TOWARDS ROBUST REASONING TRAJECTORIES: CERTIFYING CHAIN-OF-THOUGHT PREFIX STABILITY UNDER LOCAL PERTURBATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Chain-of-thought (CoT) reasoning enables large language models (LLMs) to solve complex tasks through intermediate reasoning steps, but extended CoT generation also makes inference increasingly sensitive to local perturbations. Small changes in input prompts can alter the reasoning trajectory, making both its stability and computational behavior difficult to characterize. Existing work has largely studied this phenomenon empirically, leaving a fundamental question unresolved: how long can a CoT prefix be certified to remain unchanged under bounded perturbations? We address this question by formalizing trajectory stability as exact prefix preservation under bounded local representation perturbations. We derive finite-perturbation propagation bounds for causal Transformers and combine them with clean next-token decision margins to obtain sufficient conditions for preserving every token in a target reasoning prefix. The resulting certificate explicitly links perturbation location, perturbation magnitude, and certifiable prefix length, providing a trajectory-level robustness guarantee for autoregressive reasoning. We evaluate the theory on four reasoning models and three benchmarks, and observe stability patterns that consistently follow the predicted effects of perturbation location, magnitude, and prefix length. Together, our theoretical and empirical results provide a principled framework for certifying the robustness of reasoning trajectories in autoregressive language models.

1 INTRODUCTION

Chain-of-thought (CoT) has become an established approach for eliciting intermediate reasoning in large language models (Wei et al., 2022) and a central component of recent reasoning models (Guo et al., 2025; Yang et al., 2025). However, the resulting reasoning trajectories can be sensitive to local input changes (Liu et al., 2026). Controlled problem variants and localized perturbations have shown that small changes can redirect subsequent reasoning behavior (Mirzadeh et al., 2025; Li et al., 2024; Shi et al., 2023; Jeong et al., 2026). Because CoT generation is autoregressive, such changes can influence the context for subsequent token decisions and propagate through the reasoning trajectory, motivating trajectory-level guarantees of reasoning stability.

Recent work has begun to study this sensitivity at the trajectory level. Embedding interventions show how local changes can propagate and cause initially matching generations to diverge (Jaca et al., 2026), while interventions on an ongoing CoT reveal position-dependent recovery behavior (von Recum et al., 2026). These studies characterize trajectory changes empirically, but do not establish how much of the original reasoning trajectory is guaranteed to remain unchanged. Existing theoretical analyses provide important robustness perspectives, while a trajectory-level certificate that preserves perturbation location information in causal Transformers remains unexplored (Wang et al., 2026). This leaves a basic certification question unresolved:

Research question. *How long can the clean CoT prefix be certified to remain exactly unchanged under every local representation perturbation within a given budget at a specified input position?*

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

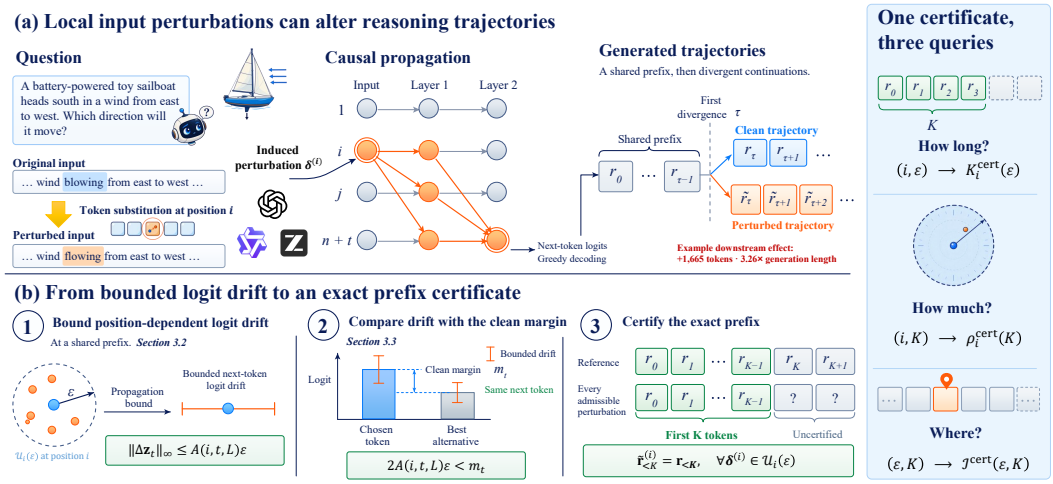

Figure 1: **Overview of CoT prefix certification under local input perturbations. (a)** A local perturbation may propagate through the causal Transformer and cause trajectory divergence. **(b)** Bounding logit drift against clean decision margins yields exact prefix guarantees on prefix length, perturbation radius, and input position.

To answer this question, we analyze deterministic greedy generation under a bounded local representation perturbation. As illustrated in Figure 1, we derive finite-perturbation bounds that retain the perturbed input position and control its effect on future next-token logits. By connecting bounded logit drift with clean next-token decision margins, we transform local perturbation propagation into an exact token-level trajectory certificate. Our analysis yields the following prefix certificate.

Theorem 1.1 (Informal prefix certificate; formalized in Section 3.3)**.** *Consider a fixed clean greedy trajectory for an input of length n and an L-layer model. Under the propagation conditions of Section 3.2 and local representation perturbations of norm at most ε at position $1 \le i < n$,*

$$K_i^{\mathrm{cert}}(\varepsilon) = \max\left\{ K : \varepsilon < c \min_{0 \le t < K} \frac{m_t (n+t-1)}{\left[1 + \log\left(\frac{n+t-1}{i}\right)\right]^{L-1}} \right\},$$

where m_t is the clean decision margin and $c > 0$ is a model-dependent constant; the maximum is over the generation horizon and is zero if no positive K qualifies. Every admissible perturbation preserves this prefix; Theorem 3.3 gives an explicit lower bound on its length.

The same certificate admits three complementary views: $K_i^{\mathrm{cert}}(\varepsilon)$ for fixed (i, ε), $\rho_i^{\mathrm{cert}}(K)$ for fixed (i, K), and $\mathcal{I}^{\mathrm{cert}}(\varepsilon, K)$ for fixed (ε, K). Together, they connect where a perturbation occurs, how large it is, and how long the reasoning trajectory can be certified unchanged.

We validate the predicted certificate geometry through the position–prefix and magnitude–prefix relationships across four reasoning models and three benchmarks. The positional ordering is consistent across models, tasks, and replacement draws, and persists under matched perturbation magnitudes. Prefix retention decreases systematically with perturbation magnitude across target prefix lengths. Finally, an end-to-end realization shows that the certificate yields explicit token-level guarantees.

Our main contributions are:

- We derive finite perturbation propagation bounds for causal Transformers with softmax attention, preserving the source location of local changes and connecting representation shifts to future token decisions.
- Building on this propagation bound, we introduce the first trajectory-level certification framework for exact CoT prefix preservation under bounded local perturbations, unifying prefix length, perturbation radius, and input position in a single certificate.
- We characterize the resulting certificate geometry across reasoning models and benchmarks, and demonstrate its concrete realization through explicit token-level certification on a causal Transformer.

2 Related Work

Robustness of CoT trajectories. Recent studies have investigated the robustness of chain-of-thought trajectories under local input or intermediate-state modifications. Controlled problem variants probe input sensitivity (Mirzadeh et al., 2025), while altered demonstrations study rationale relevance and step ordering (Wang et al., 2023). MATCHA analyzes rationale fragility through answer-conditioned perturbations (Jiang et al., 2026), and CoT interventions examine trajectory changes after intermediate reasoning modifications (von Recum et al., 2026; Zhu et al., 2026). Embedding interventions characterize alternative solutions and divergence behavior (Zhu et al., 2025; Jaca et al., 2026). Process evaluation and faithfulness studies further analyze reasoning quality and rationale reliability (Golovneva et al., 2023; Prasad et al., 2023; Zheng et al., 2025; Turpin et al., 2023; Lanham et al., 2023). These studies characterize how CoT trajectories vary under different interventions, providing an empirical foundation for studying reasoning robustness.

Propagation analysis and robustness guarantees. Theoretical analyses of Transformers study how information propagates through attention and sequence positions. Hahn (2020) bounds single-symbol influence in soft attention, while causal-path analyses characterize position-dependent sensitivity (Barbero et al., 2024; 2025). Other studies examine attention structure and positional effects (Wu et al., 2025; Herasimchyk et al., 2026), and Chowdhury (2026) derives influence kernels for simplified causal models. In reasoning models, Wang et al. (2026) relate perturbation tolerance to reasoning length and representation norms.

Certified robustness methods provide guarantees against bounded input changes, including word substitutions and embedding perturbations (Huang et al., 2019; Jia et al., 2019; Shi et al., 2020). Margin-based methods connect output sensitivity with unchanged predictions (Tsuzuku et al., 2018), while continuity analyses study next-token stability in decoder-only Transformers (Pasten et al., 2025). Unlike prior work in these settings, we derive finite-perturbation bounds that retain perturbation position and combine them with clean margins to certify exact prefixes of greedy CoT generation.

3 Robustness Guarantees for Autoregressive CoT Generation

Given a local representation perturbation of budget ε at input position i, our goal is to certify how many initial tokens of the clean trajectory remain exactly unchanged. For this purpose, we trace the perturbation through the causal Transformer to bound its effect on future next-token logits, and then compare this drift with clean decision margins to obtain an exact prefix guarantee.

3.1 Notation and Problem Setup

Autoregressive CoT Generation. Let $\mathcal{V}$ be the model vocabulary and let $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{V}^n$, with $n \geq 2$, denote the complete tokenized input, including task instructions and chat-template tokens. The decoder-only Transformer Z_θ computes next-token logits using input representations, L causal decoder layers, and an output projection. Appendix A.1 specifies the model used in the analysis. Input positions are indexed from one and generation steps from zero. Greedy decoding uses fixed deterministic tie-breaking throughout: $\mathbf{z}_t = Z_\theta(\mathbf{x}, \mathbf{r}_{<t})$ and $r_t = \arg\max_{v \in \mathcal{V}} \mathbf{z}_t(v)$, where $\mathbf{z}_t(v)$ is the logit of token v, $\mathbf{r}_{<t} = (r_0, \ldots, r_{t-1})$, and $\mathbf{r}_{<0} = \varnothing$. Token r_t is predicted from the last context position $n + t$.

Local Input Perturbations. We measure local input changes in the representations supplied to the decoder. Let $\mathbf{h}_j^{(\ell)} \in \mathbb{R}^d$ denote the hidden state at position j and layer $\ell \in \{0, \ldots, L\}$, with layer zero denoting the input to the first decoder block. Hidden states and perturbation vectors are column vectors. For the original input, write $\mathbf{H}^{(0)}(\mathbf{x}) = [\mathbf{h}_1^{(0)}(\mathbf{x}) \ \cdots \ \mathbf{h}_n^{(0)}(\mathbf{x})]^\top \in \mathbb{R}^{n \times d}$. Fix a position $i \in \{1, \ldots, n-1\}$ and a budget $\varepsilon \geq 0$. We restrict $i < n$ because our propagation bound applies to positions $j > i$, while the first generated token is predicted from position n. The admissible perturbations are

$$\mathcal{U}_i(\varepsilon) = \left\{ \boldsymbol{\delta}^{(i)} = [\boldsymbol{\delta}_1^{(i)} \ \cdots \ \boldsymbol{\delta}_n^{(i)}]^\top \in \mathbb{R}^{n \times d} : \boldsymbol{\delta}_j^{(i)} = \mathbf{0} \ (j \neq i), \quad \|\boldsymbol{\delta}_i^{(i)}\|_2 \leq \varepsilon \right\}. \tag{1}$$

Here $\boldsymbol{\delta}_j^{(i)} \in \mathbb{R}^d$ is the perturbation vector at position j, stored as row $(\boldsymbol{\delta}_j^{(i)})^\top$ of $\boldsymbol{\delta}^{(i)}$. For each $\boldsymbol{\delta}^{(i)} \in \mathcal{U}_i(\varepsilon)$, we generate from $\mathbf{H}^{(0)}(\mathbf{x}) + \boldsymbol{\delta}^{(i)}$, keeping the perturbation fixed on the original input positions throughout generation.

Let $\widetilde{\mathbf{x}}^{(i)}$ be the input obtained by replacing only x_i with another token, leaving all other tokens unchanged. Let $\boldsymbol{\delta}^{(i)} = \mathbf{H}^{(0)}(\widetilde{\mathbf{x}}^{(i)}) - \mathbf{H}^{(0)}(\mathbf{x})$. Its components are

$$\boldsymbol{\delta}_j^{(i)} = \begin{cases} \mathbf{h}_i^{(0)}(\widetilde{\mathbf{x}}^{(i)}) - \mathbf{h}_i^{(0)}(\mathbf{x}), & j = i, \\ \mathbf{0}, & j \neq i, \end{cases} \qquad D_i = \|\boldsymbol{\delta}_i^{(i)}\|_2. \tag{2}$$

The induced perturbation belongs to $\mathcal{U}_i(\varepsilon)$ whenever $D_i \leq \varepsilon$.

Exact Prefix Robustness. For each $\boldsymbol{\delta}^{(i)} \in \mathcal{U}_i(\varepsilon)$, let $\widetilde{\mathbf{z}}_t^{(i)}$ denote the next-token logits under the fixed layer-zero intervention $\mathbf{H}^{(0)}(\mathbf{x}) + \boldsymbol{\delta}^{(i)}$. Starting from $\widetilde{\mathbf{r}}_{<0}^{(i)} = \varnothing$, the perturbed trajectory is generated by $\widetilde{r}_t^{(i)} = \arg\max_{v \in \mathcal{V}} \widetilde{\mathbf{z}}_t^{(i)}(v)$. For the token-induced perturbation in Equation 2, this gives $\widetilde{\mathbf{z}}_t^{(i)} = Z_\theta(\widetilde{\mathbf{x}}^{(i)}, \widetilde{\mathbf{r}}_{<t}^{(i)})$, where $\widetilde{\mathbf{r}}_{<t}^{(i)} = (\widetilde{r}_0^{(i)}, \ldots, \widetilde{r}_{t-1}^{(i)})$.

For a target prefix length $K \in \mathbb{N}_+$ within the clean generation horizon, a position i, and a budget ε, we say that $\mathbf{x}$ is **(i, ε, K)-robust** if:

$$\widetilde{r}_t^{(i)} = r_t \quad \text{for every } \boldsymbol{\delta}^{(i)} \in \mathcal{U}_i(\varepsilon) \text{ and } 0 \leq t < K. \tag{3}$$

We treat termination as part of the token-level decision, so a termination mismatch counts as a divergence within the target prefix horizon.

The length of the unchanged prefix is determined by the first divergence. For a fixed clean input $\mathbf{x}$ and perturbation $\boldsymbol{\delta}^{(i)}$, define

$$\tau(\boldsymbol{\delta}^{(i)}) = \inf\{t \in \mathbb{N}_0 : r_t \neq \widetilde{r}_t^{(i)}\}. \tag{4}$$

Here $\tau(\boldsymbol{\delta}^{(i)}) = \infty$ if no divergence occurs, including when both runs terminate identically. A finite value k means that the first k tokens agree and the next decisions differ. When comparing inputs, we write it as $\tau(\mathbf{x}; \boldsymbol{\delta}^{(i)})$. Thus (i, ε, K)-robustness is equivalent to $\inf_{\boldsymbol{\delta}^{(i)} \in \mathcal{U}_i(\varepsilon)} \tau(\boldsymbol{\delta}^{(i)}) \geq K$. The outer infimum selects the earliest divergence over all admissible perturbations. Before this divergence, the two runs share the same generated prefix, allowing their computations to be compared with this prefix held fixed.

Problem Setup. For a fixed model and input $\mathbf{x}$, we seek conditions on the position i, budget ε, and target prefix length K that guarantee (i, ε, K)-robustness for $\mathbf{x}$. The key step is to connect a bound on representation changes to agreement of discrete token decisions.

3.2 Propagation of Local Input Perturbations

We first trace the layer-zero perturbation through attention over earlier positions $w < j$, while the generated prefixes agree. The local state enters the layer update separately. Appendix A.3 extends the bound to attention that also includes the current position, with the corresponding assumptions and constants.

Propagation Conditions. Fix a clean input and a bounded perturbation domain containing zero. We use common constants over this domain, the input positions considered, and the common-prefix steps in a fixed finite generation horizon. For each budget considered, the domain contains $\mathcal{U}_i(\varepsilon)$ at every input position being compared.

The layer update is Lipschitz in its local state and attention outputs (A1), and hidden-state norms and value-projection operator norms are bounded (A2). Attention scores are Lipschitz in the query and key states (A3). To control how influence spreads across positions, we also bound each attention weight at $j \geq 2$ by $c_\alpha/(j-1)$, for a constant $c_\alpha \geq 1$. This bound holds along the line segment between the clean and perturbed score vectors (A4). Appendix A.1 states A1–A4 formally.

For a fixed input perturbation $\boldsymbol{\delta}^{(i)}$, let $\Delta\mathbf{h}_j^{(\ell)}(\mathbf{x})$ denote the perturbed hidden state minus the clean hidden state at the same generated prefix. For the token substitution above, the resulting layer-wise hidden-state difference is given by $\Delta\mathbf{h}_j^{(\ell)}(\mathbf{x}) = \mathbf{h}_j^{(\ell)}(\widetilde{\mathbf{x}}^{(i)}) - \mathbf{h}_j^{(\ell)}(\mathbf{x})$.

To retain the dependence on the source position, define for $j > i$ $R_i(j) = 1 + \log \frac{j-1}{i}$. For a fixed downstream position j, $R_i(j)$ decreases as the source i approaches j. The following theorem shows how this position factor enters the propagation bound at each decoder layer.

Theorem 3.1 (Local perturbation propagation)**.** *Under A1–A4, each layer $\ell \in \{1, \dots, L\}$ admits a constant $C_\ell > 0$ depending on the layer and the uniform assumption constants. Let $\varepsilon \geq 0$, $i \in \{1, \dots, n-1\}$, and $\boldsymbol{\delta}^{(i)} \in \mathcal{U}_i(\varepsilon)$. At any step t satisfying $\widetilde{\mathbf{r}}_{<t}^{(i)} = \mathbf{r}_{<t}$, every position j with $i < j \leq n + t$ satisfies*

$$\left\| \Delta \mathbf{h}_j^{(\ell)}(\mathbf{x}) \right\|_2 \leq C_\ell \left\| \boldsymbol{\delta}_i^{(i)} \right\|_2 \frac{R_i(j)^{\ell-1}}{j-1} \leq C_\ell \varepsilon \frac{R_i(j)^{\ell-1}}{j-1}. \tag{5}$$

The bound separates two effects. The factor $1/(j-1)$ captures the dilution of a single source among earlier positions under diffuse causal attention. The factor $R_i(j)^{\ell-1}$ captures the accumulation of influence through intermediate positions across deeper layers. For a fixed downstream position j, an earlier source leaves a longer range of intermediate positions over which such influence can accumulate, yielding a larger position factor. This makes the source position explicit in the propagation bound. The first inequality retains the realized perturbation magnitude $\|\boldsymbol{\delta}_i^{(i)}\|_2$, while replacing it with ε gives a uniform bound over $\mathcal{U}_i(\varepsilon)$. Appendix A.3 gives the proof.

At generation step t, the next token is predicted from position $n + t$. Applying the propagation bound at this position and passing it through the output map gives a position-dependent bound on the next-token logits $\mathbf{z}_t = \mathbf{W}_{\text{out}} g\left(\mathbf{h}_{n+t}^{(L)}(\mathbf{x})\right) + \mathbf{b}_{\text{out}}$. Here g is the transformation before the output projection, such as a final normalization layer. It is the identity if the projection uses the hidden state directly.

Suppose g is L_g-Lipschitz in the Euclidean norm and define $C_{\text{out}} = L_g \max_{v \in \mathcal{V}} \|(\mathbf{W}_{\text{out}})_{v,:}\|_2$. Here $(\mathbf{W}_{\text{out}})_{v,:}$ is the output-matrix row for token v. The constant C_{out} bounds how much a final-state change can affect any individual logit. Combining it with the propagation bound at $j = n + t$ gives the coefficient

$$A(i, t, L) = C_{\text{out}} C_L \frac{\left(1 + \log \frac{n+t-1}{i}\right)^{L-1}}{n + t - 1}. \tag{6}$$

For fixed t and common model constants, $A(i, t, L)$ decreases with i when $L \geq 2$, yielding tighter logit-drift bounds at later source positions.

Corollary 3.2 (Local perturbation logit bound)**.** *Let $\Delta \mathbf{z}_t(\mathbf{x}) = \widetilde{\mathbf{z}}_t^{(i)} - \mathbf{z}_t$. Under the conditions of Theorem 3.1 and the Lipschitz condition on g above, at any generation step t for which $\widetilde{\mathbf{r}}_{<t}^{(i)} = \mathbf{r}_{<t}$, we have*

$$\boxed{\left\| \Delta \mathbf{z}_t(\mathbf{x}) \right\|_\infty \leq A(i, t, L) \left\| \boldsymbol{\delta}_i^{(i)} \right\|_2 \leq A(i, t, L) \varepsilon}. \tag{7}$$

3.3 Prefix Stability Certificates

A logit bound yields a token-level guarantee when the drift is sufficiently small relative to the clean model's preference for its selected token. At generation step t, define the *clean decision margin*

$$m_t = \mathbf{z}_t(r_t) - \max_{v \neq r_t} \mathbf{z}_t(v) \geq 0. \tag{8}$$

Combining the clean decision margins with the position-dependent logit bound in Corollary 3.2 yields a unified prefix certificate. We characterize this certificate along three complementary dimensions: certified prefix length, certified perturbation radius, and certified input positions.

For a target prefix length K, let $a_K = \min_{0 \leq t < K} m_t$. To expose the dependence on the generation step, define

$$f_i(t) = \frac{\left(1 + \log \frac{n+t-1}{i}\right)^{L-1}}{n + t - 1}. \tag{9}$$

Then $A(i, t, L) = C_{\text{out}} C_L f_i(t)$, and a uniform sufficient condition for preserving the first K generated tokens is

$$2 C_{\text{out}} C_L \varepsilon f_i(t) < a_K, \qquad 0 \leq t < K. \tag{10}$$

Certified prefix length under a fixed position and perturbation budget. For fixed i and ε, we solve the prefix condition for the maximum certified prefix length. Let $K_{\max}$ denote the maximum prefix length considered within the fixed analysis horizon. For fixed i and ε, define

$$K_i^{\mathrm{cert}}(\varepsilon) = \max\left(\{0\} \cup \{K \in \{1, \dots, K_{\max}\} : 2A(i,t,L)\varepsilon < m_t,\ \forall\, 0 \le t < K\}\right). \tag{11}$$

For $L \ge 2$, the continuous extension of $f_i(t)$ is unimodal with stationary point $t_i^\star = ie^{L-2} - n + 1$. Hence $f_i(t)$ is decreasing when $t_i^\star \le 0$, increasing and then decreasing when $0 < t_i^\star < K_{\max} - 1$, and increasing throughout the analysis horizon when $t_i^\star \ge K_{\max} - 1$.

If the full-horizon condition in Equation 10 does not hold but the condition at $t = 0$ does, let $t_i^\dagger$ denote the first point on the increasing branch satisfying $2C_{\mathrm{out}}C_L\varepsilon f_i(t_i^\dagger) = a_{K_{\max}}$.

Theorem 3.3 (Certified prefix length)**.** *Assume $L \ge 2$, and that the propagation conditions and the Lipschitz condition on g hold throughout the fixed analysis horizon. For $\varepsilon > 0$, we have*

$$\boxed{K_i^{\mathrm{cert}}(\varepsilon) \ge \begin{cases} K_{\max}, & 2C_{\mathrm{out}}C_L\varepsilon \max\limits_{t\in\{0,\dots,K_{\max}-1\}} f_i(t) < a_{K_{\max}}, \\ \left\lceil t_i^\dagger \right\rceil, & 2C_{\mathrm{out}}C_L\varepsilon f_i(0) < a_{K_{\max}} \le 2C_{\mathrm{out}}C_L\varepsilon \max\limits_{t\in\{0,\dots,K_{\max}-1\}} f_i(t), \\ 0, & \textit{otherwise}. \end{cases}} \tag{12}$$

Moreover, $\tau(\boldsymbol{\delta}^{(i)}) \ge K_i^{\mathrm{cert}}(\varepsilon)$ for every $\boldsymbol{\delta}^{(i)} \in \mathcal{U}_i(\varepsilon)$.

For a fixed perturbation budget, later input positions reduce $f_i(t)$ and therefore cannot yield a shorter certified prefix under this sufficient condition. The first case certifies the complete analysis horizon. In the second case, the uniform-margin bound certifies at least $\lceil t_i^\dagger \rceil$ tokens before reaching its first threshold crossing. The last case indicates that this relaxation provides no nontrivial positive lower bound. Appendix A.7 gives the monotonicity analysis and the explicit solution for $t_i^\dagger$.

Certified perturbation radius under a fixed position and target prefix length. For fixed i and K, we solve the same condition for the certified perturbation-radius threshold.

Proposition 3.1 (Certified perturbation radius)**.** *Under the propagation conditions and the Lipschitz condition on g over the target prefix horizon,*

$$\boxed{\rho_i^{\mathrm{cert}}(K) = \min_{0\le t<K} \frac{m_t}{2A(i,t,L)} \ge \frac{a_K}{2C_{\mathrm{out}}C_L \max\limits_{t\in\{0,\dots,K-1\}} f_i(t)}.} \tag{13}$$

Therefore, if

$$\varepsilon < \frac{a_K}{2C_{\mathrm{out}}C_L \max\limits_{t\in\{0,\dots,K-1\}} f_i(t)},$$

then $\tau(\boldsymbol{\delta}^{(i)}) \ge K$ for every $\boldsymbol{\delta}^{(i)} \in \mathcal{U}_i(\varepsilon)$.

For fixed i and K, perturbations are certified whenever $\varepsilon < \rho_i^{\mathrm{cert}}(K)$; the sufficient radius bound cannot increase as the target prefix length grows. The maximum of $f_i(t)$ is determined by $t_i^\star = ie^{L-2} - n + 1$: it occurs at $t = 0$ when $t_i^\star \le 0$, at $t = K - 1$ when $t_i^\star \ge K - 1$, and otherwise at one of the two integers nearest to $t_i^\star$.

Certified input positions under a fixed perturbation budget and target prefix length. For fixed ε and K, we solve the condition for the input positions whose perturbations can be certified:

$$\mathcal{I}^{\mathrm{cert}}(\varepsilon, K) = \left\{i \in \{1, \dots, n-1\} : \varepsilon < \rho_i^{\mathrm{cert}}(K)\right\}. \tag{14}$$

Solving Equation 10 for i gives an explicit sufficient range of certified input positions.

Proposition 3.2 (Certified input positions)**.** *Assume $L \ge 2$, $\varepsilon > 0$, and that the propagation conditions and the Lipschitz condition on g hold throughout the target prefix horizon. If*

$$\boxed{i > \max_{t\in\{0,\dots,K-1\}} \left\{(n+t-1)\exp\left[1 - \left(\frac{a_K(n+t-1)}{2C_{\mathrm{out}}C_L\varepsilon}\right)^{\frac{1}{L-1}}\right]\right\},} \tag{15}$$

then $i \in \mathcal{I}^{\mathrm{cert}}(\varepsilon, K)$, and hence $\tau(\boldsymbol{\delta}^{(i)}) \ge K$ for every $\boldsymbol{\delta}^{(i)} \in \mathcal{U}_i(\varepsilon)$.

For fixed ε and K, the proposition gives a lower threshold on i, so the positions certified by this condition form a contiguous suffix of the input. The maximization over t is analyzed in Appendix A.7.

Remark 1. *The analysis also extends to simultaneous perturbations at multiple input positions. When A1–A4 and the output-map Lipschitz condition hold uniformly over the joint perturbation domain and all common-prefix steps considered, the logit drift is bounded by the sum of the position-weighted perturbation magnitudes. The same margin argument gives a prefix guarantee. Appendix A.8 states and proves the extension.*

4 EXPERIMENTS

Section 3.3 establishes how the sufficient conditions for exact prefix stability depend on input position i and perturbation magnitude ε. We evaluate two central structural predictions of this framework across reasoning models and benchmarks:

- **RQ1.** How does prefix stability vary with input position, and does the positional pattern persist under matched perturbation magnitudes?
- **RQ2.** How does prefix stability vary with perturbation magnitude across target prefix lengths K?

Section 4.2 addresses RQ1 and Section 4.3 addresses RQ2; we then instantiate the full certificate end to end, from explicit propagation constants to exact token-level prefix guarantees.

4.1 EXPERIMENTAL DESIGN

Models and benchmarks. We evaluate four reasoning models: Qwen3-8B (Yang et al., 2025), DeepSeek-R1-Distill-Qwen-14B (Guo et al., 2025), GLM-Z1-9B-0414 (GLM et al., 2024), and GPT-OSS-20B (OpenAI, 2025). Experiments cover ARC-Challenge (Clark et al., 2018), GSM8K (Cobbe et al., 2021), and MMLU-Pro (Wang et al., 2024), yielding twelve model–benchmark settings.

Perturbation protocol. For each example, we separately perturb one token at each of five question locations and record its input position i. A complete substitution induces a representation perturbation $\boldsymbol{\delta}^{(i)}$ with magnitude $D_i = \|\boldsymbol{\delta}_i^{(i)}\|_2$. Position analysis uses two replacement draws. Magnitude analysis scales the primary replacement direction to $\varepsilon/D_i \in \{0.25, 0.5, 0.75, 1\}$. Matched-magnitude controls are described in Appendix B.4.

Prefix-stability metrics. We quantify trajectory stability using the first-divergence index τ from Equation 4. The median τ measures the realized preserved-prefix length, and the prefix-retention rate $\Pr[\tau \geq K]$ measures the fraction of trajectories preserving a target prefix length K. All clean and perturbed trajectories are generated using greedy decoding. Aggregation details are provided in Appendix B.2.

4.2 POSITION–PREFIX RELATIONSHIP

Position dependence under complete substitutions. Later input positions preserve substantially longer prefixes under complete token substitutions (Figure 2). Across all twelve model–benchmark settings, the median preserved-prefix length increases strictly over the five tested locations for both replacement draws. Under the primary draw, the earliest-to-latest median increase ranges from 46 to 111 generated tokens. The agreement across models, tasks, and replacement draws reveals a highly consistent positional ordering of reasoning-prefix stability.

Matched-magnitude validation. The positional pattern remains strong when the absolute representation-space perturbation magnitude is matched across input locations. At $K = 64$ and $K = 128$, pooled prefix retention increases monotonically from earlier to later locations at all four magnitude levels (Table 1). The late-minus-early gap ranges from 5.7 to 60.3 percentage points at $K = 64$ and from 13.3 to 47.0 at $K = 128$. Across 48 benchmark–magnitude–target-length configurations, 47 endpoint contrasts are positive and one is zero (Appendix B.4).

Together, the complete-substitution and matched-magnitude results exhibit a positional pattern consistent with the sufficient conditions in Theorem 3.3 and Proposition 3.2: later input positions are associated with stronger prefix stability under the tested perturbations.

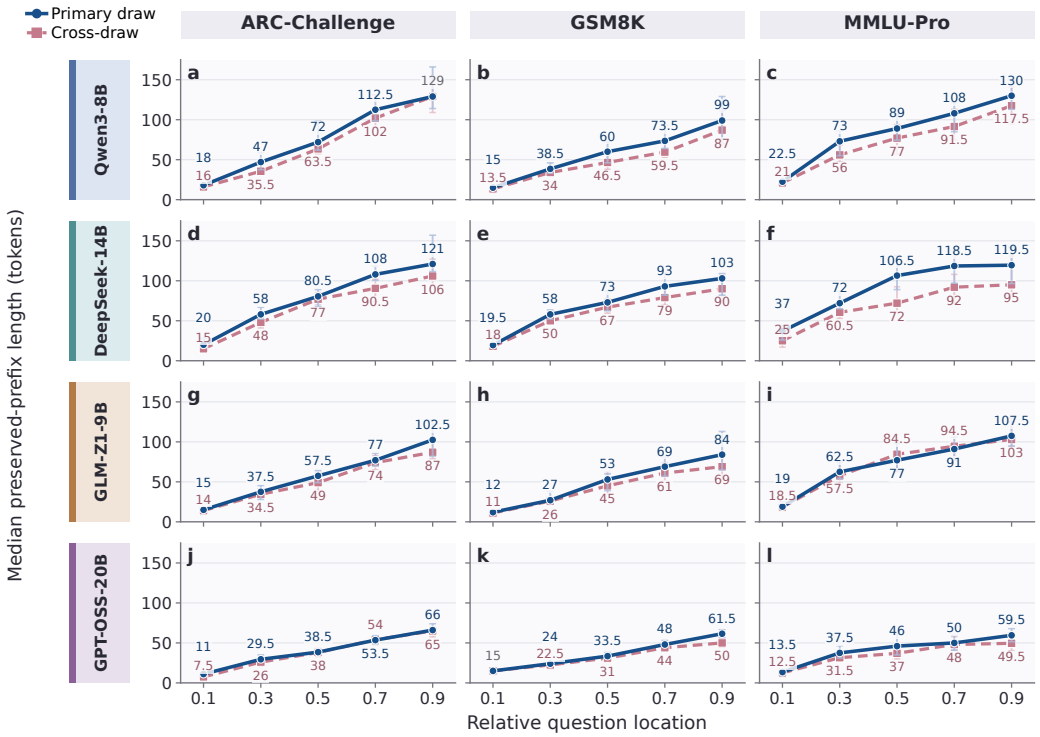

Figure 2: **Preserved prefix length across input positions.** Each panel shows one model–benchmark setting. Solid and dashed curves report the median preserved-prefix length under the primary and cross-draw complete substitutions, respectively, over five locations spanning the user question from early to late. Each selected token is mapped to its actual position i in the complete formatted input before intervention.

Table 1: **Prefix retention under matched perturbation magnitudes.** Results for Qwen3-8B pooled across three benchmarks. Values are prefix retention rates (%); gaps report late-minus-early differences, with absolute perturbation magnitude matched across positions within each example.

	K = 64						**K = 128**					
$\varepsilon_{\mathbf{x}}/D_{\min}(\mathbf{x})$	**0.1**	0.3	0.5	0.7	**0.9**	**Gap (pp)**	**0.1**	0.3	0.5	0.7	**0.9**	**Gap (pp)**
0.25	91.7	93.3	95.7	96.3	97.3	**5.7**↑	82.3	87.3	90.3	92.3	95.7	**13.3**↑
0.50	72.3	83.3	88.0	91.7	93.7	**21.3**↑	55.7	68.0	73.7	82.7	86.0	**30.3**↑
0.75	33.3	49.3	63.7	73.3	82.7	**49.3**↑	16.7	28.3	39.0	48.7	62.3	**45.7**↑
1.00	13.0	40.3	51.3	66.3	73.3	**60.3**↑	4.7	19.7	27.3	38.0	51.7	**47.0**↑

4.3 Magnitude–Prefix Relationship

Increasing perturbation magnitude consistently reduces prefix retention across target prefix lengths (Figure 3). All 48 model–benchmark–K profiles have lower retention at $\varepsilon = D_i$ than at $\varepsilon = 0.25D_i$, and 44 decrease strictly across all four tested magnitude levels. At $K = 64$, pooled retention falls from 76.5% to 39.4% as the perturbation increases from $0.25D_i$ to D_i. Across target horizons, retention at the complete-substitution magnitude decreases from 82.3% for $K = 16$ to 16.0% for $K = 128$. These results are consistent with the magnitude dependence captured by Proposition 3.1: larger perturbation budgets reduce the range of prefixes that can be preserved, while longer target prefixes impose increasingly stringent stability requirements.

4.4 Certificate Realization

The preceding experiments validate the predicted stability patterns. We next instantiate the certificate using explicit propagation constants and clean decision margins for a two-layer decoder-only

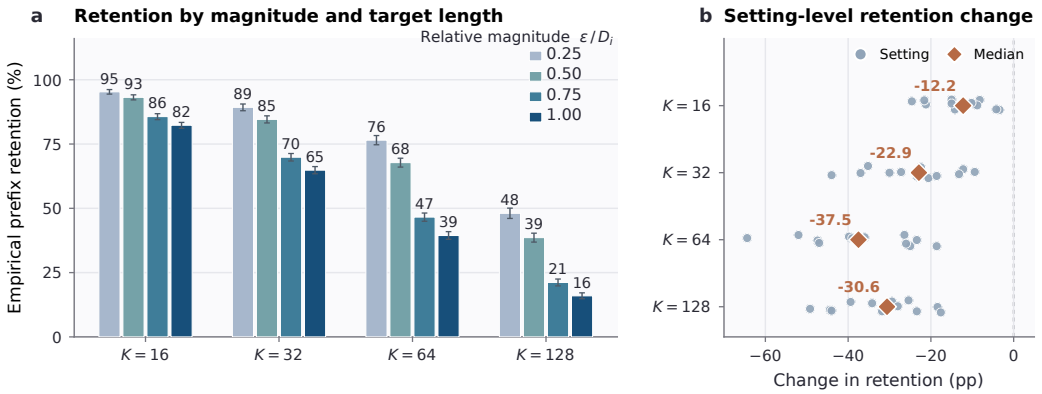

Figure 3: **Prefix retention across perturbation magnitudes and target lengths.** (a) Mean retention across twelve model–benchmark settings and five input positions at $\varepsilon/D_i \in \{0.25, 0.5, 0.75, 1\}$. (b) Setting-level changes from $\varepsilon = 0.25D_i$ to $\varepsilon = D_i$.

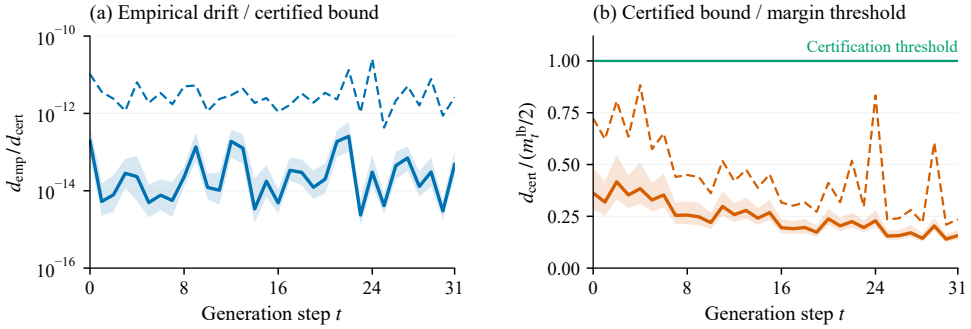

Figure 4: **Mechanism of the prefix certificate.** (a) Empirical drift relative to the certified bound. (b) Certified bound relative to the certified margin threshold. Solid curves denote medians, shaded regions denote interquartile ranges, and dashed curves denote maxima over example–position pairs.

causal Transformer trained on a synthetic multi-step arithmetic task (Appendix C). Figure 4 illustrates the intermediate quantities that connect perturbation propagation to token-level prefix certification. The empirical drift remains below the certified bound at every generation step, while the bound remains below the certified margin threshold throughout the trajectory. Together, these results realize the propagation and prefix guarantees established by Theorem 3.1 and Theorem 3.3. Across 168 held-out examples and eight admissible input positions, we compute certified radii for all 1,344 example–position pairs at $K \in \{4, 8, 16, 32\}$. A common perturbation budget certifies exact preservation of the full 32-token trajectory for every evaluated pair. We further evaluate the realized certificates through 12,480 direct generation checks within the certified radii and observe no prefix violations. Full numerical certification statistics and bound-comparison results are provided in Appendix C.

5 Conclusion

We develop a certification framework for exact CoT prefix preservation under bounded local representation perturbations. By retaining perturbation position in finite-perturbation propagation bounds and comparing logit drift with clean decision margins, we derive sufficient conditions for preserving an exact autoregressive reasoning prefix. The certificate admits three complementary views: certified prefix length, perturbation radius, and input position. Experiments across models and tasks exhibit the predicted positional and magnitude dependencies, while an end-to-end realization demonstrates that the guarantees can be concretely instantiated. These findings suggest that robustness assessment for long-horizon reasoning systems should account for where perturbations occur, how large they are, and how long the reasoning process must remain stable. Together, these results move reasoning-trajectory robustness beyond empirical sensitivity analysis toward token-level guarantees, establishing exact prefix stability as a certifiable property of autoregressive reasoning.

STATEMENTS

AI USE STATEMENT

In this work, we used generative AI tools for writing assistance, literature organization, and manuscript preparation. All content was independently verified by the authors, who take full responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

Complete proofs and assumptions are provided in the appendix, together with the experimental setup, evaluation protocols, and implementation details needed to reproduce the theoretical and empirical results.

REFERENCES

Federico Barbero, Andrea Banino, Steven Kapturowski, Dharshan Kumaran, João G.M. Araújo, Alex Vitvitskyi, Razvan Pascanu, and Petar Veličković. Transformers need glasses! information over-squashing in language tasks. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 98111–98142. Curran Associates, Inc., 2024. doi: 10.52202/079017-3114. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/b1d35561c4a4a0e0b6012b2af531e149-Paper-Conference.pdf.

Federico Barbero, Alvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Petar Veličković, Razvan Pascanu, and Michael M. Bronstein. Why do LLMs attend to the first token? In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=tu4dFUsW5z.

Borun D Chowdhury. Lost in the middle at birth: An exact theory of transformer position bias, 2026. URL https://arxiv.org/abs/2603.10123.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. doi: 10.48550/arXiv.2110.14168. URL https://arxiv.org/abs/2110.14168.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024. URL https://arxiv.org/abs/2406.12793.

Olga Golovneva, Moya Peng Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. ROSCOE: A suite of metrics for scoring step-by-step reasoning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=xYlJRpzZtsY.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei

Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645 (8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Michael Hahn. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171, 2020. doi: 10.1162/tacl_a_00306. URL https://aclanthology.org/2020.tacl-1.11/.

Hanna Herasimchyk, Robin Labryga, Tomislav Prusina, and Sören Laue. A structural theory of position bias in transformers, 2026. URL https://arxiv.org/abs/2602.16837.

Po-Sen Huang, Robert Stanforth, Johannes Welbl, Chris Dyer, Dani Yogatama, Sven Gowal, Krishnamurthy Dvijotham, and Pushmeet Kohli. Achieving verified robustness to symbol substitutions via interval bound propagation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4083–4093, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1419. URL https://aclanthology.org/D19-1419/.

Gregorio Jaca, Kristóf Benedek, and János Török. Chaos in reason: How chain-of-thought llms can look for an answer, 2026. URL https://arxiv.org/abs/2607.27805.

Seungwon Jeong, Jiwoo Jeong, Hyeonjin Kim, Yunseok Lee, and Woojin Lee. Slot-gcg: Exploiting the positional vulnerability in llms for jailbreak attacks. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 69776–69805, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/71ec377d5df1fc61ee7770857820519b-Paper-Conference.pdf.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. Certified robustness to adversarial word substitutions. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4129–4142, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1423. URL https://aclanthology.org/D19-1423/.

Enyi Jiang, Changming Xu, Nischay Singh, and Gagandeep Singh. Robust answers, fragile logic: Probing the decoupling hypothesis in LLM reasoning. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=pMhTFUdM4G.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning, 2023. URL https://arxiv.org/abs/2307.13702.

Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. GSM-plus: A comprehensive benchmark for evaluating the robustness of LLMs as mathematical problem solvers. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2961–2984, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.163. URL https://aclanthology.org/2024.acl-long.163/.

Shuaitong Liu, Renjue Li, Lijia Yu, Lijun Zhang, Zhiming Liu, and Gaojie Jin. Badthink: Triggered overthinking attacks on chain-of-thought reasoning in large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 32141–32149, 2026. URL https://doi.org/10.1609/aaai.v40i38.40486.

Iman Mirzadeh, Keivan Alizadeh-Vahid, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 94743–94765, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/ec2e7a896f8250986b3907f57621ce94-Paper-Conference.pdf.

OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

Hector Pasten, Felipe Urrutia, Hector Orellana, Cristian Buc Calderon, Cristobal Rojas, and Alexander Kozachinskiy. Continuity and isolation lead to doubts or dilemmas in large language models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 17725–17753. Curran Associates, Inc., 2025. doi: 10.52202/085713-0599. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/19ded4cfc36a7feb7fce975393d378fd-Paper-Conference.pdf.

Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. ReCEval: Evaluating reasoning chains via correctness and informativeness. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10066–10086, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.622. URL https://aclanthology.org/2023.emnlp-main.622/.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31210–31227. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/shi23a.html.

Zhouxing Shi, Huan Zhang, Kai-Wei Chang, Minlie Huang, and Cho-Jui Hsieh. Robustness verification for transformers. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJxwPJHFwS.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/485843481a7edacbfce101ecb1e4d2a8-Paper.pdf.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 74952–74965. Curran Associates, Inc., 2023. doi: 10.52202/075280-3275. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ed3fea9033a80fea1376299fa7863f4a-Paper-Conference.pdf.

Alexander von Recum, Leander Girrbach, and Zeynep Akata. Are reasoning llms robust to interventions on their chain-of-thought? In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 98544–98595, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/a03037317560b8c5f2fb4b6466d4c439-Paper-Conference.pdf.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. Towards understanding chain-of-thought prompting: An empirical study of what matters. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2717–2739, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.153. URL https://aclanthology.org/2023.acl-long.153/.

Dingzirui Wang, Xuanliang Zhang, Keyan Xu, Qingfu Zhu, Wanxiang Che, and Yang Deng. Bounds of chain-of-thought robustness: Reasoning steps, embed norms, and beyond. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 67627–67655, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/6e72128ef96c9ede89e541f42d21f3aa-Paper-Conference.pdf.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 95266–95290. Curran Associates, Inc., 2024. doi: 10.52202/079017-3018. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ad236edc564f3e3156e1b2feafb99a24-Paper-Datasets_and_Benchmarks_Track.pdf.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

Xinyi Wu, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. On the emergence of position bias in transformers. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 67756–67781. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/wu25ad.html.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. ProcessBench: Identifying process errors in mathematical reasoning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1009–1024, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.50. URL https://aclanthology.org/2025.acl-long.50/.

Qinglin Zhu, Runcong Zhao, Hanqi Yan, Yulan He, Yudong Chen, and Lin Gui. Soft reasoning: Navigating solution spaces in large language models through controlled embedding exploration. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 80427–80447. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/zhu25ae.html.

Wei Zhu, Jian Zhang, Lixing Yu, Kun Yue, and Zhiwen Tang. Dissecting failure dynamics in large language model reasoning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8893–8914, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.401. URL https://aclanthology.org/2026.acl-long.401/.

APPENDIX CONTENTS

Appendix A. Proofs for Robustness Bounds under Local Input Perturbations 16

A.1 Model and Uniform Assumptions 16
A.2 Auxiliary Bounds 17
A.3 Hidden-State Propagation 19
A.4 From Hidden-State Propagation to Logit Drift 21
A.5 Decision-Margin Stability 22
A.6 Exact Prefix Robustness 22
A.7 Closed-Form Analysis of the Prefix Certificates 23
A.8 Simultaneous Input Perturbations 25
A.9 Interpretation of the Bound 27

Appendix B. Additional Experimental Details 28

B.1 Experimental Protocol and Position Mapping 28
B.2 Aggregation Details 29
B.3 Position–Prefix: Complete-Substitution Results 29
B.4 Position–Prefix: Matched-Magnitude Validation 30
B.5 Magnitude–Prefix Results 32
B.6 Additional Position–Magnitude Interaction 33

Appendix C. Certificate Realization Details 34

C.1 Model and Multi-Step Reasoning Task 34
C.2 Certificate Computation and Numerical Verification 35
C.3 Tightness and Position-Awareness of the Certificate 36
C.4 Certified Radius and Prefix Guarantees 38
C.5 Independent Generation Checks 39

A Proofs for Robustness Bounds under Local Input Perturbations

We prove the results in Section 3 for deterministic greedy generation under layer-zero input perturbations. We first derive the position-dependent hidden-state propagation bound and map it to next-token logit drift. Comparing this drift with the clean decision margin yields a one-step token guarantee, which is extended to the generated prefix by induction. The main derivation uses attention over earlier positions, with the local state entering the layer update separately; Appendix A.3 also gives the corresponding extension to attention that includes the current position.

Fix a clean input $\mathbf{x}$, a bounded perturbation domain containing zero, and a finite generation horizon. The assumptions below hold with common constants over this domain, the input positions considered, and all common-prefix steps in this horizon. For each budget considered, $\mathcal{U}_i(\varepsilon)$ lies in the domain at every input position being compared.

A.1 Model and Uniform Assumptions

The model Z_θ has an input representation map, L causal decoder layers with H attention heads each, and a next-token output map. Layer-zero states are the representations supplied to the first decoder block. Positional information may be encoded in these states or enter the attention computation, depending on the architecture. Each attention head uses a linear value projection. The layer update combines the local state and head outputs through position-wise operations, such as residual additions, feed-forward maps, and normalization after attention. The final transformation g and the output projection in Equation 31 map the final hidden state to logits.

Fix an input position $i < n$, a budget $\varepsilon \geq 0$, and a perturbation $\boldsymbol{\delta}^{(i)} \in \mathcal{U}_i(\varepsilon)$. At a generation step t where

$$\mathbf{r}_{<t} = \widetilde{\mathbf{r}}^{(i)}_{<t},$$

the clean and perturbed states $\mathbf{h}_j^{(\ell)} \in \mathbb{R}^d$ and $\widetilde{\mathbf{h}}_j^{(\ell)} \in \mathbb{R}^d$ at position j and layer $\ell \in \{0, \ldots, L\}$ differ by

$$\Delta\mathbf{h}_j^{(\ell)} = \widetilde{\mathbf{h}}_j^{(\ell)} - \mathbf{h}_j^{(\ell)}. \tag{16}$$

The fixed intervention changes one original input position. Since the runs share the generated prefix, the appended tokens have identical layer-zero representations. Thus, the initial difference is confined to position i:

$$\Delta\mathbf{h}_j^{(0)} = \begin{cases} \boldsymbol{\delta}_i^{(i)}, & j = i, \\ \mathbf{0}, & j \neq i. \end{cases} \tag{17}$$

By definition of the perturbation set,

$$\left\| \boldsymbol{\delta}_i^{(i)} \right\|_2 \leq \varepsilon. \tag{18}$$

For head h in layer ℓ, let $s_{j,w,h}^{(\ell)}$ be the attention score between the query at j and the key at w. It depends on their previous-layer states and fixed positional coordinates. For $j \geq 2$, the strict-predecessor attention weights are

$$\alpha_{j,w,h}^{(\ell)} = \frac{\exp(s_{j,w,h}^{(\ell)})}{\sum_{u<j} \exp(s_{j,u,h}^{(\ell)})}, \qquad w < j. \tag{19}$$

The head output is

$$\mathbf{b}_{j,h}^{(\ell)} = \sum_{w<j} \alpha_{j,w,h}^{(\ell)} \mathbf{V}_h^{(\ell)} \mathbf{h}_w^{(\ell-1)}. \tag{20}$$

The corresponding perturbed quantity is

$$\widetilde{\mathbf{b}}_{j,h}^{(\ell)} = \sum_{w<j} \widetilde{\alpha}_{j,w,h}^{(\ell)} \mathbf{V}_h^{(\ell)} \widetilde{\mathbf{h}}_w^{(\ell-1)}. \tag{21}$$

The local state enters through the layer update

$$\mathbf{h}_j^{(\ell)} = f_\ell \left(\mathbf{h}_j^{(\ell-1)}, \mathbf{b}_{j,1}^{(\ell)}, \ldots, \mathbf{b}_{j,H}^{(\ell)} \right), \tag{22}$$

with the same f_ℓ applied to the perturbed states. It combines the local state and attention outputs through the position-wise operations above. At $j = 1$, all attention outputs are zero.

We use the following uniform conditions.

A1. For each layer ℓ, the update f_ℓ is Lipschitz. There exists $L_\ell > 0$ such that

$$\left\| f_\ell\left(\mathbf{u}, \mathbf{b}_1, \ldots, \mathbf{b}_H\right) - f_\ell\left(\widetilde{\mathbf{u}}, \widetilde{\mathbf{b}}_1, \ldots, \widetilde{\mathbf{b}}_H\right) \right\|_2 \leq L_\ell \left(\|\mathbf{u} - \widetilde{\mathbf{u}}\|_2 + \sum_{h=1}^{H} \left\| \mathbf{b}_h - \widetilde{\mathbf{b}}_h \right\|_2 \right).$$

A2. Hidden states and value projections are uniformly bounded. There exist $M, B_V > 0$ such that

$$\left\|\mathbf{h}_j^{(\ell)}\right\|_2, \left\|\widetilde{\mathbf{h}}_j^{(\ell)}\right\|_2 \leq M, \qquad \left\|\mathbf{V}_h^{(\ell)}\right\|_{2\leftarrow 2} \leq B_V.$$

A3. Attention scores are Lipschitz in their query and key states. There exists $L_s > 0$ such that, for every $w < j$,

$$\left|\widetilde{s}_{j,w,h}^{(\ell)} - s_{j,w,h}^{(\ell)}\right| \leq L_s \left(\left\|\Delta\mathbf{h}_j^{(\ell-1)}\right\|_2 + \left\|\Delta\mathbf{h}_w^{(\ell-1)}\right\|_2 \right).$$

A4. Fix a layer, head, and position $j \geq 2$. Let $\mathbf{s}$ and $\widetilde{\mathbf{s}}$ be the clean and perturbed score vectors over $w < j$, and define

$$\boldsymbol{\alpha}_\eta = \mathrm{softmax}\left((1-\eta)\mathbf{s} + \eta\widetilde{\mathbf{s}}\right), \qquad \eta \in [0,1].$$

There exists $c_\alpha \geq 1$ such that, at every layer, head, and sequence position $j \geq 2$,

$$\|\boldsymbol{\alpha}_\eta\|_\infty \leq \frac{c_\alpha}{j-1}. \tag{23}$$

The proof uses the constants L_ℓ, M, B_V, L_s, and c_α specified in A1–A4.

For example, A4 holds if every interpolated score lies in $[-B_s, B_s]$, since

$$\|\boldsymbol{\alpha}_\eta\|_\infty \leq \frac{e^{2B_s}}{j-1}.$$

A.2 Auxiliary Bounds

The propagation proof uses three auxiliary bounds. The first controls changes in the attention weights, the second bounds the hidden-state change at the perturbed position, and the third controls accumulation over intermediate positions in deeper layers.

Lemma A.1 (Softmax stability under diffuse weights)**.** *Let $\boldsymbol{\alpha} = \mathrm{softmax}(\mathbf{s})$ and $\widetilde{\boldsymbol{\alpha}} = \mathrm{softmax}(\widetilde{\mathbf{s}})$ be distributions over m positions. Suppose every softmax distribution along the line segment between $\mathbf{s}$ and $\widetilde{\mathbf{s}}$ satisfies*

$$\|\boldsymbol{\alpha}_\eta\|_\infty \leq \frac{c}{m}.$$

Then

$$\|\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|_1 \leq \frac{2c}{m} \|\widetilde{\mathbf{s}} - \mathbf{s}\|_1. \tag{24}$$

Proof. For $\mathbf{p} = \mathrm{softmax}(\mathbf{s})$, the softmax Jacobian is

$$\mathbf{J}(\mathbf{p}) = \mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top.$$

The sum of the absolute values of the entries in column k is

$$p_k(1-p_k) + \sum_{r\neq k} p_r p_k = 2p_k(1-p_k) \leq 2\|\mathbf{p}\|_\infty.$$

Therefore,

$$\|\mathbf{J}(\mathbf{p})\|_{1\leftarrow 1} \leq \frac{2c}{m}$$

along the entire score segment. Integrating the Jacobian along this segment gives Equation 24. □

Lemma A.2 (Source-position stability). *Under A1–A4, each layer $\ell \in \{0, \dots, L\}$ has a constant $B_\ell > 0$, depending only on the layer and the uniform assumption constants, such that*

$$\left\|\Delta \mathbf{h}_i^{(\ell)}\right\|_2 \le B_\ell \left\|\boldsymbol{\delta}_i^{(i)}\right\|_2 . \tag{25}$$

Also,

$$\Delta \mathbf{h}_j^{(\ell)} = \mathbf{0} \qquad \text{for every } j < i.$$

Proof. By causality, positions $j < i$ remain unchanged at every layer. At layer zero, Equation 25 holds with $B_0 = 1$.

Suppose

$$\left\|\Delta \mathbf{h}_i^{(\ell-1)}\right\|_2 \le B_{\ell-1} \left\|\boldsymbol{\delta}_i^{(i)}\right\|_2 .$$

Causality leaves all positions before i unchanged, so the value vectors entering attention at i are identical in the two runs. The perturbed query can still change the attention weights. A3 bounds the resulting score difference:

$$\left\|\widetilde{\mathbf{s}}_{i,\cdot,h}^{(\ell)} - \mathbf{s}_{i,\cdot,h}^{(\ell)}\right\|_1 \le (i-1) L_s \left\|\Delta \mathbf{h}_i^{(\ell-1)}\right\|_2 .$$

For $i > 1$, Lemma A.1 with $m = i - 1$ yields

$$\left\|\widetilde{\boldsymbol{\alpha}}_{i,\cdot,h}^{(\ell)} - \boldsymbol{\alpha}_{i,\cdot,h}^{(\ell)}\right\|_1 \le 2 c_\alpha L_s \left\|\Delta \mathbf{h}_i^{(\ell-1)}\right\|_2 .$$

Using A2,

$$\left\|\widetilde{\mathbf{b}}_{i,h}^{(\ell)} - \mathbf{b}_{i,h}^{(\ell)}\right\|_2 \le 2 c_\alpha B_V M L_s \left\|\Delta \mathbf{h}_i^{(\ell-1)}\right\|_2 .$$

For $i = 1$, the attention term is zero. Applying A1 gives

$$\left\|\Delta \mathbf{h}_i^{(\ell)}\right\|_2 \le L_\ell \left(1 + 2H c_\alpha B_V M L_s\right) \left\|\Delta \mathbf{h}_i^{(\ell-1)}\right\|_2 .$$

The claim follows recursively with

$$B_\ell = L_\ell \left(1 + 2H c_\alpha B_V M L_s\right) B_{\ell-1}.$$

□

To track accumulation between the perturbed position and a downstream position $j > i$, define

$$R_i(j) = 1 + \log \frac{j-1}{i}. \tag{26}$$

Lemma A.3 (Logarithmic accumulation). *For every integer $r \ge 0$, there exists $c_r > 0$, independent of i and j, such that*

$$\sum_{w=i+1}^{j-1} \frac{R_i(w)^r}{w-1} \le c_r R_i(j)^{r+1}, \qquad j > i+1. \tag{27}$$

Proof. Let $k = w - 1$. The sum becomes

$$\sum_{k=i}^{j-2} \frac{\left(1 + \log \frac{k}{i}\right)^r}{k}.$$

For $x \in [k, k+1]$, we have $x \le 2k$ and

$$1 + \log \frac{x}{i} \ge 1 + \log \frac{k}{i}.$$

Hence

$$\frac{\left(1+\log\frac{k}{i}\right)^r}{k} \le 2\int_k^{k+1} \frac{\left(1+\log\frac{x}{i}\right)^r}{x}\,dx.$$

Summing gives

$$\begin{aligned}\sum_{k=i}^{j-2} \frac{\left(1+\log\frac{k}{i}\right)^r}{k} &\le 2\int_i^{j-1} \frac{\left(1+\log\frac{x}{i}\right)^r}{x}\,dx \\ &= \frac{2}{r+1}\left(R_i(j)^{r+1}-1\right) \\ &\le \frac{2}{r+1}R_i(j)^{r+1}.\end{aligned}$$

Thus, the result holds with $c_r = 2/(r+1)$. □

A.3 Hidden-State Propagation

Theorem A.4 (Local perturbation propagation)**.** *Under A1–A4, each layer $\ell \in \{1,\dots,L\}$ has a constant $C_\ell > 0$ that depends only on the layer and the uniform assumption constants. Let $\varepsilon \ge 0$ and $\boldsymbol{\delta}^{(i)} \in \mathcal{U}_i(\varepsilon)$. At a step t with a common generated prefix, every position $i < j \le n+t$ satisfies*

$$\left\|\Delta\mathbf{h}_j^{(\ell)}\right\|_2 \le C_\ell \left\|\boldsymbol{\delta}_i^{(i)}\right\|_2 \frac{R_i(j)^{\ell-1}}{j-1} \le C_\ell\varepsilon\frac{R_i(j)^{\ell-1}}{j-1}. \tag{28}$$

Proof. We proceed by induction over ℓ. At the first layer, the perturbation enters through a single input position. In deeper layers, the perturbation can also arrive through intermediate positions. We bound these contributions using the preceding lemmas.

Base layer. At layer zero, only position i differs. For $j > i$, decompose the attention-output difference for head h as

$$\begin{aligned}&\widetilde{\mathbf{b}}_{j,h}^{(1)} - \mathbf{b}_{j,h}^{(1)} \\ &= \underbrace{\sum_{w<j} \alpha_{j,w,h}^{(1)} \mathbf{V}_h^{(1)} \Delta\mathbf{h}_w^{(0)}}_{T_{j,h,\mathrm{val}}^{(1)}} \\ &\quad + \underbrace{\sum_{w<j} \left(\widetilde{\alpha}_{j,w,h}^{(1)} - \alpha_{j,w,h}^{(1)}\right) \mathbf{V}_h^{(1)} \widetilde{\mathbf{h}}_w^{(0)}}_{T_{j,h,\mathrm{wt}}^{(1)}}.\end{aligned}$$

A2 and A4 give

$$\left\|T_{j,h,\mathrm{val}}^{(1)}\right\|_2 \le \frac{c_\alpha B_V}{j-1}\left\|\boldsymbol{\delta}_i^{(i)}\right\|_2.$$

For $j > i$, the query is unchanged and only the key at i differs. A3 gives

$$\left\|\widetilde{\mathbf{s}}_{j,\cdot,h}^{(1)} - \mathbf{s}_{j,\cdot,h}^{(1)}\right\|_1 \le L_s\left\|\boldsymbol{\delta}_i^{(i)}\right\|_2.$$

By Lemma A.1 and A2,

$$\left\|T_{j,h,\mathrm{wt}}^{(1)}\right\|_2 \le \frac{2c_\alpha B_V M L_s}{j-1}\left\|\boldsymbol{\delta}_i^{(i)}\right\|_2.$$

The local layer-zero state at j is unchanged. Applying A1 and summing over the H heads gives

$$\left\|\Delta\mathbf{h}_j^{(1)}\right\|_2 \le C_1\frac{\left\|\boldsymbol{\delta}_i^{(i)}\right\|_2}{j-1},$$

which proves Equation 28 for $\ell = 1$.

Inductive step. Assume the result holds at layer $\ell-1$, where $\ell \geq 2$. For $j > i$, use the decomposition

$$\begin{aligned}\widetilde{\mathbf{b}}_{j,h}^{(\ell)} - \mathbf{b}_{j,h}^{(\ell)} &= \underbrace{\sum_{w<j} \alpha_{j,w,h}^{(\ell)} \mathbf{V}_h^{(\ell)} \Delta\mathbf{h}_w^{(\ell-1)}}_{T_{j,h,\mathrm{val}}^{(\ell)}} \\ &\quad + \underbrace{\sum_{w<j} \left(\widetilde{\alpha}_{j,w,h}^{(\ell)} - \alpha_{j,w,h}^{(\ell)}\right) \mathbf{V}_h^{(\ell)} \widetilde{\mathbf{h}}_w^{(\ell-1)}}_{T_{j,h,\mathrm{wt}}^{(\ell)}}.\end{aligned}$$

Positions before i remain unchanged. The sum over predecessors therefore contains the source contribution at i and the contributions from positions between i and j. Lemma A.2 bounds the source term. The induction hypothesis and Lemma A.3 bound the remaining sum, giving

$$\begin{aligned}\sum_{w<j} \left\|\Delta\mathbf{h}_w^{(\ell-1)}\right\|_2 &\leq B_{\ell-1} \left\|\boldsymbol{\delta}_i^{(i)}\right\|_2 \\ &\quad + C_{\ell-1} \left\|\boldsymbol{\delta}_i^{(i)}\right\|_2 \sum_{w=i+1}^{j-1} \frac{R_i(w)^{\ell-2}}{w-1} \\ &\leq C \left\|\boldsymbol{\delta}_i^{(i)}\right\|_2 R_i(j)^{\ell-1},\end{aligned}$$

where C depends only on the uniform constants and the layer. When $j = i+1$, the intermediate-position sum is empty and the same result follows from $R_i(j) = 1$.

Using A2 and A4,

$$\left\|T_{j,h,\mathrm{val}}^{(\ell)}\right\|_2 \leq C \left\|\boldsymbol{\delta}_i^{(i)}\right\|_2 \frac{R_i(j)^{\ell-1}}{j-1}. \tag{29}$$

For the attention-score difference, A3 gives

$$\begin{aligned}&\left\|\widetilde{\mathbf{s}}_{j,\cdot,h}^{(\ell)} - \mathbf{s}_{j,\cdot,h}^{(\ell)}\right\|_1 \\ &\leq L_s \left((j-1)\left\|\Delta\mathbf{h}_j^{(\ell-1)}\right\|_2 + \sum_{w<j} \left\|\Delta\mathbf{h}_w^{(\ell-1)}\right\|_2\right) \\ &\leq C \left\|\boldsymbol{\delta}_i^{(i)}\right\|_2 R_i(j)^{\ell-1}.\end{aligned}$$

Lemma A.1, together with A2 and A4, then yields

$$\left\|T_{j,h,\mathrm{wt}}^{(\ell)}\right\|_2 \leq C \left\|\boldsymbol{\delta}_i^{(i)}\right\|_2 \frac{R_i(j)^{\ell-1}}{j-1}. \tag{30}$$

Finally, A1 combines the local hidden-state difference with the attention output differences. The induction hypothesis gives

$$\left\|\Delta\mathbf{h}_j^{(\ell-1)}\right\|_2 \leq C_{\ell-1} \left\|\boldsymbol{\delta}_i^{(i)}\right\|_2 \frac{R_i(j)^{\ell-2}}{j-1} \leq C_{\ell-1} \left\|\boldsymbol{\delta}_i^{(i)}\right\|_2 \frac{R_i(j)^{\ell-1}}{j-1}.$$

Absorbing the layer- and head-dependent factors into C_ℓ completes the induction. The second inequality in Equation 28 follows from $\|\boldsymbol{\delta}_i^{(i)}\|_2 \leq \varepsilon$. □

Explicit constants. Write $e_j^{(\ell)} = \|\Delta\mathbf{h}_j^{(\ell)}\|_2$. The preceding estimates give

$$e_j^{(\ell)} \leq a_\ell e_j^{(\ell-1)} + \frac{b_\ell}{j-1} \sum_{w<j} e_w^{(\ell-1)}, \qquad j \geq 2,$$

where

$$a_\ell = L_\ell(1 + 2Hc_\alpha B_V M L_s), \qquad b_\ell = L_\ell H c_\alpha B_V (1 + 2ML_s).$$

At $j = 1$, the sum is omitted. Set $B_0 = 1$, $B_\ell = a_\ell B_{\ell-1}$, and $C_1 = b_1$. For $\ell \geq 2$, an explicit choice is

$$C_\ell = a_\ell C_{\ell-1} + b_\ell \left(B_{\ell-1} + \frac{2}{\ell - 1} C_{\ell-1} \right),$$

using $c_{\ell-2} = 2/(\ell-1)$ from Lemma A.3. These constants depend on the uniform bounds in A1–A4.

Attention including the current position. The same position-dependent bound also holds when attention includes the current position. For this variant, use

$$\alpha^{(\ell)}_{j,w,h} = \frac{\exp(s^{(\ell)}_{j,w,h})}{\sum_{u \leq j} \exp(s^{(\ell)}_{j,u,h})}, \qquad w \leq j,$$

and sum the value vectors over $w \leq j$. Assume A3 for every $w \leq j$, and require

$$\|\boldsymbol{\alpha}_\eta\|_\infty \leq \frac{c_\alpha}{j}$$

along each interpolated score segment. A1 and A2 are unchanged. Bounded interpolated scores in $[-B_s, B_s]$ again suffice, with $c_\alpha = e^{2B_s}$.

Write $e^{(\ell)}_j = \|\Delta \mathbf{h}^{(\ell)}_j\|_2$. The value-and-weight decomposition and Lemma A.1 give, for $j \geq 2$,

$$\|\widetilde{\mathbf{b}}^{(\ell)}_{j,h} - \mathbf{b}^{(\ell)}_{j,h}\|_2 \leq 2c_\alpha B_V M L_s e^{(\ell-1)}_j + \frac{c_\alpha B_V (1 + 2ML_s)}{j} \sum_{w \leq j} e^{(\ell-1)}_w.$$

Applying A1 and separating the term at $w = j$ yields

$$\begin{aligned} e^{(\ell)}_j &\leq a_\ell e^{(\ell-1)}_j + \frac{b_\ell}{j} \sum_{w \leq j} e^{(\ell-1)}_w \\ &\leq (a_\ell + b_\ell) e^{(\ell-1)}_j + \frac{b_\ell}{j-1} \sum_{w < j} e^{(\ell-1)}_w. \end{aligned}$$

At $j = 1$, the only attention weight is one, so

$$e^{(\ell)}_1 \leq L_\ell(1 + HB_V) e^{(\ell-1)}_1 \leq (a_\ell + b_\ell) e^{(\ell-1)}_1.$$

Thus, replace a_ℓ by $a_\ell + b_\ell$ in the recursions for B_ℓ and C_ℓ. The source-position bound and the logarithmic accumulation argument then give Equation 28 with these constants. The position factor is unchanged.

A.4 From Hidden-State Propagation to Logit Drift

We now convert the hidden-state bound into a bound on the logits used for generation. At step t, the next token is predicted from position $n + t$, which lies after the perturbed input position. With a common generated prefix, this position has the same layer-zero representation in both runs, so the propagation theorem applies. The output map is

$$\mathbf{z}_t = \mathbf{W}_{\text{out}} g\left(\mathbf{h}^{(L)}_{n+t}\right) + \mathbf{b}_{\text{out}}, \qquad \widetilde{\mathbf{z}}_t = \mathbf{W}_{\text{out}} g\left(\widetilde{\mathbf{h}}^{(L)}_{n+t}\right) + \mathbf{b}_{\text{out}}. \tag{31}$$

Here g is the final transformation, such as normalization or the identity. The logit drift is

$$\Delta \mathbf{z}_t = \widetilde{\mathbf{z}}_t - \mathbf{z}_t. \tag{32}$$

Suppose g is L_g-Lipschitz in the Euclidean norm and define

$$C_{\text{out}} = L_g \max_{v \in \mathcal{V}} \|(\mathbf{W}_{\text{out}})_{v,:}\|_2 . \tag{33}$$

Here $(\mathbf{W}_{\text{out}})_{v,:}$ is the output row for token v. The constant C_{out} bounds how much the output map can amplify a hidden-state change. Combining it with the layerwise propagation bound gives the coefficient

$$A(i, t, L) = C_{\text{out}} C_L \frac{\left(1 + \log \frac{n+t-1}{i}\right)^{L-1}}{n + t - 1}. \tag{34}$$

Corollary A.5 (Local perturbation logit bound)**.** *Under the conditions of Theorem A.4 and the Lipschitz condition on g above, fix $\varepsilon \geq 0$ and $\boldsymbol{\delta}^{(i)} \in \mathcal{U}_i(\varepsilon)$. If the runs share the generated prefix before step t, then*

$$\|\Delta \mathbf{z}_t\|_\infty \leq A(i,t,L) \left\|\boldsymbol{\delta}_i^{(i)}\right\|_2 \leq A(i,t,L)\varepsilon. \tag{35}$$

Proof. Applying Cauchy–Schwarz to each output row and using the Lipschitz property of g gives

$$\begin{aligned}\|\Delta \mathbf{z}_t\|_\infty &\leq L_g \max_{v\in\mathcal{V}} \|(\mathbf{W}_{\text{out}})_{v,:}\|_2 \left\|\Delta \mathbf{h}_{n+t}^{(L)}\right\|_2 \\ &= C_{\text{out}} \left\|\Delta \mathbf{h}_{n+t}^{(L)}\right\|_2 .\end{aligned}$$

Applying Theorem A.4 at $j = n + t$ proves the first inequality. The second follows from $\|\boldsymbol{\delta}_i^{(i)}\|_2 \leq \varepsilon$. □

A.5 Decision-Margin Stability

A logit-drift bound yields an exact token-level guarantee when the drift is sufficiently small relative to the clean token's lead over its competitors. We measure this lead by the clean decision margin

$$m_t = \mathbf{z}_t(r_t) - \max_{v \neq r_t} \mathbf{z}_t(v) \geq 0. \tag{36}$$

The margin is positive for a unique maximizer and zero for a tie.

Lemma A.6 (Greedy next-token stability)**.** *Suppose the clean and perturbed runs share the generated prefix before step t. If*

$$2\|\Delta \mathbf{z}_t\|_\infty < m_t, \tag{37}$$

then

$$\arg\max_{v\in\mathcal{V}} \widetilde{\mathbf{z}}_t(v) = \arg\max_{v\in\mathcal{V}} \mathbf{z}_t(v) = r_t.$$

Proof. For every $v \neq r_t$,

$$\begin{aligned}&\widetilde{\mathbf{z}}_t(r_t) - \widetilde{\mathbf{z}}_t(v) \\ &= \mathbf{z}_t(r_t) - \mathbf{z}_t(v) + \Delta \mathbf{z}_t(r_t) - \Delta \mathbf{z}_t(v) \\ &\geq m_t - 2\|\Delta \mathbf{z}_t\|_\infty > 0.\end{aligned}$$

Thus, r_t remains the unique maximizer. □

Corollary A.5 and Lemma A.6 yield the following certificate.

Corollary A.7 (Uniform one-step stability)**.** *Fix $\varepsilon \geq 0$. If*

$$2A(i,t,L)\varepsilon < m_t, \tag{38}$$

then every $\boldsymbol{\delta}^{(i)} \in \mathcal{U}_i(\varepsilon)$ whose trajectory shares the clean prefix before step t selects the clean token r_t at that step.

A.6 Exact Prefix Robustness

The one-step certificate applies whenever the two runs share a generated prefix. Starting from the empty prefix, each certified token preserves this agreement and allows the certificate to be applied at the next step. This induction gives a prefix guarantee for every perturbation in $\mathcal{U}_i(\varepsilon)$.

Theorem A.8 (Prefix stability certificate)**.** *Fix a clean input $\mathbf{x}$, a position $i \in \{1, \ldots, n-1\}$, and a budget $\varepsilon \geq 0$. Let $K \in \mathbb{N}_+$ be within the clean generation horizon. Assume the uniform propagation conditions and the Lipschitz condition on g hold at every common-prefix step $0 \leq t < K$. If*

$$2A(i,t,L)\varepsilon < m_t \qquad \text{for every } t \in \{0, \ldots, K-1\}, \tag{39}$$

then $\mathbf{x}$ is (i, ε, K)-robust:

$$\widetilde{r}_t^{(i)} = r_t \qquad \text{for every } \boldsymbol{\delta}^{(i)} \in \mathcal{U}_i(\varepsilon) \text{ and every } t \in \{0, \ldots, K-1\}. \tag{40}$$

Equivalently,

$$\tau(\boldsymbol{\delta}^{(i)}) \geq K \qquad \text{for every } \boldsymbol{\delta}^{(i)} \in \mathcal{U}_i(\varepsilon). \tag{41}$$

Proof. Fix an arbitrary $\boldsymbol{\delta}^{(i)} \in \mathcal{U}_i(\varepsilon)$. At $t = 0$, both runs condition on the empty generated prefix. Equation 39 and Corollary A.7 imply

$$\widetilde{r}_0^{(i)} = r_0.$$

Suppose the first t generated tokens agree. Applying Corollary A.7 to this common prefix gives

$$\widetilde{r}_t^{(i)} = r_t.$$

Induction proves equality for all $t \in \{0, \dots, K-1\}$ and all $\boldsymbol{\delta}^{(i)} \in \mathcal{U}_i(\varepsilon)$, including matching termination decisions within the target horizon. □

For a fixed clean trajectory and target prefix length K, this certificate requires the margin condition at every step. Taking the smallest stepwise threshold gives the certified radius

$$\boxed{\rho_i^{\mathrm{cert}}(K) = \min_{0 \le t < K} \frac{m_t}{2A(i,t,L)}.} \tag{42}$$

Corollary A.9 (Radius-based robustness)**.** *Under the assumptions of Theorem A.8, if*

$$\varepsilon < \rho_i^{\mathrm{cert}}(K), \tag{43}$$

then $\mathbf{x}$ *is* (i, ε, K)*-robust.*

For a target prefix length K, a token substitution with representation distance D_i is certified whenever

$$D_i < \rho_i^{\mathrm{cert}}(K).$$

Alternatively, fix the budget and let $K_{\max}$ be the maximum prefix length considered within the fixed analysis horizon. The longest prefix satisfying the radius condition is

$$K_i^{\mathrm{cert}}(\varepsilon) = \max\left(\{0\} \cup \left\{K \in \{1, \dots, K_{\max}\} : \varepsilon < \rho_i^{\mathrm{cert}}(K)\right\}\right). \tag{44}$$

Every $\boldsymbol{\delta}^{(i)} \in \mathcal{U}_i(\varepsilon)$ then satisfies

$$\tau(\boldsymbol{\delta}^{(i)}) \ge K_i^{\mathrm{cert}}(\varepsilon). \tag{45}$$

A.7 Closed-Form Analysis of the Prefix Certificates

We derive the explicit forms used in Section 3.3. Recall

$$a_K = \min_{0 \le t < K} m_t, \tag{46}$$

and

$$f_i(t) = \frac{\left(1 + \log \frac{n+t-1}{i}\right)^{L-1}}{n+t-1}. \tag{47}$$

Since

$$A(i,t,L) = C_{\mathrm{out}} C_L f_i(t),$$

and $m_t \ge a_K$, the condition

$$2C_{\mathrm{out}} C_L \varepsilon f_i(t) < a_K \qquad \text{for every } 0 \le t < K \tag{48}$$

implies the original stepwise condition $2A(i,t,L)\varepsilon < m_t$.

Dependence on the generation step. Treat t temporarily as a continuous variable. Differentiating $f_i(t)$ gives

$$f_i'(t) = \frac{\left(1 + \log \frac{n+t-1}{i}\right)^{L-2}}{(n+t-1)^2} \left(L - 2 - \log \frac{n+t-1}{i}\right). \tag{49}$$

For $L \ge 2$, the stationary point satisfies

$$\log \frac{n+t-1}{i} = L - 2,$$

and hence

$$t_i^\star = ie^{L-2} - n + 1. \tag{50}$$

It follows that $f_i(t)$ is decreasing for $t \geq 0$ when $t_i^\star \leq 0$. When $t_i^\star > 0$, it increases before $t_i^\star$ and decreases afterward. Restricted to the discrete steps $\{0, \ldots, K-1\}$, its maximum is therefore

$$\max_{t\in\{0,\ldots,K-1\}} f_i(t) = \begin{cases} f_i(0), & t_i^\star \leq 0, \\ \max\left\{f_i(\lfloor t_i^\star \rfloor), f_i(\lceil t_i^\star \rceil)\right\}, & 0 < t_i^\star < K-1, \\ f_i(K-1), & t_i^\star \geq K-1. \end{cases} \tag{51}$$

At an interior continuous maximum,

$$f_i(t_i^\star) = \frac{(L-1)^{L-1}}{ie^{L-2}}. \tag{52}$$

Certified prefix length. Consider the full analysis horizon $K_{\max}$. If

$$2C_{\text{out}}C_L\varepsilon \max_{t\in\{0,\ldots,K_{\max}-1\}} f_i(t) < a_{K_{\max}}, \tag{53}$$

then Equation 48 holds throughout the entire horizon, and therefore

$$K_i^{\text{cert}}(\varepsilon) = K_{\max}.$$

If instead

$$2C_{\text{out}}C_L\varepsilon f_i(0) \geq a_{K_{\max}}, \tag{54}$$

the uniform-margin condition does not certify even the first generated token, giving the lower bound $K_i^{\text{cert}}(\varepsilon) \geq 0$.

In the remaining case, the continuous curve starts below the threshold but reaches it before or at its maximum. Let $t_i^\dagger$ be its first crossing:

$$2C_{\text{out}}C_L\varepsilon f_i(t_i^\dagger) = a_{K_{\max}}. \tag{55}$$

Every integer generation step strictly before $\lceil t_i^\dagger \rceil$ lies before this first crossing and therefore satisfies Equation 48. Hence

$$K_i^{\text{cert}}(\varepsilon) \geq \left\lceil t_i^\dagger \right\rceil. \tag{56}$$

Combining the three cases proves Equation 12.

The crossing point also admits an explicit expression. Let

$$b = \frac{a_{K_{\max}}}{2C_{\text{out}}C_L\varepsilon}. \tag{57}$$

Solving

$$f_i(t) = b$$

with $x = n + t - 1$ gives

$$\frac{\left(1 + \log\frac{x}{i}\right)^{L-1}}{x} = b.$$

Therefore

$$\left(1 + \log\frac{x}{i}\right)\exp\left[-\frac{1 + \log(x/i)}{L-1}\right] = \left(\frac{bi}{e}\right)^{\frac{1}{L-1}}.$$

The first crossing lies on the increasing branch and is obtained from the principal Lambert-W branch:

$$\boxed{t_i^\dagger = i\exp\left[-(L-1)W_0\left(-\frac{1}{L-1}\left(\frac{bi}{e}\right)^{\frac{1}{L-1}}\right) - 1\right] - n + 1.} \tag{58}$$

When a second root exists on the decreasing branch, it corresponds to W_{-1} and does not determine the first point at which the uniform-margin sufficient condition fails.

Certified perturbation radius. For fixed i and K,

$$\rho_i^{\mathrm{cert}}(K) = \min_{0 \le t < K} \frac{m_t}{2C_{\mathrm{out}} C_L f_i(t)}.$$

Since

$$m_t \ge a_K$$

and

$$f_i(t) \le \max_{s \in \{0,\dots,K-1\}} f_i(s),$$

we have, for every $t < K$,

$$\frac{m_t}{2C_{\mathrm{out}} C_L f_i(t)} \ge \frac{a_K}{2C_{\mathrm{out}} C_L \max_{s \in \{0,\dots,K-1\}} f_i(s)}.$$

Taking the minimum over t proves

$$\rho_i^{\mathrm{cert}}(K) \ge \frac{a_K}{2C_{\mathrm{out}} C_L \max\limits_{t \in \{0,\dots,K-1\}} f_i(t)}, \tag{59}$$

which gives Proposition 3.1.

Certified input positions. For fixed $\varepsilon > 0$, K, and generation step t, the uniform-margin condition is

$$2C_{\mathrm{out}} C_L \varepsilon \frac{\left(1 + \log \frac{n+t-1}{i}\right)^{L-1}}{n+t-1} < a_K. \tag{60}$$

For $L \ge 2$, taking the $(L-1)$-th root gives

$$1 + \log \frac{n+t-1}{i} < \left(\frac{a_K (n+t-1)}{2C_{\mathrm{out}} C_L \varepsilon} \right)^{\frac{1}{L-1}}.$$

Rearranging yields

$$i > (n+t-1) \exp\left[1 - \left(\frac{a_K (n+t-1)}{2C_{\mathrm{out}} C_L \varepsilon} \right)^{\frac{1}{L-1}} \right]. \tag{61}$$

For the first K tokens to be certified, this condition must hold for every $0 \le t < K$. Taking the maximum of the right-hand side therefore gives

$$i > \max_{t \in \{0,\dots,K-1\}} \left\{ (n+t-1) \exp\left[1 - \left(\frac{a_K (n+t-1)}{2C_{\mathrm{out}} C_L \varepsilon} \right)^{\frac{1}{L-1}} \right] \right\}, \tag{62}$$

which proves Proposition 3.2.

For $a_K > 0$, the expression inside the maximum is unimodal as a function of $x = n + t - 1$. Its stationary point satisfies

$$x = \frac{2C_{\mathrm{out}} C_L \varepsilon (L-1)^{L-1}}{a_K}. \tag{63}$$

Hence its discrete maximum can again be found from the relevant boundary or the nearest integers around this stationary point. If $a_K = 0$, the expression inside the maximum reduces to $e(n+t-1)$, and the uniform-margin sufficient condition certifies no input position.

A.8 Simultaneous Input Perturbations

Corollary A.10 (Simultaneous input perturbations)**.** *Let $S \subseteq \{1, \dots, n-1\}$ be nonempty, and fix a budget $\varepsilon_i \ge 0$ for each $i \in S$. Consider a joint perturbation $\boldsymbol{\delta}^{(S)}$ supported on S, satisfying*

$$\boldsymbol{\delta}_j^{(S)} = \mathbf{0} \quad \text{for } j \notin S, \qquad \left\| \boldsymbol{\delta}_i^{(S)} \right\|_2 \le \varepsilon_i \quad \text{for } i \in S.$$

Assume A1–A4 and the Lipschitz bound for g hold uniformly on this joint perturbation domain at every common-prefix step in the analysis horizon. Use the corresponding constants in C_ℓ and $A(i,t,L)$. If $\widetilde{\mathbf{r}}_{<t}^{(S)} = \mathbf{r}_{<t}$, then for $\ell \in \{1,\dots,L\}$ and $\max S < j \le n+t$,

$$\left\|\Delta \mathbf{h}_j^{(\ell)}\right\|_2 \le \frac{C_\ell}{j-1} \sum_{i\in S} R_i(j)^{\ell-1} \left\|\boldsymbol{\delta}_i^{(S)}\right\|_2 . \tag{64}$$

In particular, the next-token logit drift satisfies

$$\left\|\Delta \mathbf{z}_t\right\|_\infty \le \sum_{i\in S} A(i,t,L) \left\|\boldsymbol{\delta}_i^{(S)}\right\|_2 . \tag{65}$$

In this bound, each source contributes its perturbation magnitude weighted by its propagation coefficient. The sum follows from a nonnegative recurrence for the error bounds, as shown below.

The same prefix-induction argument applies to simultaneous changes. Assume the conditions of Corollary A.10 hold throughout the target horizon. Every joint perturbation satisfying the per-position budgets preserves the first K tokens if

$$2\sum_{i\in S} A(i,t,L)\varepsilon_i < m_t \qquad \text{for every } 0 \le t < K. \tag{66}$$

Proof of Corollary A.10. Fix a common-prefix step and write

$$e_j^{(\ell)} = \left\|\Delta \mathbf{h}_j^{(\ell)}\right\|_2 .$$

At layer zero,

$$e_j^{(0)} = \left\|\boldsymbol{\delta}_j^{(S)}\right\|_2$$

for $j \in S$, and it is zero elsewhere, including all generated positions. All constants below are uniform on the joint perturbation domain specified in the corollary.

For $j \ge 2$, split the attention difference into value and weight terms as in the single-position proof. A2–A4 and Lemma A.1 give

$$\left\|\widetilde{\mathbf{b}}_{j,h}^{(\ell)} - \mathbf{b}_{j,h}^{(\ell)}\right\|_2 \le 2c_\alpha B_V M L_s e_j^{(\ell-1)} + \frac{c_\alpha B_V (1+2ML_s)}{j-1} \sum_{w<j} e_w^{(\ell-1)}.$$

A1 then gives

$$e_j^{(\ell)} \le a_\ell e_j^{(\ell-1)} + \frac{b_\ell}{j-1} \sum_{w<j} e_w^{(\ell-1)}, \tag{67}$$

where a_ℓ and b_ℓ are defined in Appendix A.3, using uniform constants on the joint perturbation domain. At $j = 1$, the sum term is omitted and the same local bound holds.

To separate the contributions of the perturbed positions, let

$$p_{j,i}^{(0)} = \mathbf{1}_{\{j=i\}}$$

for each source i, and define $p_{j,i}^{(\ell)}$ by equality in Equation 67, with the same convention at $j = 1$. This is the response of the comparison recurrence to a unit initial error at i. Its linearity lets us add the source contributions, and its nonnegative coefficients preserve the upper bound. Induction therefore gives

$$e_j^{(\ell)} \le \sum_{i\in S} p_{j,i}^{(\ell)} \left\|\boldsymbol{\delta}_i^{(S)}\right\|_2 .$$

Here $p_{j,i}^{(\ell)} = 0$ for $j < i$, and $p_{i,i}^{(\ell)} = B_\ell$, where $B_0 = 1$ and $B_\ell = a_\ell B_{\ell-1}$. For $j > i$, the base layer satisfies

$$p_{j,i}^{(1)} = \frac{b_1}{j-1}.$$

Suppose the downstream bound holds at layer $\ell - 1$. For $\ell \geq 2$, Lemma A.3 and $R_i(j) \geq 1$ give

$$p_{j,i}^{(\ell)} \leq \left[a_\ell C_{\ell-1} + b_\ell \left(B_{\ell-1} + c_{\ell-2} C_{\ell-1}\right)\right] \frac{R_i(j)^{\ell-1}}{j-1}.$$

The bound also holds at $j = i + 1$, where the intermediate-position sum is empty. The bracketed coefficient is C_ℓ from the recursion in Appendix A.3. Substitution proves Equation 64.

The output-map inequality

$$\|\Delta \mathbf{z}_t\|_\infty \leq C_{\text{out}} e_{n+t}^{(L)},$$

with $n + t > \max S$, then proves Equation 65.

Given a common generated prefix, Equation 66 gives, for every $v \neq r_t$,

$$\begin{aligned} \widetilde{\mathbf{z}}_t(r_t) - \widetilde{\mathbf{z}}_t(v) &\geq m_t - 2\,\|\Delta \mathbf{z}_t\|_\infty \\ &\geq m_t - 2 \sum_{i \in S} A(i, t, L)\varepsilon_i > 0. \end{aligned}$$

Induction from the empty prefix proves the guarantee for every joint perturbation within the per-position budgets. □

For attention that includes the current position, use the corresponding conditions in Appendix A.3 and replace a_ℓ by $a_\ell + b_\ell$ throughout the comparison recurrence, including the recursions for B_ℓ and C_ℓ. The same nonnegative comparison and margin argument give Equations 64–65 and the prefix condition in Equation 66.

A.9 Interpretation of the Bound

The certificate depends jointly on the propagation coefficient $A(i, t, L)$, the perturbation budget ε, and the clean decision margin m_t. For fixed t, uniform model constants, and $L \geq 2$, $A(i, t, L)$ is nonincreasing in i, giving later input positions a more favorable sufficient condition for prefix preservation. Actual certification also depends on the perturbation magnitude and the clean decision margins. For a fixed clean input and model, the margin at each generation step is independent of the perturbation position.

The certificate is sufficient: failure of the condition does not imply that the reasoning prefix must diverge. Exact prefix preservation provides a token-level process criterion for reasoning robustness by certifying an invariant initial segment of the clean greedy trajectory under all admissible perturbations. The large-model experiments in Section 4 examine the position- and magnitude-dependent structural patterns associated with this certificate, while Appendix C instantiates the formal certified radii end to end.

A two-layer example. Consider two layers that uniformly average over earlier positions, with identity value projections and no additional transformations. Let the attention output at the first position be zero. A perturbation at position i gives, for $j > i + 1$,

$$\Delta \mathbf{h}_j^{(1)} = \frac{\boldsymbol{\delta}_i^{(i)}}{j-1}, \qquad \Delta \mathbf{h}_j^{(2)} = \frac{\boldsymbol{\delta}_i^{(i)}}{j-1} \sum_{w=i+1}^{j-1} \frac{1}{w-1}.$$

The second layer accumulates the perturbation through intermediate positions. The resulting harmonic sum illustrates the logarithmic position dependence captured by the general propagation bound.

Relation to closest theoretical analyses. Prior analyses study single-token influence, causal-path sensitivity, next-token continuity, or perturbation dynamics in simplified or continuous settings (Hahn, 2020; Barbero et al., 2024; Pasten et al., 2025; Wang et al., 2026). Our analysis combines finite local representation perturbations, input-dependent softmax changes, explicit source-position dependence, and an exact-prefix guarantee for greedy autoregressive generation.

B Additional Experimental Details

B.1 Experimental Protocol and Position Mapping

Implementation environment. Large-model experiments were run on a workstation with two NVIDIA RTX PRO 6000 Blackwell Workstation Edition GPUs, using Python 3.11.15 and one GPU per inference process. Complete substitutions and the original magnitude sweeps used vLLM 0.23.0, PyTorch 2.11.0, and Transformers 5.12.1, with tensor parallelism set to one. Qwen3-8B, DeepSeek-R1-Distill-Qwen-14B, and GLM-Z1-9B-0414 used BF16 with FlashAttention 2, whereas GPT-OSS-20B used BF16 computation with MXFP4-quantized weights and the Triton attention backend. The matched-magnitude Qwen3-8B experiment and its corresponding control used FP32 with Transformers 5.12.1, PyTorch 2.11.0+cu130, and SDPA, with TF32 disabled. Decoding was greedy throughout: vLLM used temperature zero and top-p one, while Transformers used `do_sample=False` and one beam; the vLLM sampling seed and the Transformers PyTorch seed were set to 11457.

Cohorts, Positions, and Decoding. Each of the twelve model–benchmark settings uses a fixed cohort of 100 examples. Cohort membership is held fixed across all input positions and perturbation magnitudes within a setting.

We select five locations spanning the tokenized user question at relative sampling coordinates

$$p \in \{0.1, 0.3, 0.5, 0.7, 0.9\}.$$

Here p is used only to select comparable locations within user questions of different lengths; the theoretical perturbation position is the actual position i in the complete formatted model input.

Let $n_{\mathrm{usr}}(\mathbf{x})$ denote the tokenized user-question length. For each sampling coordinate p, the protocol selects an eligible zero-based question-token index $j_p^{\mathrm{usr}}(\mathbf{x})$ satisfying

$$\left| \frac{j_p^{\mathrm{usr}}(\mathbf{x})}{n_{\mathrm{usr}}(\mathbf{x}) - 1} - p \right| \leq 0.05.$$

After task formatting and tokenization, we locate the selected token in the complete model input, verify that the clean and substituted inputs differ at exactly one token position, and record its zero-based index as $j_p^{\mathrm{prompt}}(\mathbf{x})$. The corresponding theoretical position is

$$i_p(\mathbf{x}) = j_p^{\mathrm{prompt}}(\mathbf{x}) + 1.$$

Thus, all theoretical quantities involving i use $i = i_p(\mathbf{x})$, while p serves only as a relative sampling coordinate. The theoretical input length n counts the complete formatted model input. All reported trajectories use deterministic greedy decoding.

For compactness, the tables abbreviate DeepSeek-R1-Distill-Qwen-14B as DeepSeek-14B and GLM-Z1-9B-0414 as GLM-Z1-9B.

Complete Token Substitutions. For a selected position i, the complete token substitution induces the single-position layer-zero perturbation $\boldsymbol{\delta}^{(i)}$ defined in Equation 2, with magnitude

$$D_i = \left\| \boldsymbol{\delta}_i^{(i)} \right\|_2.$$

A strict-valid substitution changes exactly one token after retokenization, preserves the tokenized input length, and leaves all other input tokens unchanged.

The primary draw selects one strict-valid replacement for every example and sampled location. In the cross-draw experiment, we keep the example, selected position, and original target token fixed, select a different strict-valid replacement, and regenerate the reasoning trajectory. Each draw therefore contains

$$12 \times 100 \times 5 = 6{,}000$$

clean–perturbed trajectory comparisons. Complete-substitution trajectories are generated with a maximum of $16{,}384$ new tokens. All comparisons in both draws exhibit a finite first-divergence index before the generation limit, so the reported medians reflect observed divergences rather than horizon censoring.

Layer-Zero Magnitude Paths.

For every primary substitution, we scale its replacement-induced direction while keeping the perturbed position fixed. Given its complete-substitution magnitude D_i, we evaluate

$$\frac{\varepsilon}{D_i} \in \{0.25, 0.5, 0.75, 1\}.$$

At magnitude ε, the applied perturbation is

$$\boldsymbol{\delta}^{(i)}(\varepsilon) = \frac{\varepsilon}{D_i}\boldsymbol{\delta}^{(i)}, \qquad \left\|\boldsymbol{\delta}_i^{(i)}(\varepsilon)\right\|_2 = \varepsilon. \tag{68}$$

All four interventions therefore follow the same replacement-induced direction and differ only in magnitude. The endpoint $\varepsilon = D_i$ recovers the complete-substitution representation, while smaller values interpolate toward the clean representation.

The intervened trajectory is generated from

$$\mathbf{H}^{(0)}(\mathbf{x}) + \boldsymbol{\delta}^{(i)}(\varepsilon),$$

with all other layer-zero input positions unchanged. This construction provides the empirical magnitude path used to examine the representation-space perturbation budget in the theoretical formulation.

For each model–benchmark setting, the magnitude sweep contains 100 clean generations and 2,000 intervened generations, obtained from 100 examples, five sampled positions, and four nonzero magnitudes. Across all twelve settings, this yields 25,200 generation records and 24,000 clean–intervention comparisons.

Generation uses deterministic greedy decoding with a 128-token observation horizon, and prefix retention is evaluated for

$$K \in \{16, 32, 64, 128\}.$$

When no divergence occurs within the observation horizon, the experiment establishes $\tau \geq 128$; therefore every event $\tau \geq K$ evaluated in this sweep is directly determined.

B.2 Aggregation Details

For the complete-substitution experiment, we report the sample median of the first-divergence index τ at each tested input location. Confidence intervals in Figure 2 are obtained from 10,000 sample-level percentile-bootstrap replicates, applied separately to each model–benchmark–position combination and perturbation draw.

For the magnitude sweep, prefix retention is the empirical frequency of the event

$$\tau \geq K.$$

Figure 3(a) first averages the five sampled input positions within each model–benchmark setting and then weights the twelve settings equally. Its confidence intervals use 5,000 stratified bootstrap replicates. Within each replicate, the 100 examples are resampled independently in each setting, while all observations associated with the same example across the five positions, four perturbation magnitudes, and four target prefix lengths remain in the same resampling block. This preserves the paired structure of the magnitude and target-length comparisons.

Figure 3(b) reports, for each setting and target prefix length, the change in pooled retention from $\varepsilon = 0.25D_i$ to $\varepsilon = D_i$. Negative values indicate lower retention at the complete-substitution magnitude, and diamonds denote the median change across the twelve settings.

For the matched-magnitude position experiment in Appendix B.4, confidence intervals for the late-minus-early retention contrast use paired example-level bootstrap resampling, preserving all five positional observations from an example within the same resampling block.

B.3 Position–Prefix: Complete-Substitution Results

Table 2 reports the exact medians underlying Figure 2. For both replacement draws, the median preserved-prefix length increases strictly across all five sampled locations in every one of the twelve model–benchmark settings. Across the 60 corresponding profile points, the primary and cross-draw medians have Pearson correlation $r = 0.980$, showing that the positional pattern is highly consistent across separately selected substitutions.

Table 2: **Complete-substitution preserved-prefix profiles.** Median preserved-prefix length at the five sampled question locations under the primary and cross-draw substitutions. Both draws produce a strictly increasing profile in every model–benchmark setting.

		Primary draw					Cross-draw				
Model	Benchmark	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
Qwen3-8B	ARC-Challenge	18.0	47.0	72.0	112.5	129.0	16.0	35.5	63.5	102.0	129.0
Qwen3-8B	GSM8K	15.0	38.5	60.0	73.5	99.0	13.5	34.0	46.5	59.5	87.0
Qwen3-8B	MMLU-Pro	22.5	73.0	89.0	108.0	130.0	21.0	56.0	77.0	91.5	117.5
DeepSeek-14B	ARC-Challenge	20.0	58.0	80.5	108.0	121.0	15.0	48.0	77.0	90.5	106.0
DeepSeek-14B	GSM8K	19.5	58.0	73.0	93.0	103.0	18.0	50.0	67.0	79.0	90.0
DeepSeek-14B	MMLU-Pro	37.0	72.0	106.5	118.5	119.5	25.0	60.5	72.0	92.0	95.0
GLM-Z1-9B	ARC-Challenge	15.0	37.5	57.5	77.0	102.5	14.0	34.5	49.0	74.0	87.0
GLM-Z1-9B	GSM8K	12.0	27.0	53.0	69.0	84.0	11.0	26.0	45.0	61.0	69.0
GLM-Z1-9B	MMLU-Pro	19.0	62.5	77.0	91.0	107.5	18.5	57.5	84.5	94.5	103.0
GPT-OSS-20B	ARC-Challenge	11.0	29.5	38.5	53.5	66.0	7.5	26.0	38.0	54.0	65.0
GPT-OSS-20B	GSM8K	15.0	24.0	33.5	48.0	61.5	15.0	22.5	31.0	44.0	50.0
GPT-OSS-20B	MMLU-Pro	13.5	37.5	46.0	50.0	59.5	12.5	31.5	37.0	48.0	49.5

B.4 Position–Prefix: Matched-Magnitude Validation

We further evaluate the position–prefix relationship under matched representation-space perturbation magnitudes. Using Qwen3-8B on ARC-Challenge, GSM8K, and MMLU-Pro, we retain the same sampled input locations and replacement-induced directions as in the main experiments, while assigning the same absolute L_2 perturbation magnitude to all five positions within each example.

For an example $\mathbf{x}$ and sampled coordinate p, let $i = i_p(\mathbf{x})$ denote the corresponding theoretical input position. We write $\boldsymbol{\delta}_{\mathbf{x}}^{(i)}$ for the complete-substitution perturbation at this position and define its magnitude as

$$D_{\mathbf{x},i} = \left\| \left(\boldsymbol{\delta}_{\mathbf{x}}^{(i)} \right)_i \right\|_2 . \tag{69}$$

The smallest complete-substitution magnitude among the five sampled positions is

$$D_{\min}(\mathbf{x}) = \min_p D_{\mathbf{x},i_p(\mathbf{x})}. \tag{70}$$

For each example, we evaluate four matched absolute perturbation magnitudes

$$\varepsilon_{\mathbf{x}} \in \{0.25 D_{\min}(\mathbf{x}), 0.50 D_{\min}(\mathbf{x}), 0.75 D_{\min}(\mathbf{x}), D_{\min}(\mathbf{x})\} .$$

At input position i, the matched perturbation is

$$\boldsymbol{\delta}_{\mathbf{x},\text{match}}^{(i)}(\varepsilon_{\mathbf{x}}) = \frac{\varepsilon_{\mathbf{x}}}{D_{\mathbf{x},i}} \boldsymbol{\delta}_{\mathbf{x}}^{(i)}. \tag{71}$$

Consequently,

$$\left\| \left(\boldsymbol{\delta}_{\mathbf{x},\text{match}}^{(i)}(\varepsilon_{\mathbf{x}}) \right)_i \right\|_2 = \varepsilon_{\mathbf{x}},$$

so all five tested positions receive the same absolute L_2 perturbation magnitude for a fixed example and matched-magnitude level. Since $\varepsilon_{\mathbf{x}} \leq D_{\mathbf{x},i}$, every intervention remains on the interpolation path between the clean representation and its complete-substitution endpoint.

The matched-magnitude experiment contains

$$3 \times 100 \times 5 \times 4 = 6{,}000$$

comparisons. The realized perturbation norms agree across the five positions up to numerical precision, confirming the intended magnitude matching. All generations use deterministic greedy decoding and the same 128-token observation horizon as the magnitude sweep.

Table 3: **Late-minus-early prefix-retention gap under matched perturbation magnitudes.** Values report the equally weighted mean across ARC-Challenge, GSM8K, and MMLU-Pro at $K = 64$.

$\varepsilon_{\mathbf{x}}/D_{\min}(\mathbf{x})$	Late-minus-early gap (pp)	95% CI
0.25	5.7	[2.3, 9.0]
0.50	21.3	[16.0, 26.7]
0.75	49.3	[42.7, 56.0]
1.00	60.3	[54.7, 66.0]

Position dependence under matched perturbation magnitudes. Figure 5 reports the benchmark-level prefix-retention profiles at $K = 64$. Across the three benchmarks and four matched-magnitude levels, prefix retention generally increases toward later sampled question locations, with some local decreases. Across all $3 \times 4 \times 4 = 48$ benchmark–magnitude–target-length configurations with $K \in \{16, 32, 64, 128\}$, the late-minus-early retention contrast is positive in 47 cases and zero in one.

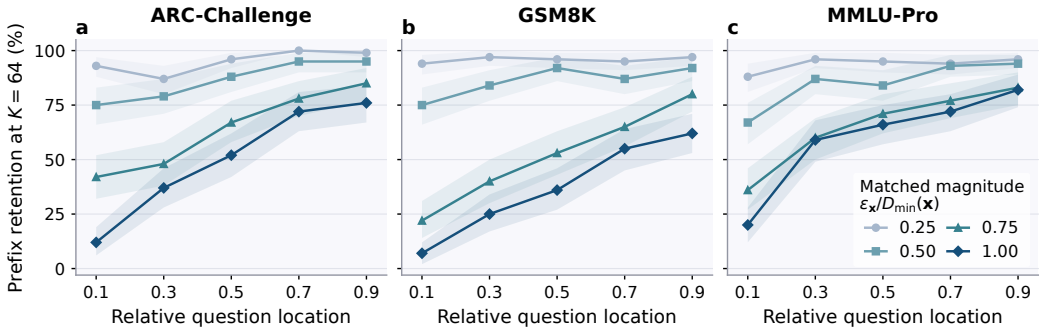

Figure 5: **Benchmark-level prefix-retention profiles under matched perturbation magnitudes.** Results are shown at $K = 64$ for Qwen3-8B on ARC-Challenge, GSM8K, and MMLU-Pro. Within each example and magnitude level, all five sampled input locations receive the same absolute L_2 perturbation magnitude. Shaded bands indicate 95% confidence intervals.

The pooled late-minus-early retention gaps at $K = 64$ are summarized in Table 3.

Figure 6 further reports the median preserved-prefix profiles at the largest matched perturbation magnitude. All three benchmarks show a monotone increase from earlier to later input positions.

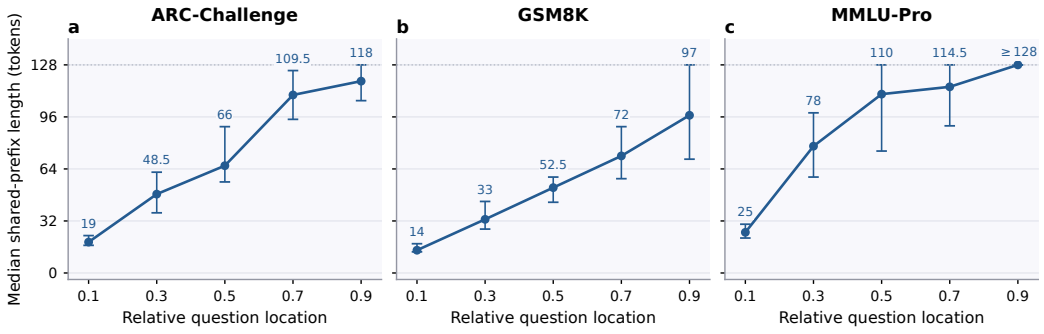

Figure 6: **Preserved-prefix profiles at the largest matched perturbation magnitude.** Median preserved-prefix length is shown at $\varepsilon_{\mathbf{x}} = D_{\min}(\mathbf{x})$. All three benchmarks exhibit a monotone increase from earlier to later input positions. Prefixes that remain identical throughout the 128-token observation horizon are reported as at least 128 tokens.

We additionally compare the matched-magnitude intervention with the corresponding original scaling using the same FP32 inference backend. At $K = 64$, the pooled late-minus-early gaps remain similar after absolute perturbation magnitudes are equalized, as shown in Table 4.

Table 4: **Position gap before and after perturbation-magnitude matching.** Both conditions use the same FP32 inference backend. Values are equally weighted across the three benchmarks at $K = 64$.

Relative perturbation level	Original scaling (pp)	Matched magnitude (pp)
0.25	5.7	5.7
0.50	24.3	21.3
0.75	56.0	49.3
1.00	59.3	60.3

The matched-magnitude results reproduce the positional stability pattern under equalized absolute perturbation magnitudes across the five tested locations.

B.5 Magnitude–Prefix Results

Tables 5 and 6 report prefix-retention rates for every model–benchmark setting after pooling the five sampled input locations. Each entry aggregates 500 comparisons at a fixed relative perturbation magnitude ε/D_i and target prefix length K, and gives the setting-level values underlying Figure 3.

Table 5: **Setting-level prefix-retention rates (%) at lower perturbation magnitudes.** Results are pooled over the five sampled input locations.

		$\varepsilon/D_i = 0.25$				$\varepsilon/D_i = 0.50$			
Model	Benchmark	$K = 16$	$K = 32$	$K = 64$	$K = 128$	$K = 16$	$K = 32$	$K = 64$	$K = 128$
Qwen3-8B	ARC-Challenge	97.6	93.4	81.0	53.0	95.8	90.2	75.6	46.4
Qwen3-8B	GSM8K	96.2	92.0	84.8	61.4	94.0	86.0	76.2	50.6
Qwen3-8B	MMLU-Pro	93.6	84.0	72.8	48.8	95.2	83.8	69.4	40.8
DeepSeek-14B	ARC-Challenge	96.0	87.2	75.0	43.2	94.8	87.4	72.0	40.4
DeepSeek-14B	GSM8K	99.0	95.8	90.4	65.2	98.0	96.0	87.6	59.4
DeepSeek-14B	MMLU-Pro	95.6	87.2	70.4	45.4	94.6	86.4	67.6	43.4
GLM-Z1-9B	ARC-Challenge	96.0	91.0	77.2	42.2	92.8	82.8	62.8	29.0
GLM-Z1-9B	GSM8K	98.4	90.6	80.4	64.4	94.4	82.0	64.0	44.0
GLM-Z1-9B	MMLU-Pro	95.8	89.2	76.6	46.6	93.4	83.0	67.6	38.8
GPT-OSS-20B	ARC-Challenge	91.2	90.8	74.8	35.6	87.8	83.0	62.0	22.2
GPT-OSS-20B	GSM8K	93.6	91.4	78.2	41.8	89.2	81.8	61.0	26.4
GPT-OSS-20B	MMLU-Pro	90.6	78.6	56.4	29.0	87.8	72.6	47.4	21.8

Table 6: **Setting-level prefix-retention rates (%) at higher perturbation magnitudes.** Results are pooled over the five sampled input locations.

		$\varepsilon/D_i = 0.75$				$\varepsilon/D_i = 1.00$			
Model	Benchmark	$K = 16$	$K = 32$	$K = 64$	$K = 128$	$K = 16$	$K = 32$	$K = 64$	$K = 128$
Qwen3-8B	ARC-Challenge	88.8	74.8	52.2	26.4	87.4	70.0	45.0	18.8
Qwen3-8B	GSM8K	85.8	69.0	45.8	24.2	81.2	62.0	37.8	17.4
Qwen3-8B	MMLU-Pro	90.6	72.2	53.2	29.6	89.4	70.8	49.4	20.8
DeepSeek-14B	ARC-Challenge	89.8	77.2	53.8	24.0	87.8	75.0	50.0	19.8
DeepSeek-14B	GSM8K	91.8	80.0	60.2	30.4	89.4	75.2	50.6	20.8
DeepSeek-14B	MMLU-Pro	92.8	81.6	60.2	31.8	92.2	77.8	51.8	27.0
GLM-Z1-9B	ARC-Challenge	84.0	68.0	43.6	15.4	81.0	63.8	38.2	12.8
GLM-Z1-9B	GSM8K	79.6	58.6	40.8	21.0	73.8	53.6	33.0	15.2
GLM-Z1-9B	MMLU-Pro	88.8	74.2	53.2	25.8	87.0	70.6	50.2	21.2
GPT-OSS-20B	ARC-Challenge	79.2	68.4	37.6	7.6	70.0	55.6	22.8	3.8
GPT-OSS-20B	GSM8K	77.6	56.4	26.2	6.4	72.0	47.4	13.8	2.4
GPT-OSS-20B	MMLU-Pro	79.2	58.0	32.2	11.4	76.4	56.2	30.4	11.4

Across the 48 model–benchmark–K profiles, retention at $\varepsilon = D_i$ is lower than at $\varepsilon = 0.25D_i$ in every case, and 44 profiles decrease strictly across all four tested magnitudes. The setting-level results

therefore reproduce the magnitude dependence summarized in Figure 3 across models, benchmarks, and target prefix lengths.

B.6 Additional Position–Magnitude Interaction

Table 7 reports the pooled retention rates across input position, perturbation magnitude, and target prefix length. Positional separation becomes substantially larger as the perturbation magnitude increases. For every fixed sampled location and perturbation magnitude, retention also decreases as K increases because the events $\{\tau \geq K\}$ are nested.

Table 7: **Pooled prefix retention across sampled input locations, perturbation magnitudes, and target prefix lengths.** Values are percentages averaged equally across the twelve model–benchmark settings.

ε/D_i	K	0.1	0.3	0.5	0.7	0.9
0.25	16	95.6	94.9	95.8	94.8	95.4
0.25	32	89.2	88.5	89.6	89.2	89.9
0.25	64	75.3	75.9	77.5	76.4	77.4
0.25	128	46.2	48.0	50.4	47.2	48.5
0.50	16	88.0	93.0	95.1	94.7	95.0
0.50	32	76.4	82.7	87.2	88.4	88.3
0.50	64	58.8	64.7	70.3	71.8	73.3
0.50	128	31.9	36.4	40.5	41.7	42.5
0.75	16	64.4	87.5	90.2	93.2	93.1
0.75	32	38.2	65.3	76.8	84.8	84.2
0.75	64	21.0	40.8	48.7	59.8	62.7
0.75	128	8.9	17.1	21.2	27.7	30.9
1.00	16	53.6	85.8	87.5	92.6	92.1
1.00	32	24.8	60.5	72.7	83.8	82.5
1.00	64	9.3	32.1	42.3	54.1	59.3
1.00	128	3.1	10.3	16.7	22.9	26.8

At $K = 64$, the latest sampled location has higher prefix retention than the earliest location in all twelve model–benchmark settings for

$$\varepsilon/D_i \in \{0.5, 0.75, 1\}.$$

The complete pooled table and setting-level profiles therefore show that the magnitude of the positional separation itself varies with perturbation strength.

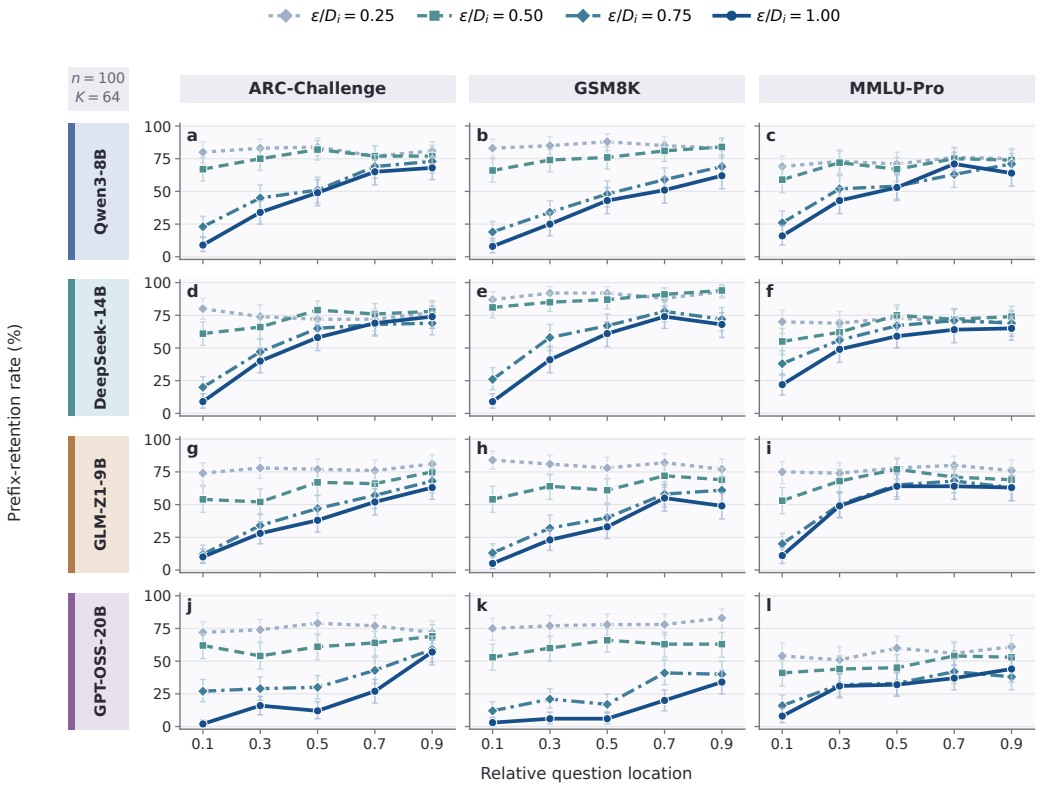

Figure 7: **Setting-level position–magnitude interaction at $K = 64$.** Each panel represents one model–benchmark setting. Curves show empirical prefix retention over the five sampled question locations at $\varepsilon/D_i \in \{0.25, 0.5, 0.75, 1\}$. Error bars denote 95% confidence intervals. At $\varepsilon/D_i \in \{0.5, 0.75, 1\}$, retention at the latest sampled location exceeds that at the earliest location in all twelve settings.

C Certificate Realization Details

We provide additional details for the certificate realization introduced in Section 4.4. We first describe the controlled reasoning model and numerical certificate construction, then analyze the tightness and position dependence of the resulting bounds, report certified radii and prefix guarantees, and finally perform independent generation checks.

C.1 Model and Multi-Step Reasoning Task

We use a two-layer decoder-only causal Transformer with two attention heads, hidden dimension $d = 64$, head dimension 32, and feed-forward dimension 128. Each attention head uses strict-predecessor softmax causal attention, i.e., the query at position j attends only to positions $w < j$. Therefore, the certification constants are computed directly from the strict-predecessor propagation recurrence in Appendix A.3. The model uses residual connections and tanh feed-forward transformations. Attention scores are bounded within $[-1, 1]$, providing a global bound for the diffuse-attention factor required by A4.

The model is trained on a synthetic multi-step arithmetic task. Each input contains four operands,

$$a_1 + a_2 + a_3 + a_4,$$

and the target trajectory explicitly generates three successive intermediate sums followed by the final result. Every target trajectory contains 32 generated tokens. The complete task contains $6^4 = 1{,}296$ inputs, divided into 960 training, 168 validation, and 168 held-out test examples using a fixed split. We train the model in FP32 using AdamW with a learning rate of 2×10^{-3}, weight decay 10^{-2}, batch

size 128, and random seed 20260918. Validation greedy exact-sequence accuracy is evaluated every 500 steps, and the checkpoint with the highest validation accuracy is selected, with ties broken by lower validation cross-entropy. Training stops at step 4,000, which is also the selected checkpoint. The selected checkpoint achieves 100% greedy exact-sequence accuracy on both the validation and held-out test sets.

Each formatted input contains nine tokens,

$$\langle \mathrm{bos} \rangle,\ a_1,\ +,\ a_2,\ +,\ a_3,\ +,\ a_4,\ \rightarrow .$$

Since the theory permits perturbations at every input position $1 \le i < n$ and here $n = 9$, we certify all eight admissible positions $i \in \{1, \dots, 8\}$. The final arrow token is excluded because it is the last input position.

For each held-out example, we certify perturbations at all eight admissible input positions. Certified radii are summarized at target prefix lengths

$$K \in \{4, 8, 16, 32\},$$

while the budget-to-prefix analysis evaluates every prefix length $K \in \{1, \dots, 32\}$.

The certificate-realization model was trained on one GPU of the same type using Python 3.11.15 and PyTorch 2.11.0+cu130, with FP32 training and FP64 generation evaluation. Verified constants and decision margins were computed on CPUs using outward-rounded binary64 interval arithmetic and mpmath 1.3.0, with 50-digit precision for transcendental-function endpoints.

C.2 Certificate Computation and Numerical Verification

Verified numerical setup. We compute numerical enclosures using custom interval routines with binary64 endpoints and the interval context of `mpmath` 1.3.0 at 50 decimal digits of precision. Elementary binary64 operations are outward-enclosed using `nextafter`, while exponential, logarithmic, and square-root endpoints are evaluated with directed rounding. Matrix products use explicit interval multiply–add operations, and hyperbolic tangents are enclosed using their monotone exponential representation.

The certification domain is fixed before computing the certificates as

$$\left\| \boldsymbol{\delta}_i^{(i)} \right\|_2 \le \bar{\varepsilon}, \qquad \bar{\varepsilon} = 1.$$

The uniform bounds cover all vocabulary tokens, all relevant positional embeddings, the eight evaluated input positions, and every common-prefix step within the 32-token generation horizon. All 5,376 closed-form certified radii used in the main realization lie within this verified domain.

Certified model constants. Let M bound the hidden-state norm, B_V bound the value-projection operator norm, L_s bound the attention-score sensitivity, and c_α bound the diffuse-attention factor. Since the attention scores are bounded within $[-1, 1]$,

$$c_\alpha = e^2.$$

The final transformation before the output projection is the identity, so $L_g = 1$ and

$$C_{\mathrm{out}} = \max_{v \in \mathcal{V}} \left\| (\mathbf{W}_{\mathrm{out}})_{v,:} \right\|_2 .$$

For the strict-predecessor architecture with $H = 2$ attention heads, the comparison recurrence in Appendix A.3 uses

$$a_\ell = L_\ell \left(1 + 2H c_\alpha B_V M L_s\right), \qquad b_\ell = L_\ell H c_\alpha B_V \left(1 + 2M L_s\right).$$

With $B_0 = 1$, $B_\ell = a_\ell B_{\ell-1}$, and $C_1 = b_1$,

$$C_\ell = a_\ell C_{\ell-1} + b_\ell \left(B_{\ell-1} + \frac{2}{\ell - 1} C_{\ell-1} \right), \qquad \ell \ge 2.$$

Table 8 summarizes the primitive and derived constants used in the realization. Displayed values are rounded upward; all computations use the full stored certified bounds.

Table 8: **Certified constants used in the certificate realization.** Displayed values are rounded upward from the stored certified bounds.

Constant	Certified upper bound
M	4.4789
B_V	3.1727
L_s	18.934
c_α	7.3891
L_1	20.650
L_2	92.458
C_{out}	5.4120
a_1	1.6423×10^5
b_1	1.6518×10^5
a_2	7.3531×10^5
b_2	7.3955×10^5
C_2	4.8721×10^{11}

Certified margins and radii. For every clean trajectory, interval evaluation yields a certified lower bound $\underline{m}_t$ on the clean decision margin,

$$\underline{m}_t \le \mathbf{z}_t(r_t) - \max_{v \neq r_t} \mathbf{z}_t(v) = m_t.$$

Using the certified upper bounds above, the closed-form propagation coefficient is evaluated as

$$\overline{A}_{\mathrm{cf}}(i,t) = \mathrm{rd}_{\uparrow}\left[C_{\mathrm{out}} C_L \frac{\left(1 + \log \frac{n+t-1}{i}\right)^{L-1}}{n+t-1} \right],$$

where $\mathrm{rd}_{\uparrow}$ denotes the computed upper enclosure. The corresponding certified radius is

$$\rho_{i,\mathrm{cf}}(K) = \min_{0 \le t < K} \mathrm{rd}_{\downarrow}\left[\frac{\underline{m}_t}{2\overline{A}_{\mathrm{cf}}(i,t)} \right]. \tag{72}$$

This is the numerical realization of $\rho_i^{\mathrm{cert}}(K)$ used in the main analysis. For a fixed perturbation budget ε, the complementary certified prefix length is

$$K_i^{\mathrm{cert}}(\varepsilon) = \max\left(\{0\} \cup \{K \in \{1,\dots,32\} : \varepsilon < \rho_{i,\mathrm{cf}}(K)\}\right). \tag{73}$$

Mechanism diagnostics. To expose the two inequalities underlying the realized prefix certificate, we use the common full-prefix budget

$$\varepsilon_0 = 3 \times 10^{-12}.$$

For each held-out example, we sample one Gaussian direction in $\mathbb{R}^{64}$ using a fixed seed, normalize it, and reuse the same direction across all eight evaluated input positions. The perturbations are scaled to remain within ε_0; the realized representation-space norms are approximately 2.99×10^{-12}.

At each generation step, we independently run the clean and perturbed greedy decoders in float64 and compute

$$d_{\mathrm{emp}}(t) = \left\|\widetilde{\mathbf{z}}_t - \mathbf{z}_t\right\|_\infty, \qquad d_{\mathrm{cert}}(t) = \overline{A}_{\mathrm{cf}}(i,t)\varepsilon_0.$$

Figure 4 reports

$$\frac{d_{\mathrm{emp}}(t)}{d_{\mathrm{cert}}(t)}, \qquad \frac{d_{\mathrm{cert}}(t)}{\underline{m}_t/2}.$$

For every generation step, the solid curves, shaded regions, and dashed curves show the median, interquartile range, and maximum, respectively, over all $1{,}344$ held-out example–position pairs.

C.3 Tightness and Position-Awareness of the Certificate

We next quantify two properties of the realized certificate: how the closed-form bound compares with the underlying discrete propagation recurrence, and the certified-radius benefit obtained by retaining the perturbed source position. All comparisons use the same checkpoint, certified clean-margin lower bounds, uniform model constants, and verified perturbation domain.

Direct-recurrence certificate. Before applying the logarithmic accumulation relaxation used in Theorem 3.1, the propagation bound can be evaluated directly from the nonnegative comparison recurrence. For a unit perturbation at source position i, initialize

$$p_{j,i}^{(0)} = \mathbf{1}_{\{j=i\}},$$

and for $j \geq 2$ compute

$$p_{j,i}^{(\ell)} = a_\ell p_{j,i}^{(\ell-1)} + \frac{b_\ell}{j-1} \sum_{w<j} p_{w,i}^{(\ell-1)}. \tag{74}$$

At $j = 1$, we set

$$p_{1,i}^{(\ell)} = a_\ell p_{1,i}^{(\ell-1)}.$$

Causality preserves $p_{j,i}^{(\ell)} = 0$ for $j < i$, while the source response propagates through the local-state term.

At generation step t, the direct-recurrence propagation coefficient is enclosed by

$$\overline{A}_{\mathrm{rec}}(i,t) = \mathrm{rd}_{\uparrow}\left[C_{\mathrm{out}} p_{n+t,i}^{(L)}\right],$$

which gives

$$\rho_{i,\mathrm{rec}}(K) = \min_{0\leq t<K} \mathrm{rd}_{\downarrow}\left[\frac{m_t}{2\overline{A}_{\mathrm{rec}}(i,t)}\right]. \tag{75}$$

An independent exact-rational evaluation of the two-layer expansion confirms that all direct-recurrence coefficients lie within their computed interval enclosures.

Position-agnostic certificate. To isolate the contribution of source-position information, we construct a position-agnostic baseline by replacing the position-dependent closed-form coefficient with its worst case over the eight admissible positions:

$$\overline{A}_{\mathrm{agn}}(t) = \max_{1\leq i\leq 8} \overline{A}_{\mathrm{cf}}(i,t).$$

Since the closed-form coefficient decreases with i, this maximum is attained at $i = 1$. The resulting certified radius is

$$\rho_{\mathrm{agn}}(K) = \min_{0\leq t<K} \mathrm{rd}_{\downarrow}\left[\frac{m_t}{2\overline{A}_{\mathrm{agn}}(t)}\right]. \tag{76}$$

This baseline removes the source-position information while leaving the certified margins, uniform model constants, and target prefix length unchanged.

We quantify the two effects using

$$G_{\mathrm{rec/cf}}(i,K) = \frac{\rho_{i,\mathrm{rec}}(K)}{\rho_{i,\mathrm{cf}}(K)}, \qquad G_{\mathrm{pos}}(i,K) = \frac{\rho_{i,\mathrm{cf}}(K)}{\rho_{\mathrm{agn}}(K)}.$$

Comparison results. Across all $168 \times 8 \times 4 = 5{,}376$ example–position–prefix combinations, the certified radii satisfy

$$\rho_{i,\mathrm{rec}}(K) \geq \rho_{i,\mathrm{cf}}(K) \geq \rho_{\mathrm{agn}}(K),$$

with no violations. Table 9 summarizes the resulting gains.

Table 9: **Tightness and position-awareness of the certificate.** For each target prefix length K, entries report the median [P10, P90] certified-radius ratio over the $1{,}344$ held-out example–position pairs.

K	$\rho_{i,\mathrm{rec}}/\rho_{i,\mathrm{cf}}$	$\rho_{i,\mathrm{cf}}/\rho_{\mathrm{agn}}$
4	$2.52\times$ [2.19, 2.73]	$1.84\times$ [1.00, 2.70]
8	$2.56\times$ [2.20, 2.73]	$1.79\times$ [1.00, 2.69]
16	$2.56\times$ [2.20, 2.73]	$1.79\times$ [1.00, 2.69]
32	$2.56\times$ [2.21, 2.73]	$1.72\times$ [1.00, 2.68]

The direct recurrence yields certified radii that are approximately 2.52–$2.56\times$ larger in median than the closed-form certificate, quantifying the effect of the analytical relaxations used to obtain its explicit position-dependent form. Retaining the source position yields a median certified-radius gain of 1.72–$1.84\times$ over the position-agnostic certificate. The gain equals one at $i = 1$ by construction and increases monotonically with the source position: all $4{,}704$ adjacent-position comparisons are strictly increasing, reaching approximately $2.70\times$ at $i = 8$.

Sources of closed-form looseness. For clarity, we describe the following decomposition at the algebraic coefficient level; all reported certified-radius ratios are evaluated using the same outward-rounding procedure as above.

For the two-layer realization, let $j = n + t$ and define

$$R_i(j) = 1 + \log \frac{j-1}{i}, \qquad H_i(j) = \sum_{r=i}^{j-2} \frac{1}{r},$$

together with

$$S = a_2 b_1 + b_2 a_1, \qquad U = b_2 b_1.$$

The corresponding unrounded direct-recurrence coefficient has the form

$$A_{\mathrm{rec}}(i,t) = \frac{C_{\mathrm{out}}}{j-1} \left[S + U H_i(j) \right].$$

To isolate the logarithmic accumulation relaxation, define

$$A_{\log}(i,t) = \frac{C_{\mathrm{out}}}{j-1} \left[S + 2U R_i(j) \right],$$

using the bound

$$H_i(j) \le 2R_i(j).$$

Its certified enclosure is

$$\overline{A}_{\log}(i,t) = \mathrm{rd}_{\uparrow} \left[A_{\log}(i,t) \right],$$

and the corresponding certified radius is

$$\rho_{i,\log}(K) = \min_{0 \le t < K} \mathrm{rd}_{\downarrow} \left[\frac{m_t}{2\overline{A}_{\log}(i,t)} \right].$$

The remaining closed-form step factors the common position term. Since $R_i(j) \ge 1$,

$$S + 2UR_i(j) \le (S + 2U) R_i(j).$$

For the two-layer realization,

$$C_2 = S + 2U,$$

so this step yields the closed-form coefficient underlying $\overline{A}_{\mathrm{cf}}(i,t)$.

We therefore define

$$G_{\mathrm{accum}}(i,K) = \frac{\rho_{i,\mathrm{rec}}(K)}{\rho_{i,\log}(K)}, \qquad G_{C_2}(i,K) = \frac{\rho_{i,\log}(K)}{\rho_{i,\mathrm{cf}}(K)}.$$

Across the four target prefix lengths, the logarithmic accumulation relaxation contributes a median factor of approximately 1.95–$1.96\times$, while the subsequent global-C_2 factorization contributes approximately 1.28–$1.31\times$. Together, these two steps account for the observed gap between the direct-recurrence and closed-form certificates.

C.4 Certified Radius and Prefix Guarantees

We retain the closed-form certificate for the absolute-radius analysis because it is the explicit position-dependent certificate characterized in Section 3.3; Appendix C.3 quantifies the tighter numerical radius available from the underlying direct recurrence. We now report the absolute certified radii obtained from the closed-form certificate. For each $K \in \{4, 8, 16, 32\}$, certified radii are computed for all $168 \times 8 = 1{,}344$ held-out example–position pairs. Table 10 summarizes their distribution.

At $K = 32$, the minimum certified radius across all evaluated example–position pairs is 3.397×10^{-12}. Thus, the common budget $\varepsilon_0 = 3 \times 10^{-12}$ certifies exact preservation of all 32 generated tokens for every evaluated pair.

Table 10: **Certified perturbation radii on the held-out test set.** Radii are shown in units of 10^{-12}.

K	Min	Median	P90
4	3.723	7.202	10.276
8	3.397	7.141	10.268
16	3.397	7.141	10.268
32	3.397	7.140	10.167

Figure 8 shows the complementary budget-to-prefix view of the same certificate. As the perturbation budget increases, certified prefix length decreases, while later input positions support larger certified regions.

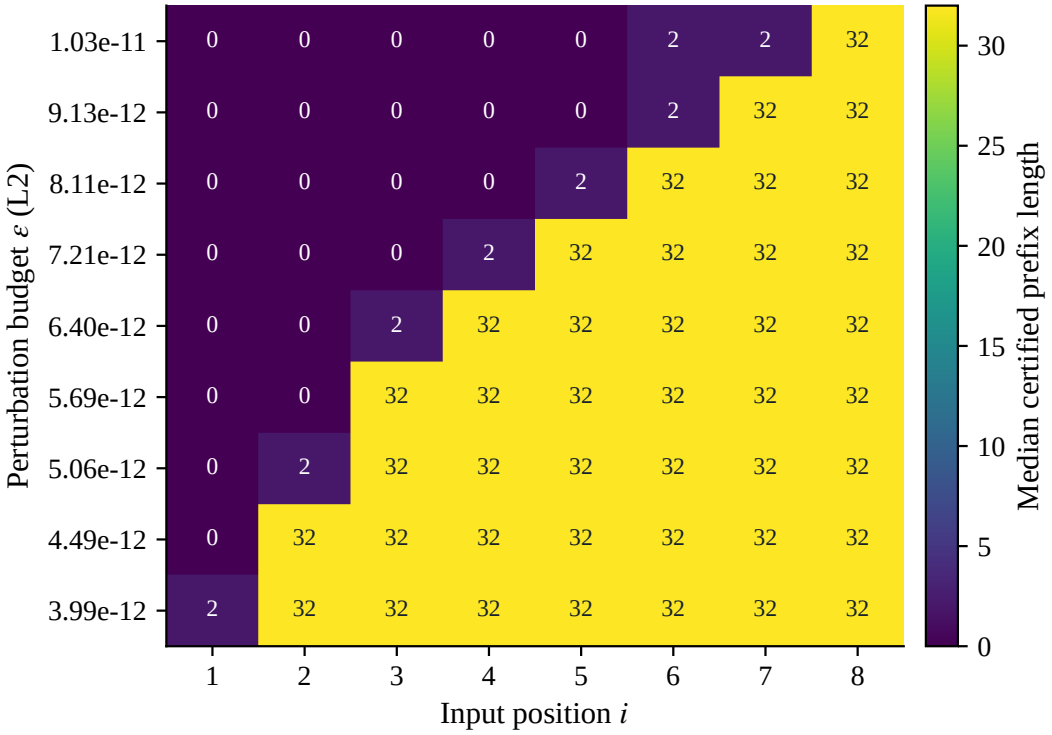

Figure 8: **Certified prefix length across input positions and perturbation budgets.** Each cell reports the median closed-form $K_i^{\mathrm{cert}}(\varepsilon)$ across the 168 held-out examples. Perturbation budgets are shown in units of 10^{-12}.

C.5 Independent Generation Checks

We additionally test the realized certificates through direct generation. We use the first three examples in the fixed held-out test split, positions $i \in \{1, 8\}$, and target prefix lengths $K \in \{4, 8, 16, 32\}$. Let $\rho = \rho_{i,\mathrm{cf}}(K)$ denote the computed lower enclosure of $\rho_i^{\mathrm{cert}}(K)$. We evaluate perturbation budgets $\varepsilon \in \{0.5\rho, 0.9\rho\}$.

For each example–position–prefix–budget combination, we sample 256 random perturbations in the L_2 ball using normalized Gaussian directions and radial factors $U^{1/d}$, with $U \sim \mathrm{Uniform}[0, 1)$.

For the projected-gradient search, the clean generated prefix $\mathbf{r}_{<t}$ is held fixed while optimizing the perturbation. The objective is

$$J_K\left(\boldsymbol{\delta}^{(i)}\right) = \max_{0 \le t < K} \left[\max_{v \neq r_t} \widetilde{\mathbf{z}}_t^{(i)}(v) - \widetilde{\mathbf{z}}_t^{(i)}(r_t) \right], \tag{77}$$

where the perturbed logits are evaluated under the fixed clean prefix. We run four float64 restarts, each with 40 objective evaluations and normalized gradient ascent. The strongest evaluated iterate from each restart is then tested with free greedy generation.

This yields $3 \times 2 \times 4 \times 2 = 48$ evaluated groups, comprising $12{,}288$ random and 192 projected-gradient generation checks. Across all $12{,}480$ checks, every perturbed trajectory preserves its certified prefix, with all realized perturbation norms remaining within the corresponding budget ε.