# Mathematical Appendix: $A_0$ and RP for Autoregressive Transformer LLMs (LLaMA-class)

## How to Read This Appendix

This appendix provides a **formal, theorem–proof** foundation for the $A_0$ invariant and its operationalization as Reality Protocol (RP) in **discrete autoregressive transformer** language models (a LLaMA-class decoder-only transformer is used as the canonical representative). It is designed to complement:

- **Reality Protocol (RP): An $A_0$-Invariant Stabilization Framework for Generative Artificial Intelligence** (main paper)
- **Reality Protocol (RP): Clarifications, Scope, and Anticipated Critiques** (boundary/specification)

This appendix proves statements at the level of **mathematical abstraction** (Markov processes, Lyapunov functions, stability, and constrained local relaxation). It does **not** require access to model weights.

---

## 1. Preliminaries and Notation

### 1.1 Tokenization and Sequences

Let $\mathcal{V}$ be a finite vocabulary of tokens, and let $\mathrm{EOS} \in \mathcal{V}$ denote the termination token.

A generated output is a token sequence $x_{1:T}$ with $x_t \in \mathcal{V}$. Generation terminates when $x_t = \mathrm{EOS}$.

### 1.2 Autoregressive Model

An autoregressive language model defines conditional distributions

$$\pi_\theta(x_t \mid x_{<t}) \in \Delta(\mathcal{V}),$$

where $\Delta(\mathcal{V})$ is the probability simplex.

A decoder-only transformer (LLaMA-class) is a deterministic mapping from prefix $x_{<t}$ to logits $\ell_t \in \mathbb{R}^{|\mathcal{V}|}$, then

$$\pi_\theta(x_t \mid x_{<t}) = \mathrm{softmax}(\ell_t/\tau),$$

possibly followed by decoding transformations (temperature $\tau$, top-k/top-p, masking).

### 1.3 State Space

Define the **generation state** as

$$S_t := (x_{<t}, h_t),$$

where $h_t$ is any internal representation sufficient for next-token prediction (e.g., hidden activations). For theoretical results we may equivalently take $S_t := x_{<t}$ since $h_t$ is a deterministic function of the prefix for a fixed model and inference configuration.

Let $\mathcal{S}$ be the set of reachable states.

---

# 2. Decoding as a Controlled Markov Process

## 2.1 Controlled Transition Kernel

A decoding procedure (sampling, greedy, beam, etc.) determines a transition kernel

$$P(x_t = a \mid S_t) = \Pi(a \mid S_t),$$

where $\Pi$ is the effective policy induced by the base model $\pi_\theta$ and decoding rules.

The resulting process $\{S_t\}$ is a **controlled Markov process** with control realized by decoding constraints.

## 2.2 Termination as an Absorbing Set

Define the terminating set

$$\mathcal{S}_{\text{term}} := \{S : \text{last token is EOS}\}.$$

**Assumption A (Absorption).** Once EOS is emitted, the process remains terminated (no further tokens).

This is standard in LLM generation.

---

# 3. Admissibility and Local Cost (RP Core Objects)

RP is defined through (i) an admissible action set and (ii) a local transition cost.

## 3.1 Admissible Actions

For each state $S$, let $\mathcal{A}(S) = \mathcal{V}$ be the set of possible next tokens. Define an admissible subset

$$\mathcal{A}'(S) := \{a \in \mathcal{A}(S) : a \text{ satisfies admissibility constraints at } S\}.$$

Admissibility constraints are abstracted as a predicate $\mathrm{Adm}(S, a) \in \{0, 1\}$ and can include policy constraints, safety constraints, and stability constraints (see main paper).

### 3.2 Local Transition Cost

Define a local cost functional

$$\Xi(S, a) := \alpha Z(S, a) + \beta H(S, a) + \gamma T(S, a),$$

with $\alpha, \beta, \gamma \geq 0$.

Operational interpretations:

- $Z$: execution impedance (expected continuation length, token cost)
- $H$: informational entropy/drift (uncertainty inflation or semantic drift)
- $T$: constraint sensitivity/risk margin

This appendix does **not** require unique definitions of $Z, H, T$; it requires only mild regularity conditions stated below.

### 3.3 RP Local Relaxation Rule

The RP transition rule selects (or biases toward) an admissible minimizer:

$$\label{eq:rp} rpa^*(S) \in \arg\min_{a \in \mathcal{A}'(S)} \Xi(S, a). \tag{RP}$$

When $\mathcal{A}'(S) = \varnothing$, RP collapses to termination (silence/EOS).

---

## 4. The $A_0$ Invariant (Discrete Form)

### 4.1 Definition ($A_0$)

**$A_0$ (Minimal Local Discharge, Discrete).** A generated trajectory $\{S_t\}$ is **$A_0$-consistent** if at each step $t$, the realized transition $x_t$ is an admissible local minimizer of $\Xi$ at $S_t$, i.e., it satisfies \eqref{eq:rp} (or terminates when no admissible action exists).

Interpretation: $A_0$ states **local relaxation**, not global planning: $\arg\min$ is a fixed point of local admissible relaxation.

---

## 5. Stability Theorem: $A_0$ as Lyapunov-Style Descent

This section proves that if $\Xi$ induces a Lyapunov descent on a suitable potential, then non-$A_0$ trajectories are unstable (in the precise sense below).

### 5.1 Potential Function and Descent Condition

Define a nonnegative **instability potential** $\Phi : \mathcal{S} \to \mathbb{R}_{\geq 0}$. Think of $\Phi(S)$ as "local instability" or "structural contradiction pressure."

**Assumption B (Local Descent Link).** There exists $\Phi$ and a function $\Delta\Phi(S, a)$ such that for admissible actions:

$$\Delta\Phi(S, a) := \mathbb{E}[\Phi(S') \mid S, a] - \Phi(S)$$

and the cost $\Xi$ is **monotone aligned** with $\Delta\Phi$: for any admissible $a_1, a_2$,

$$\Xi(S, a_1) < \Xi(S, a_2) \implies \Delta\Phi(S, a_1) \leq \Delta\Phi(S, a_2).$$

This formalizes "lower impedance actions do not increase instability more than higher impedance actions." It is weaker than requiring explicit equality.

**Assumption C (Strict Progress When Possible).** If there exists an admissible action with $\Delta\Phi(S, a) < 0$, then any minimizer $a^*(S)$ satisfies $\Delta\Phi(S, a^*(S)) \leq 0$, with strict negativity for at least one step along any nonterminal trajectory until a stable set is reached.

### 5.2 Stable Set

Define the **stable set**

$$\mathcal{S}_* := \{S \in \mathcal{S} : \forall a \in \mathcal{A}'(S),\ \Delta\Phi(S, a) \geq 0\} \cup \mathcal{S}_{\text{term}}.$$

This includes states where no admissible action decreases instability, plus termination.

---

## 6. Main Theorem ($A_0$ Collapse of Unstable Trajectories)

### Theorem 1 ($A_0$ Descent and Convergence to Stable Set)

Under Assumptions A–C, any trajectory generated by the RP rule \eqref{eq:rp} satisfies:

1. $\Phi(S_t)$ is a supermartingale along the trajectory until $\mathcal{S}_*$ is reached:

$$\mathbb{E}[\Phi(S_{t+1}) \mid S_t] \leq \Phi(S_t).$$

1. The trajectory reaches $\mathcal{S}_*$ in finite time with probability 1, or converges to it in the limit (depending on whether $\Phi$ admits a positive drift barrier).

2. Any alternative policy that repeatedly chooses actions with strictly larger $\Xi$ when a minimizer exists has strictly higher expected $\Phi$ at some future time and therefore is unstable in the sense of not remaining within low-$\Phi$ neighborhoods.

**Proof**

(1) By Assumption B, selecting $a^*(S_t)$ that minimizes $\Xi$ yields an action whose $\Delta\Phi$ is minimal among admissible actions, hence $\Delta\Phi(S_t, a^*) \leq 0$ whenever any non-increasing option exists (Assumption C). Thus

$$\mathbb{E}[\Phi(S_{t+1}) \mid S_t] - \Phi(S_t) = \Delta\Phi(S_t, a^*) \leq 0,$$

so $\Phi$ is a supermartingale until $\mathcal{S}_*$.

(2) Since $\Phi \geq 0$ and is a supermartingale, it converges almost surely. If there is strict descent whenever outside $\mathcal{S}_*$ (Assumption C), the trajectory cannot remain indefinitely outside $\mathcal{S}_*$ without decreasing $\Phi$; thus it reaches or converges to $\mathcal{S}_*$.

(3) Consider a policy that deviates by selecting $\tilde{a}$ with $\Xi(S, \tilde{a}) > \Xi(S, a^*)$ in a state $S$ where a minimizer exists. By Assumption B, $\Delta\Phi(S, \tilde{a}) \geq \Delta\Phi(S, a^*)$, with strict inequality in at least one step in any sustained deviation scenario. Therefore the deviating policy yields higher expected $\Phi$ after some horizon, hence cannot stay within low-instability neighborhoods as reliably as the $A_0$ policy. □

---

# 7. Corollaries: RP Invariants as Consequences of Theorem 1

Theorem 1 provides a formal base for several RP invariants (see the main paper's IVL).

### Corollary 1 (Non-Selection)

Under the RP rule, behavior is determined by instability elimination: transitions are not "selected" as global optima; rather, trajectories with higher $\Xi$ induce higher expected $\Phi$ and fail to persist within stable neighborhoods.

### Corollary 2 (No-Agency)

The state evolution is fully specified by $(\mathcal{S}, \mathcal{A}', \Xi, \Pi)$. No additional primitive variable corresponding to "agency" appears in the dynamics or in the stability proof.

### Corollary 3 (Bounded Generation)

If $Z(S, a)$ includes an expected continuation-length penalty and admissibility excludes ungrounded drift (high $H$ without $\Phi$ reduction), then long narrative continuations have strictly higher $\Xi$ and therefore are suppressed; termination becomes a stable alternative when no admissible descent exists.

### Corollary 4 (Silence as Fixed Point)

If $\mathrm{EOS} \in \mathcal{A}'(S)$ whenever admissibility is indeterminate, and if $\Xi(S, \mathrm{EOS})$ is minimal among admissible actions (e.g., $Z$ and $H$ are minimal), then termination is an absorbing stable state in $\mathcal{S}_*$.

---

# 8. Mapping to LLaMA-Class Transformers (Canonical Instance)

This section shows that a LLaMA-class decoder-only transformer naturally fits the controlled Markov abstraction.

## 8.1 Locality

Next-token logits depend only on the current prefix and model parameters. Thus the transition kernel $\Pi(a \mid S)$ is locally defined.

## 8.2 Stochasticity

Sampling-based decoding induces a stochastic kernel. Greedy decoding is recovered as a deterministic special case.

## 8.3 Admissibility as Masking/Filtering

Admissibility predicates can be instantiated by:

- token masks (hard exclusions),
- logit penalties (soft exclusions),
- refusal/constraint boundaries.

This aligns with the admissibility-first ordering used in RP.

## 8.4 Constructing $Z, H, T$ from Decoding Primitives (Operational Choices)

The following operational instantiations are sufficient for the proof assumptions (not unique):

- $H(S, a)$: local uncertainty inflation proxy, e.g. logit dispersion or entropy of $\Pi(\cdot \mid S)$ conditioned on selecting $a$.
- $Z(S, a)$: expected additional token mass after choosing $a$, approximated by continuation heuristics (e.g., penalize open-ended discourse markers).
- $T(S, a)$: policy proximity margin, approximated by rule-based risk indicators or constraint-sensitive token classes.

These definitions need only preserve **ordering alignment** with $\Delta\Phi$ (Assumption B), not absolute correctness.

---

# 9. Prompt-Level RP as an Attractor (Soft-Field Approximation)

When RP is implemented purely as a system prompt (no external controller), the admissibility predicate $\mathrm{Adm}(S, a)$ and cost shaping $\Xi(S, a)$ are not computed numerically; they are induced **implicitly** by the model's learned conditional distribution under constraint pressure.

Formally, prompt-level RP is a **soft-field approximation** of \eqref{eq:rp}:

$$\Pi_{\mathrm{RP}}(a \mid S) \propto \Pi_0(a \mid S)\, \exp(-\lambda\, \widehat{\Xi}(S, a)),$$

where $\widehat{\Xi}$ is an implicit internal proxy induced by instruction-following and $\lambda$ is an effective strength of constraint pressure.

In this regime, Theorem 1 applies **conditionally**: to the extent that $\widehat{\Xi}$ preserves the ordering alignment of Assumption B, the process converges to low-instability regimes (attractors) and suppresses unstable expansions.

---

## 10. What This Appendix Does and Does Not Prove

### Proven (within stated assumptions)

- $A_0$ as a discrete local relaxation invariant.
- RP as admissibility-first, local-cost minimization.
- Stability/"collapse of unstable trajectories" in Lyapunov/supermartingale sense.
- Existence of stable sets including termination.
- RP core invariants as corollaries.

### Not claimed

- Universal correctness of any specific numerical $Z, H, T$ estimator.
- Guaranteed factual truth.
- Independence from external policies (policies may shrink $\mathcal{A}'$).

---

## 11. Minimal Checklist for a Fully Self-Contained Paper Proof

To make the main paper fully self-contained, include this appendix and ensure the main text:

1. States Assumptions A–C explicitly.
2. Defines $\Phi$ as an instability potential.
3. Defines admissibility and the RP rule.
4. Cites Theorem 1 as the formal basis for $A_0$ and IVL invariants.

---

## End

This appendix provides a complete mathematical scaffold for $A_0$/RP at the level of discrete stochastic dynamical systems, with LLaMA-class transformers as canonical instances via standard autoregressive decoding abstractions.