ and $T \ge 2L$, it follows that

$$\frac{\mathrm{SDC}^{\mathcal{Q}}_{1,L}(M^-_{w,L}; u, T)}{\mathrm{SDC}^{\mathcal{Q}}_{1,L}(M^+_{a,L}; u, T)} \ge \frac{(1-\varepsilon)LT}{2(T + L - 1)} = \Omega(L).$$

Every round emits between 1 and L tokens, so each SDC lies between $\lceil T/L \rceil$ and T. Consequently, the ratio is also at most L, proving the $\Theta(L)$ separation in the stated parameter range. For the description-length claim, choose $\varepsilon = 1/2$ and $b = \max\{6, \lceil \log_2(2L) \rceil\}$. Then $L2^{-b} \le 1/2$, and the constant number of parameters and readout entries, together with the fixed positional-parity rule, require $O(b) = O(\log L)$ bits. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and section 1 state $\sigma_L \le \Omega_L$, the fixed-LKM-residual counterexample, the full-sum unbounded separation in Proposition 4, the admissible-class SDC bounds, the path-level KL results, and the architecture separation. The uniform Θ characterization assumes exact optimization, $T \ge L$, a positive reachable-rate floor, and bounded $(\sigma^* + 1)/(\sigma_* + 1)$. Corollary 1 separately states the local conditional-overlap condition sufficient for sequential speedup. The architecture result retains the three-element proposal family and $L2^{-b} \le \varepsilon \le 1/2$, $T \ge 2L$.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.

- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: section 7 limits the SDC characterization to admissible clipped-ratio rounds and states its bounded reachable-rate heterogeneity condition. Tightness of Ω_L is left open specifically for the larger rejection-style class retaining the fixed LKM residual, distinct from general block coupling. The architecture separation uses only three proposals; richer families and empirical wall-clock performance are not analyzed. The local overlap condition in Corollary 1 is separate from the path-level Pinsker condition, and Theorem 6 rules out an unconditional KL converse. The analysis counts one-target-call verification rounds and does not cover general multi-call protocols.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The assumptions distinguish fixed-precision model tables and arithmetic from exact path probabilities and verifier randomization. SDC optimizes over conditionally admissible rounds. Theorem 3 gives explicit bounds for general reachable-rate ratios and requires bounded heterogeneity only for the uniform Θ conclusion; Theorem 2 treats ε_b as an assumed uniform optimization tolerance. Corollary 1 states reachable-state uniformity and its local overlap condition. Proposition 4 proves the complete-sum separation with fixed-precision conditionals. The architecture bound retains $L2^{-b} \le \varepsilon \le 1/2$ and accounts for finite-horizon overshoot; the $O(\log L)$-bit description claim uses the explicit $\varepsilon = 1/2$, $b = \max\{6, \lceil \log_2(2L) \rceil\}$ choice. Proofs appear in sections A to D.

   Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: The paper is a purely theoretical contribution. There are no experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: No experimental code or data are required for this purely theoretical paper.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper contains no experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The paper contains no experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The paper contains no experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is purely theoretical; it does not involve human subjects, datasets, or deployed systems. No conflict-of-interest, dual-use, or fairness concerns arise from establishing complexity-theoretic characterizations of speculative decoding speedup.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The framework characterizes target-verification-round complexity within the stated protocol class and identifies the corresponding sequential acceptance rate (section 7, "Implications for proposal choice"). Positive impact: deployed inference stacks can be analyzed with a principled theory, potentially reducing latency and energy cost. Potential negative impact: faster inference may accelerate the deployment of generative systems whose downstream uses carry their own societal questions; we view this as a second-order indirect effect of any efficiency-related result for Transformer inference.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper releases no models, datasets, or generative artifacts.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: No external assets (code, data, or models) are used in deriving the theoretical results.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper introduces no new datasets, models, or other reusable assets. Its sole contribution is theoretical.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper involves no human subjects and no crowdsourcing.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper involves no human subjects, so no IRB review is applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLM tools were used in a limited capacity to assist with writing, literature search, and exploratory problem-solving discussions during the preparation of this paper. They were not used as an important, original, or non-standard component of the core methodology, and all definitions, theorem statements, and proofs are the authors' own work.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.