Numerical Fragility in Transformers: A Layer-wise Theory for Risk Estimation and Selective Stabilization

Jinwoo Baek
Department of Computer Science, Oregon State University

Abstract

Low-precision execution can induce substantial forward discrepancies in Transformers even for fixed weights and input, yet these discrepancies are usually monitored only at the output and lack a layer-wise theoretical account. We develop a first-order decomposition of output mismatch into layer-local attention, LayerNorm, and residual-transport terms, and derive from it a practical causal risk estimator and a budgeted controller, Bound-Guided Selective Stabilization (BGSS). Controlled sweeps verify the predicted local sign, monotonicity, and transport structure. On GPT-2, the transport-aware combined predictor is positively correlated with FP32-reference mismatch in all 18 runs and improves over a no-transport ablation in 17/18 runs. Reference-patch attribution shows that the same score preserves useful layer ordering information (mean Spearman 0.362). In budget-matched mitigation, BGSS outperforms random same-budget control in onset events (10.67 vs. 11.67), final mismatch (1.243×10^{-3} vs. 1.284×10^{-3}), and worst-case mismatch (3.14×10^{-3} vs. 8.49×10^{-3}), while matching a risk-only same-budget controller on onset suppression and sharply reducing worst-case mismatch (3.14×10^{-3} vs. 5.71×10^{-3}). These results support a theory-to-algorithm account of Transformer numerical fragility in which finite-precision risk can be analyzed, estimated, localized, and selectively stabilized.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

1 INTRODUCTION

Modern Transformer systems (Vaswani et al., 2017) are routinely executed in reduced precision to increase throughput and reduce memory cost (Micikevicius et al., 2018; Kalamkar et al., 2019; Dettmers et al., 2022; Yao et al., 2022; Frantar et al., 2022; Xiao et al., 2023; Lin et al., 2024; Bondarenko et al., 2023). Yet low-precision execution can produce substantial forward discrepancies even for fixed weights and input, and these discrepancies are often treated as a global implementation artifact rather than a structured internal phenomenon. Output-level mismatch reveals that a run is numerically unstable, but not which layers are responsible, which mechanisms dominate, or where a limited stabilization budget should be spent.

This paper develops a theory-to-algorithm account of Transformer numerical fragility. We start from a first-order layer-wise decomposition of finite-precision forward error into attention-side sensitivity, LayerNorm-driven instability with explicit ε-dependence, and downstream residual transport. From this decomposition, we derive a practical causal risk estimator and BGSS, a budgeted controller that selectively increases LayerNorm ε only where predicted risk is large and LayerNorm-dominated.

The empirical results support the full pipeline: controlled sweeps verify the predicted local sign, monotonicity, and transport structure; on GPT-2, the transport-aware combined predictor is positively correlated with FP32-reference mismatch in all 18 runs and improves over a no-transport ablation in 17/18 runs; reference-patch attribution preserves useful layer ordering information; and budget-matched intervention shows that BGSS improves on random same-budget control while substantially tightening worst-case behavior relative to a risk-only same-budget controller.

The main contribution is a unified view of numerical fragility in Transformers:

1. We develop a first-order layer-wise theory that combines attention-side sensitivity, residual relaxation,

and LayerNorm ε-dependence into a unified forward-fragility decomposition.

2. We derive a practical causal estimator and a budgeted selective controller, BGSS, from that decomposition rather than treating monitoring and mitigation as unrelated engineering heuristics.
3. We validate the resulting pipeline empirically through controlled local checks, GPT-2 end-to-end predictor evaluation, exact-ish attribution fidelity, and budget-matched intervention experiments.

Section 3 formalizes numerical fragility at the output and layer level. Section 4 develops the decomposition, Section 5 turns it into a causal estimator and BGSS, and Section 6 validates the resulting pipeline.

Code availability. The official code release for this paper is available at `https://github.com/JinwooBaek00/Numerical-Fragility-in-Transformers`.

2 RELATED WORK

Low-precision execution is primarily motivated by efficiency. Mixed-precision training, BF16 execution, and recent Transformer/LLM quantization methods such as Q8BERT, LLM.int8, ZeroQuant, GPTQ, SmoothQuant, AWQ, and Quantizable Transformers (Micikevicius et al., 2018; Kalamkar et al., 2019; Zafrir et al., 2019; Dettmers et al., 2022; Yao et al., 2022; Frantar et al., 2022; Xiao et al., 2023; Lin et al., 2024; Bondarenko et al., 2023) show that large models can often run accurately under aggressive precision constraints, but they typically evaluate numerical behavior through task accuracy or output-level degradation rather than a layer-wise causal account of where fragility originates.

Our work is also connected to classical numerical stability (Goldberg, 1991; Higham, 2002; IEEE Computer Society, 2019), to architectural analyses of residual and normalization structure (He et al., 2016; Ba et al., 2016; Xiong et al., 2020; Zhang and Sennrich, 2019), and to activation-patching or causal-tracing style analyses that localize internal responsibility by replacing hidden activations (Vig et al., 2020; Meng et al., 2022). We draw on these perspectives, but our goal is different: we start from output mismatch and derive a unified layer-wise risk score together with a budgeted selective stabilizer.

3 PROBLEM FORMULATION

We formulate numerical fragility in Transformers as a layer-wise causal risk estimation problem under finite-precision execution. Our focus is not optimization, approximation, or generalization error, but the forward discrepancy induced purely by finite-precision arithmetic inside a fixed model.

At optimization step t, let $X_{L,t}$ denote the exact final hidden state of a depth-L Transformer on the current minibatch, and let $\widetilde{X}_{L,t}$ denote the corresponding finite-precision output. We define the output-level numerical mismatch by

$$m_t := \frac{\|\widetilde{X}_{L,t} - X_{L,t}\|}{\|X_{L,t}\|}. \tag{1}$$

A central premise is that numerical fragility is *layer-local but globally accumulated*. Accordingly, the problem is not merely to monitor a scalar mismatch, but to construct a layer-wise predictor

$$\widehat{G}_t = (\widehat{G}_{1,t}, \dots, \widehat{G}_{L,t}), \qquad \widehat{R}_t := \sum_{\ell=1}^{L} \widehat{G}_{\ell,t},$$

such that $\widehat{R}_t$ tracks m_t while the coordinates $\widehat{G}_{\ell,t}$ identify which layers contribute most strongly to fragility. This layer-wise formulation supports both localization and control: it distinguishes attention-side from LayerNorm-driven risk and enables selective stabilization when only a few layers can be protected. We focus on the minimal intervention class of LayerNorm stabilizers, asking for an online rule that identifies layers whose predicted contribution is both large and LayerNorm-driven and stabilizes only those layers.

4 THEORY

We develop a layer-wise first-order theory for numerical fragility in Transformers and combine attention, residual, and LayerNorm effects into a unified forward-stability theorem.

4.1 Setup and Notation

We work under the standard floating-point model

$$\mathrm{fl}(a \circ b) = (a \circ b)(1+\delta), \qquad |\delta| \le \epsilon_{\mathrm{mach}},$$

for $\circ \in \{+, -, \times, \div\}$, where ϵ_{mach} is the machine precision of the active compute format. For structured kernels such as GEMMs, reductions, softmax, and LayerNorm, first-order rounding effects are absorbed into implementation-dependent constants. Unless stated otherwise, $\|\cdot\|$ denotes the Frobenius norm and $\|\cdot\|_2$ the spectral norm. For an invertible linear map T, write $\kappa(T) := \|T\|_2\|T^{-1}\|_2$. Let X_ℓ and $\widetilde{X}_\ell$ denote exact and finite-precision hidden states, let $E_\ell := \widetilde{X}_\ell - X_\ell$, and for a residual block $x \mapsto x + f(x)$ define $\rho_f := \|J_f\|_2$. The small-gain regime $\rho_f < 1$ is used only in Theorem 2 and Corollary 1; the unified forward-error result later needs only $\|I + J_f\|_2 \le 1 + \rho_f$.

4.2 Self-Attention Forward Error

Let $Q, K, V \in \mathbb{R}^{n\times d}$, where n is the sequence length and d is the head width. Define

$$S = \frac{1}{\sqrt{d}} QK^\top \in \mathbb{R}^{n\times n}, \qquad P = \mathrm{softmax}(S),$$

$$A = PV \in \mathbb{R}^{n\times d}.$$

Let $\mathcal{S} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ denote the row-wise softmax map. For a probability row $p \in \mathbb{R}^n$, define

$$J(p) = \mathrm{Diag}(p) - pp^\top, \qquad 0 \le \|J(p)\|_2 \le \tfrac{1}{2}.$$

Since $D\mathcal{S}_S$ is block diagonal with row blocks $J(P_{i:})$, let

$$d_{\mathrm{smx}} = \|D\mathcal{S}_S\|_{F\to F} = \max_{1\le i\le n} \|J(P_{i:})\|_2.$$

We then define

$$\kappa_{\mathrm{softmax}} := \frac{\|S\|}{\|P\|}\, d_{\mathrm{smx}}, \qquad \chi_{\mathrm{score}} := \frac{\|Q\|\,\|K\|}{\|P\|\,\sqrt{d}}\, d_{\mathrm{smx}},$$

and, whenever $\|A\| > 0$,

$$\kappa_{\mathrm{val}} := \frac{\|P\|\,\|V\|_2}{\|A\|}.$$

Theorem 1 (Self-Attention Forward Error)**.** *Under the floating-point model above, the finite-precision result $\widetilde{A}$ of $A = PV$ satisfies*

$$\|\widetilde{A} - A\| \le \Big[c_{\mathrm{smx}} + \kappa_{\mathrm{softmax}} + c_{\mathrm{gemm}}\chi_{\mathrm{score}} + c'_{\mathrm{gemm}}\Big]\epsilon_{\mathrm{mach}}\,\|P\|\,\|V\|_2 + O(\epsilon^2_{\mathrm{mach}}). \tag{2}$$

If additionally $\|A\| > 0$, then

$$\frac{\|\widetilde{A} - A\|}{\|A\|} \le \Big[c_{\mathrm{smx}} + \kappa_{\mathrm{softmax}} + c_{\mathrm{gemm}}\chi_{\mathrm{score}} + c'_{\mathrm{gemm}}\Big] \times\ \epsilon_{\mathrm{mach}}\,\kappa_{\mathrm{val}} + O(\epsilon^2_{\mathrm{mach}}).$$

The bound separates direct softmax sensitivity, score-formation error, and value amplification; proof appears in App. A.2.

4.3 Residual Relaxation

Residual connections do not remove local floating-point error, but they can weaken its depth-wise accumulation. Let a residual block be $x \mapsto x + f(x)$ and write J_f for the Jacobian of f.

Theorem 2 (Residual Stabilization)**.** *If $\|J_f\|_2 < 1$, then the linearized residual map $T = I + J_f$ is invertible and satisfies*

$$\kappa(T) = \|I + J_f\|_2\,\|(I + J_f)^{-1}\|_2 \le \frac{1 + \|J_f\|_2}{1 - \|J_f\|_2}.$$

Hence, under the small-gain condition, residual connections relax depth-wise compounding by keeping the local linearized map well-conditioned.

Corollary 1 (Depth-wise relaxation)**.** *For a stack of residual blocks with Jacobians $\{J_{f_\ell}\}_{\ell=1}^L$ and $\rho_\ell := \|J_{f_\ell}\|_2 < 1$,*

$$\kappa\Big(\prod_{\ell=1}^{L}(I + J_{f_\ell})\Big) \le \prod_{\ell=1}^{L}\frac{1+\rho_\ell}{1-\rho_\ell}.$$

In the first-order predictor developed below, we use the relaxed downstream factor

$$\prod_{k=\ell+1}^{L}(1+\rho_k)$$

to capture this attenuation effect at the level of layer-wise accumulation.

The exact condition-number bound is $\prod_{\ell=1}^{L}\frac{1+\rho_\ell}{1-\rho_\ell}$, whereas the factor $\prod_{k=\ell+1}^{L}(1+\rho_k)$ used later is the corresponding first-order forward-error transport factor. Proofs appear in App. A.3.

4.4 LayerNorm Forward Error and ε-Regime Structure

For a feature vector $x \in \mathbb{R}^{d_{\mathrm{model}}}$, define the ε-dependent LayerNorm normalization path

$$Z^{\mathrm{LN}}_\varepsilon(x) := \mathrm{Diag}(\gamma)\frac{x - \mu(x)}{\sqrt{\sigma^2(x) + \varepsilon}},$$

so that the full LayerNorm output is

$$\mathrm{LN}(x) = Z^{\mathrm{LN}}_\varepsilon(x) + \beta.$$

Here $\mu(x)$ and $\sigma^2(x)$ denote the mean and variance over the normalized axis, and $\varepsilon > 0$ is the stabilizer. Because BGSS acts only through ε, we isolate the normalization path $Z^{\mathrm{LN}}_\varepsilon$ and absorb the final bias shift into the ε-independent remainder term of the unified theorem. We also track the scale-aware indicator $\rho_{\mathrm{LN}}(\varepsilon) := (\sigma^2(x)/\varepsilon)\, d_{\mathrm{model}}\,\epsilon_{\mathrm{mach}}$, which serves only as a practical proxy for entry into the ε-dominated regime.

Proposition 1 (First-order LayerNorm normalization–path forward error). Under the floating-point model, there exists a kernel- and dimension-dependent constant $a_{\mathrm{ln}} > 0$ such that

$$\|\widetilde{Z}^{\mathrm{LN}}_\varepsilon(x) - Z^{\mathrm{LN}}_\varepsilon(x)\| \le \epsilon_{\mathrm{mach}}\, M_{\mathrm{LN}}(x,\varepsilon) + O(\epsilon^2_{\mathrm{mach}}), \tag{3}$$

where

$$M_{\mathrm{LN}}(x,\varepsilon) := \|\mathrm{Diag}(\gamma)\|_2 \frac{\varepsilon + 2\sigma^2(x)}{(\sigma^2(x)+\varepsilon)^{3/2}}\big(a_{\mathrm{ln}}\|x\|_2\big). \tag{4}$$

If additionally $\|Z_\varepsilon^{\mathrm{LN}}(x)\| > 0$, then

$$\frac{\|\widetilde{Z}_\varepsilon^{\mathrm{LN}}(x) - Z_\varepsilon^{\mathrm{LN}}(x)\|}{\|Z_\varepsilon^{\mathrm{LN}}(x)\|} \le \epsilon_{\mathrm{mach}}\, C_{\mathrm{LN}}(x,\varepsilon) + O(\epsilon_{\mathrm{mach}}^2), \tag{5}$$

where

$$C_{\mathrm{LN}}(x,\varepsilon) := \frac{M_{\mathrm{LN}}(x,\varepsilon)}{\|Z_\varepsilon^{\mathrm{LN}}(x)\|}. \tag{6}$$

Moreover, for fixed x, γ, and a_{ln}, the ε-dependent factor

$$f_{\sigma^2(x)}(\varepsilon) := \frac{\varepsilon + 2\sigma^2(x)}{(\sigma^2(x)+\varepsilon)^{3/2}}$$

is strictly decreasing in $\varepsilon > 0$.

Proposition 1 gives the first-order ε-dependent normalization-path magnitude together with a relative coefficient when $\|Z_\varepsilon^{\mathrm{LN}}(x)\| > 0$; proof appears in App. A.4.

4.5 Unified Forward Stability

Let A_ℓ denote the exact attention-side coefficient

$$A_\ell := \Big[c_{\mathrm{smx}} + \kappa_{\mathrm{softmax},\ell} + c_{\mathrm{gemm}}\chi_{\mathrm{score},\ell} + c'_{\mathrm{gemm}}\Big] \times \|P_\ell\|\,\|V_\ell\|_2, \tag{7}$$

let $M_\ell^{\mathrm{LN}} := M_{\mathrm{LN}}(x_\ell, \varepsilon_\ell)$ be the exact LayerNorm normalization-path magnitude from Proposition 1. Let $\mathcal{R}_\ell$ be the finite set of remaining non-attention kernels in layer ℓ outside the ε-dependent normalization path, and for each $r \in \mathcal{R}_\ell$ assume

$$\|\Delta_{\ell,r}\| \le \epsilon_{\mathrm{mach}}\, b_{\ell,r} + O(\epsilon_{\mathrm{mach}}^2). \tag{8}$$

Let $M_\ell^{\mathrm{eff}} := \sum_{r\in\mathcal{R}_\ell} b_{\ell,r}$ denote the aggregate remainder magnitude, which upper-bounds the summed remainder perturbation by Lemma 5. The local fragility magnitude is

$$M_\ell := M_\ell^{\mathrm{eff}} + A_\ell + M_\ell^{\mathrm{LN}}. \tag{9}$$

Theorem 3 (Unified forward stability)**.** *Assume the absolute bound of Theorem 1 for every attention site and Proposition 1 for every LayerNorm site contributing to M_ℓ. For each residual block $x \mapsto x + f_\ell(x)$, let*

$$\rho_\ell := \|J_{f_\ell}(X_{\ell-1})\|_2.$$

Assume further that (8) holds for every $r \in \mathcal{R}_\ell$ and every layer ℓ, and that $\|X_L\| > 0$.

$$\frac{\|\widetilde{X}_L - X_L\|}{\|X_L\|} \le \epsilon_{\mathrm{mach}} \sum_{\ell=1}^{L} \frac{M_\ell}{\|X_L\|} \prod_{k=\ell+1}^{L} (1+\rho_k) + O(\epsilon_{\mathrm{mach}}^2). \tag{10}$$

Theorem 3 decomposes output-level mismatch into layer-wise local magnitudes transported by downstream residual factors, without any small-gain assumption since $\|I + J_{f_k}\|_2 \le 1 + \rho_k$. Proof appears in App. A.6. Any benign inter-layer norm ratios needed to express site-local absolute magnitudes relative to the final output scale are absorbed into the layer-dependent first-order constants defining M_ℓ.

4.6 From the Unified Bound to a Selective Controller

Theorem 3 naturally induces a monitored layer-wise risk score. At step t with $\|X_{L,t}\| > 0$, let

$$A_{\ell,t} := \Big[c_{\mathrm{smx}} + \kappa_{\mathrm{softmax},\ell,t} + c_{\mathrm{gemm}}\chi_{\mathrm{score},\ell,t} + c'_{\mathrm{gemm}}\Big] \times \|P_{\ell,t}\|\,\|V_{\ell,t}\|_2, \tag{11}$$

let $M_{\ell,t}^{\mathrm{LN}} := M_{\mathrm{LN}}(x_{\ell,t}, \varepsilon_{\ell,t})$ be the exact LayerNorm normalization-path magnitude at site (ℓ,t) from Proposition 1, and let $M_{\ell,t}^{\mathrm{eff}} := \sum_{r\in\mathcal{R}_\ell} b_{\ell,t,r}$ be the aggregate monitored-step remainder magnitude for the non-attention, non-normalization-path kernels in $\mathcal{R}_\ell$. Set

$$M_{\ell,t} := M_{\ell,t}^{\mathrm{eff}} + A_{\ell,t} + M_{\ell,t}^{\mathrm{LN}}, \tag{12}$$

where $\rho_{k,t} := \|J_{f_k}(X_{k-1,t})\|_2$ denotes the downstream residual Jacobian norm at step t. Define

$$G_{\ell,t} := \frac{M_{\ell,t}}{\|X_{L,t}\|} \prod_{k=\ell+1}^{L} (1+\rho_{k,t}). \tag{13}$$

Then Theorem 3 gives the first-order predictor

$$\frac{\|\widetilde{X}_{L,t} - X_{L,t}\|}{\|X_{L,t}\|} \lesssim \epsilon_{\mathrm{mach}} \sum_{\ell=1}^{L} G_{\ell,t}. \tag{14}$$

Here $G_{\ell,t}$ is the predicted layer-wise contribution to end-to-end fragility. To quantify whether this contribution is primarily LayerNorm-driven, define

$$\phi_{\ell,t} := \frac{M_{\ell,t}^{\mathrm{LN}}}{M_{\ell,t}}, \tag{15}$$

with the convention $\phi_{\ell,t} := 0$ when $M_{\ell,t} = 0$. Large $G_{\ell,t}$ indicates high risk, while large $\phi_{\ell,t}$ indicates LayerNorm dominance.

Proposition 2 (Monotone reduction of the frozen-scale LayerNorm contribution). Fix a monitored layer ℓ and step t. Let $v_{\ell,t} := \sigma^2(x_{\ell,t})$ and let $a_{\ell,t} > 0$ denote the positive constant inherited from Proposition 1 at the monitored site (ℓ,t). Hold $x_{\ell,t}$, $v_{\ell,t}$, $a_{\ell,t}$, and γ_ℓ, $M_{\ell,t}^{\mathrm{eff}}$, $A_{\ell,t}$, $\|X_{L,t}\|$, and all non-normalization-path coefficients fixed. Define

$$M_{\mathrm{LN},\ell,t}^{\mathrm{frz}}(\varepsilon) := \|\mathrm{Diag}(\gamma_\ell)\|_2 \frac{\varepsilon + 2v_{\ell,t}}{(v_{\ell,t}+\varepsilon)^{3/2}} \big(a_{\ell,t}\,\|x_{\ell,t}\|_2\big). \tag{16}$$

Then $M^{\mathrm{frz}}_{\mathrm{LN},\ell,t}$ is nonincreasing in $\varepsilon > 0$, and it is strictly decreasing whenever $\|\mathrm{Diag}(\gamma_\ell)\|_2 \, \|x_{\ell,t}\|_2 > 0$. Moreover,

$$\begin{aligned}\frac{d}{d\varepsilon} M^{\mathrm{frz}}_{\mathrm{LN},\ell,t}(\varepsilon) :=& -\frac{\|\mathrm{Diag}(\gamma_\ell)\|_2}{2}\big(a_{\ell,t}\,\|x_{\ell,t}\|_2\big) \\ &\times \frac{\varepsilon + 4v_{\ell,t}}{(v_{\ell,t}+\varepsilon)^{5/2}} \\ \le& \, 0.\end{aligned} \tag{17}$$

Consequently, the frozen-scale layer-wise contribution

$$G^{\mathrm{frz}}_{\ell,t}(\varepsilon) := \frac{M^{\mathrm{eff}}_{\ell,t} + A_{\ell,t} + M^{\mathrm{frz}}_{\mathrm{LN},\ell,t}(\varepsilon)}{\|X_{L,t}\|} \prod_{k=\ell+1}^{L} (1+\rho_{k,t}) \tag{18}$$

is nonincreasing in ε, and for any $\varepsilon' \ge \varepsilon$

$$\begin{aligned}G^{\mathrm{frz}}_{\ell,t}(\varepsilon) - G^{\mathrm{frz}}_{\ell,t}(\varepsilon') =& \frac{1}{\|X_{L,t}\|} \\ &\times \Big(M^{\mathrm{frz}}_{\mathrm{LN},\ell,t}(\varepsilon) - M^{\mathrm{frz}}_{\mathrm{LN},\ell,t}(\varepsilon')\Big) \\ &\times \prod_{k=\ell+1}^{L} (1+\rho_{k,t}).\end{aligned} \tag{19}$$

Proposition 2 shows that, in the frozen-scale first-order model, increasing ε decreases the LayerNorm part of the predicted layer-wise contribution, motivating selective intervention on layers with both large $G_{\ell,t}$ and large $\phi_{\ell,t}$. In the algorithmic implementation, common positive kernel-dependent multipliers are absorbed into controller thresholds. Proof appears in App. A.5.

5 ALGORITHM

We instantiate the theory of Section 4 as *Bound-Guided Selective Stabilization* (BGSS), an online controller that monitors layer-wise numerical fragility and selectively increases LayerNorm stabilizers. BGSS leaves model weights, optimizer states, and architectural blocks unchanged; it acts only through LayerNorm ε values. Throughout this section, hats denote quantities measured on the monitored finite-precision forward pass at step t. BGSS is theory-guided rather than exactly calibrated: the practical statistics below need not equal the exact coefficients from Section 4.6, but they are required to be nonnegative, causal, and to preserve the exact score structure

local magnitude
+ downstream residual transport
+ monotone ε-dependence
of the LayerNorm term.

Unknown hardware/runtime-dependent scale factors are therefore absorbed into the controller thresholds.

5.1 Practical Layer-Wise Risk Estimation

At a monitored step t with $\|\widehat{X}_{L,t}\| > 0$, the exact coefficients in Section 4.6 are either unavailable or unnecessarily costly to compute online. BGSS therefore replaces them with causal surrogates computed from the monitored pass. For each layer ℓ, let $\widehat{S}_{\ell,t}$, $\widehat{Q}_{\ell,t}$, $\widehat{K}_{\ell,t}$, $\widehat{P}_{\ell,t}$, $\widehat{V}_{\ell,t}$, and $\widehat{Z}^{\mathrm{LN}}_{\ell,t}$ denote the observed score, query, key, attention-probability, value, and LayerNorm normalization-path tensors. Let $\widehat{M}^{\mathrm{eff}}_{\ell,t} \ge 0$ denote any causal surrogate of the remaining non-attention, non-normalization-path local magnitude, and let $\widehat{\rho}_{k,t} \ge 0$ denote any causal surrogate of the downstream residual Jacobian norm $\rho_{k,t}$.

We first estimate the local softmax differential by

$$\widehat{d}_{\mathrm{smx},\ell,t} := \widehat{\|D\mathcal{S}\|}_{\ell,t},$$

$$\widehat{\kappa}_{\mathrm{softmax},\ell,t} := \frac{\|\widehat{S}_{\ell,t}\|}{\|\widehat{P}_{\ell,t}\|}\,\widehat{d}_{\mathrm{smx},\ell,t},$$

$$\widehat{\chi}_{\mathrm{score},\ell,t} := \frac{\|\widehat{Q}_{\ell,t}\|\,\|\widehat{K}_{\ell,t}\|}{\|\widehat{P}_{\ell,t}\|\sqrt{d}}\,\widehat{d}_{\mathrm{smx},\ell,t},$$

$$\widehat{A}_{\ell,t} := \big(\widehat{\kappa}_{\mathrm{softmax},\ell,t} + \widehat{\chi}_{\mathrm{score},\ell,t}\big)\|\widehat{P}_{\ell,t}\|\,\|\widehat{V}_{\ell,t}\|_2,$$

and use the monotone LayerNorm surrogate

$$\widehat{C}_{\mathrm{LN},\ell,t} := \frac{\varepsilon_{\ell,t} + 2\widehat{\sigma}^2_{\ell,t}}{(\widehat{\sigma}^2_{\ell,t} + \varepsilon_{\ell,t})^{3/2}}.$$

This preserves the exact monotone dependence on ε while omitting shared positive kernel-dependent multipliers. Define the surrogate LayerNorm magnitude

$$\widehat{M}^{\mathrm{LN}}_{\ell,t} := \|\widehat{Z}^{\mathrm{LN}}_{\ell,t}\|\,\widehat{C}_{\mathrm{LN},\ell,t},$$

$$\widehat{M}_{\ell,t} := \widehat{M}^{\mathrm{eff}}_{\ell,t} + \widehat{A}_{\ell,t} + \widehat{M}^{\mathrm{LN}}_{\ell,t},$$

the practical layer-wise risk score

$$\widehat{G}_{\ell,t} := \frac{\widehat{M}_{\ell,t}}{\|\widehat{X}_{L,t}\|} \prod_{k=\ell+1}^{L} (1+\widehat{\rho}_{k,t}), \tag{20}$$

and the LayerNorm dominance ratio

$$\widehat{\phi}_{\ell,t} := \frac{\widehat{M}^{\mathrm{LN}}_{\ell,t}}{\widehat{M}_{\ell,t}}. \tag{21}$$

We adopt the convention $\widehat{\phi}_{\ell,t} := 0$ when $\widehat{M}_{\ell,t} = 0$.

To detect entry into the ε-dominated regime, BGSS also tracks

$$\widehat{\rho}_{\mathrm{LN},\ell,t} := \frac{\widehat{\sigma}^2_{\ell,t}}{\varepsilon_{\ell,t}}\, d_{\mathrm{model}}\, \epsilon_{\mathrm{mach}}, \tag{22}$$

so that $\widehat{G}_{\ell,t}$ preserves the exact

$$\text{local magnitude / final-output scale} \times \text{downstream transport}$$

structure, $\widehat{\phi}_{\ell,t}$ measures how much of that surrogate risk is LayerNorm-driven, and $\widehat{\rho}_{\mathrm{LN},\ell,t}$ tests whether the layer is in the ε-dominated regime. We do not claim exact calibration of $\widehat{G}_{\ell,t}$ across kernels or runtimes; BGSS uses these quantities for ranking, thresholding, and local action selection. Estimation details are given in App. A.7.

Selection rule. BGSS forms the eligible set

$$\mathcal{E}_t := \left\{ \ell : \begin{array}{l} \widehat{G}_{\ell,t} \geq \tau_G, \\ \widehat{\phi}_{\ell,t} \geq \tau_\phi, \\ \widehat{\rho}_{\mathrm{LN},\ell,t} < 1 \end{array} \right\},$$

where τ_G is a risk threshold and τ_ϕ is a LayerNorm-dominance threshold. A layer is therefore eligible only when its predicted contribution is large, LayerNorm-dominant, and currently in the ε-dominated regime. Layers under active cooldown are removed from $\mathcal{E}_t$, and BGSS keeps the top-B remaining layers ranked by $\widehat{G}_{\ell,t}$.

Update rule. For each selected layer ℓ, BGSS applies the monotone update

$$\varepsilon_{\ell,t}^{\mathrm{cand}} := \mathrm{clip}\left(\frac{\widehat{\sigma}_{\ell,t}^2 \, d_{\mathrm{model}} \, \epsilon_{\mathrm{mach}}}{\rho_*}, \; \varepsilon_{\min}, \; \varepsilon_{\max} \right), \quad (23)$$

$$\varepsilon_{\ell,t+1} := \max\left\{ \varepsilon_{\ell,t}, \, \varepsilon_{\ell,t}^{\mathrm{cand}} \right\}, \quad (24)$$

where $\rho_* \in (0,1)$ is the target post-update LayerNorm ratio. All non-selected layers keep

$$\varepsilon_{\ell,t+1} = \varepsilon_{\ell,t}.$$

The candidate in (23) is the smallest value inside $[\varepsilon_{\min}, \varepsilon_{\max}]$ that would enforce $\widehat{\rho}_{\mathrm{LN},\ell,t} \leq \rho_*$ on the monitored statistics whenever feasible; the outer max in (24) is then the smallest monotone bounded update. If the target value exceeds $\varepsilon_{\max}$, the rule saturates at $\varepsilon_{\max}$ and therefore achieves the best feasible reduction of the monitored ratio subject to the box constraint.

Under frozen monitored statistics, the post-update monitored ratio satisfies

$$\frac{\widehat{\sigma}_{\ell,t}^2 \, d_{\mathrm{model}} \, \epsilon_{\mathrm{mach}}}{\varepsilon_{\ell,t+1}} \leq \rho_*$$

$$\text{whenever} \quad \frac{\widehat{\sigma}_{\ell,t}^2 \, d_{\mathrm{model}} \, \epsilon_{\mathrm{mach}}}{\rho_*} \leq \varepsilon_{\max},$$

and the frozen surrogate LayerNorm magnitude

$$\widehat{M}_{\ell,t}^{\mathrm{LN,frz}}(\varepsilon) := \|\widehat{Z}_{\ell,t}^{\mathrm{LN}}\| \frac{\varepsilon + 2\widehat{\sigma}_{\ell,t}^2}{(\widehat{\sigma}_{\ell,t}^2 + \varepsilon)^{3/2}}$$

Algorithm 1: Bound-Guided Selective Stabilization (BGSS)

Input: monitor interval m; thresholds (τ_G, τ_ϕ); target ratio ρ_*; budget B; cooldown c; bounds $(\varepsilon_{\min}, \varepsilon_{\max})$

Initialize cooldown counters to zero for all layers
for *optimization step* $t = 1, 2, \ldots$ **do**
  perform the usual forward/backward/update step
  if $t \bmod m = 0$ **then**
    decrement all positive cooldown counters by one
    estimate $\widehat{G}_{\ell,t}$, $\widehat{\phi}_{\ell,t}$, and $\widehat{\rho}_{\mathrm{LN},\ell,t}$ for all layers ℓ using the activations observed at step t
    form $\mathcal{E}_t = \{\ell : \widehat{G}_{\ell,t} \geq \tau_G, \; \widehat{\phi}_{\ell,t} \geq \tau_\phi, \; \widehat{\rho}_{\mathrm{LN},\ell,t} < 1\}$
    remove layers with positive cooldown counters from $\mathcal{E}_t$
    keep the top-B layers in $\mathcal{E}_t$ ranked by $\widehat{G}_{\ell,t}$
    foreach *selected layer* ℓ **do**
      $\varepsilon_{\ell,t}^{\mathrm{cand}} \leftarrow \mathrm{clip}\big(\widehat{\sigma}_{\ell,t}^2 d_{\mathrm{model}} \epsilon_{\mathrm{mach}} / \rho_*, \varepsilon_{\min}, \varepsilon_{\max}\big)$
      $\varepsilon_{\ell,t+1}^{\mathrm{new}} \leftarrow \max\{\varepsilon_{\ell,t}, \varepsilon_{\ell,t}^{\mathrm{cand}}\}$
      if $\varepsilon_{\ell,t+1}^{\mathrm{new}} > \varepsilon_{\ell,t}$ **then**
        set $\varepsilon_{\ell,t+1} \leftarrow \varepsilon_{\ell,t+1}^{\mathrm{new}}$
        set the cooldown counter of layer ℓ to c

is nonincreasing in ε. Consequently, the corresponding frozen surrogate layer score is also nonincreasing. This is the algorithmic analogue of Proposition 2.

Properties and theoretical grounding. BGSS is causal, selective, monotone, and mechanism-aware. Causality follows from using only statistics observed at step t and applying updates from step $t+1$ onward. Selectivity follows from the top-B budget on eligible layers. Monotonicity follows from (24). Mechanism-awareness follows from requiring both large overall risk ($\widehat{G}_{\ell,t}$) and large LayerNorm dominance ($\widehat{\phi}_{\ell,t}$). Finally, Theorem 3 provides the layer-wise decomposition underlying $\widehat{G}_{\ell,t}$, while Proposition 2 and the frozen-statistics argument above justify (24) as the smallest bounded monotone local risk-reducing action toward the target ratio ρ_*. Thus, BGSS is a causal surrogate-based algorithmic realization of the unified first-order fragility model under budgeted intervention constraints.

6 EXPERIMENTS

Evaluation protocol. Our empirical evaluation mirrors the theory-to-algorithm structure of the paper. Throughout, FP32 execution serves as the numerical reference. E1 uses synthetic controlled sweeps to test the local statements behind Theorem 1, Theorem 2, Corollary 1, and Proposition 1. E2 and E3 use the HuggingFace `gpt2` checkpoint with BF16/FP16 mon-

Table 1: Main GPT-2 evidence summary. Means are over completed runs; when reported, uncertainties are standard deviations across runs/seeds.

Exp.	Main evidence
E2	The transport-aware predictor achieves Pearson 0.370 ± 0.119 versus `no_transport` 0.206, improves correlation in 17/18 runs, and improves top-k retrieval in 12/18 runs.
E3	Reference-patch attribution yields mean Spearman 0.362 ± 0.156, pairwise accuracy 0.643 ± 0.061, top-3 overlap 0.505, and top-5 overlap 0.622.
E5	Relative to the random same-budget controller, BGSS reduces onset events (10.67 ± 1.15 vs. 11.67), final mismatch ($1.243 \times 10^{-3} \pm 2.87 \times 10^{-5}$ vs. 1.284×10^{-3}), and worst-case mismatch ($3.14 \times 10^{-3} \pm 1.01 \times 10^{-3}$ vs. 8.49×10^{-3}); versus the risk-only same-budget controller, BGSS matches mean onset events (10.67 vs. 10.67) while reducing worst-case mismatch (3.14×10^{-3} vs. 5.71×10^{-3}), at a slight cost in mean final mismatch (1.243×10^{-3} vs. 1.221×10^{-3}).

itored passes on WikiText-103 validation, sequence lengths $\{128, 512, 1024\}$, 3 seeds, and 96 monitored windows per run, for 18 completed runs. E3 reuses these E2 runs and evaluates layer attribution on the top-8 highest-mismatch windows per run. E5 uses stable FP32-master / FP16-shadow training on WikiText-2 train with sequence length 256, 256 monitored steps, 3 seeds, and a shared budget of 24 actions for the budget-matched controllers. Unless otherwise noted, aggregate statistics are means over completed runs. Table 1 collects the main GPT-2 evidence for the end-to-end predictor, attribution, and mitigation experiments in one place. Where variability is relevant, we report standard deviations across runs or across seeds.

6.1 E1: Controlled local validation

We begin with controlled numerical checks that isolate the three local mechanisms in the theory. The attention sweep varies score margin and value scale in a 2×2 toy attention map; the LayerNorm sweep varies ε over 10 logarithmically spaced values; and the residual sweep varies the local gain ρ over 8 values while composing depth-6 transport. These are not language-model runs; they are direct mechanism probes.

Figure 1 shows that the attention proxy tracks the measured attention-output perturbation almost perfectly (Pearson = 0.999999, Spearman = 1.0), as predicted by Theorem 1. In the LayerNorm sweep, both the measured normalization-path change and the causal proxy decrease monotonically with ε, matching Proposition 1. In the residual sweep, the measured downstream amplification remains below the predicted transport bound at all tested ρ values, matching Theorem 2 and Corollary 1. The corresponding LayerNorm and residual curves are reported in Appendix A.8. Overall, E1 verifies that the local signs, monotonicity, and transport directionality assumed by the unified decomposition are directly visible in controlled numerical measurements.

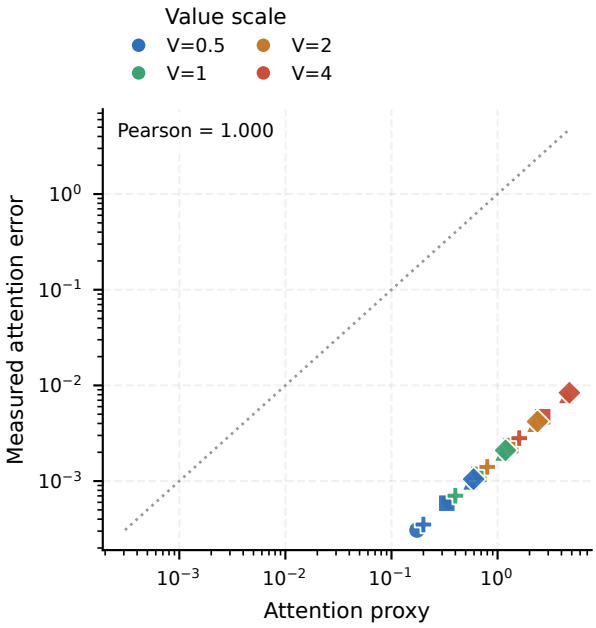

Figure 1: E1 attention-side controlled validation. Each point is one controlled attention configuration from the margin/value-scale sweep. The theory proxy and the measured attention-output perturbation align almost perfectly, confirming the sign and relative scaling predicted by Theorem 1.

6.2 E2: End-to-end predictor validation on GPT-2

We next test whether the practical transport-aware predictor tracks the final FP32-reference mismatch on real GPT-2 evaluation windows. For each monitored window, we compute the combined predictor $\widehat{R}_t = \sum_\ell \widehat{G}_{\ell,t}$ and compare it to the `no_transport` ablation, which removes the downstream residual transport factors from the same local decomposition. This is the primary E2 comparison: the unified theory specifically adds transport on top of local magnitudes, whereas the single-mechanism signals are diagnostic probes rather than the main baseline.

Across the 18 GPT-2 runs, the combined predictor is positively correlated with the final mismatch in all runs. Averaged over the full sweep, it achieves mean Pearson/Spearman correlations of 0.370 ± 0.119 and 0.351 ± 0.060, compared with 0.206/0.212 for `no_transport`. Adding transport improves correlation in 17/18 runs and improves top-k retrieval of high-mismatch windows in 12/18 runs. Figure 2 summarizes this comparison: almost all runs lie above the diagonal, showing that causal downstream transport is consis-

tently useful in practice. This is the main empirical validation of the transport-aware unified estimator from Section 5. A run-wise Δ-Pearson view and a binned risk-trend plot appear in Appendix A.8.

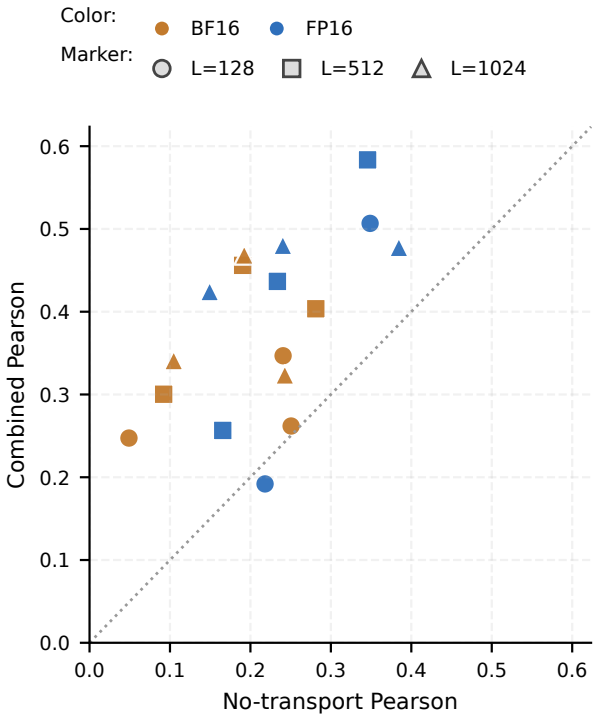

Figure 2: E2 end-to-end predictor validation on GPT-2. Each point is one run over a precision/sequence-length/seed combination. The transport-aware combined predictor almost always improves on the no-transport ablation, supporting the role of downstream residual transport in the unified risk estimator.

6.3 E3: Attribution and localization fidelity

E2 validates the scalar predictor. E3 asks whether the induced layerwise ranking is faithful to reference-patch layer importance. Starting from the E2 runs, we select the top-8 highest-mismatch windows from each run and rerun the same manual GPT-2 forward path while replacing exactly one low-precision block with its FP32 reference counterpart. The resulting reduction in final mismatch is the reference-patch effect for that layer.

The first check is alignment: the recomputed baseline mismatch agrees with the source E2 mismatch to within an average absolute gap of only $9.24\times10^{-7}\pm3.47\times10^{-7}$, so the attribution comparison is not confounded by a different forward path. The second check is ranking fidelity. Across the 18 runs, we obtain mean Spearman correlation 0.362 ± 0.156, mean pairwise ordering accuracy 0.643 ± 0.061, mean top-3 overlap 0.505 ± 0.108, and mean top-5 overlap 0.622 ± 0.084. Figure 3 shows the aggregate proxy-rank versus exact-rank heatmap. The mass is concentrated near the diagonal, indicating that the practical layerwise proxy is not a perfect oracle but does preserve useful localization structure for the most influential layers. Run-level fidelity statistics are reported in Appendix A.8.

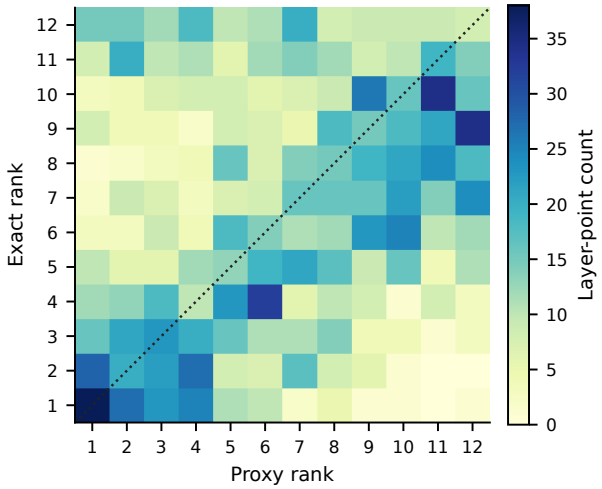

Figure 3: E3 attribution fidelity. The heatmap counts proxy-rank versus exact-ish rank over all evaluated layer-step pairs. Mass concentrated near the diagonal indicates that the practical layerwise proxy preserves meaningful exact layer ordering information.

6.4 E5: Budgeted mitigation utility

Finally, we test whether the theory-guided score is useful for intervention rather than explanation alone. We compare five policies during stable FP32-master / FP16-shadow training: no intervention, a static global LayerNorm-ε increase, a random same-budget controller, a risk-only same-budget controller that ranks layers by overall risk without the LayerNorm-dominance gate, and BGSS. The key comparisons are budget-matched: BGSS, the random controller, and the risk-only controller all use 24 actions and the same average protected layer-steps (2742), so performance gaps reflect policy quality rather than intervention volume.

Figure 4 shows the main robustness view. Relative to the random same-budget controller, BGSS reduces mean mismatch-onset events from 11.67 to 10.67 ± 1.15, lowers mean final mismatch from 1.284×10^{-3} to $1.243 \times 10^{-3} \pm 2.87 \times 10^{-5}$, and sharply reduces mean max mismatch from 8.49×10^{-3} to $3.14 \times 10^{-3} \pm 1.01 \times 10^{-3}$. The stronger mechanism-aware test is against the risk-only same-budget controller: both policies achieve the same mean onset-event count (10.67), and

the risk-only controller attains a slightly lower mean final mismatch (1.221×10^{-3} versus 1.243×10^{-3}), but BGSS reduces mean max mismatch from 5.71×10^{-3} to 3.14×10^{-3} and wins the worst-case comparison in all three seeds. Static global stabilization uses a larger protection budget (3072 protected layer-steps) yet still produces higher final mismatch and higher worst-case mismatch than BGSS. The conclusion is therefore not that BGSS minimizes every average metric, but that the theory-guided policy is the strongest *robust* budget-matched controller among the tested alternatives. A complementary tradeoff view is given in Appendix A.8.

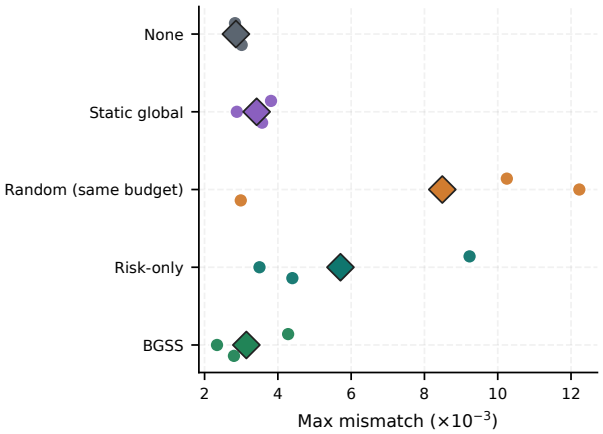

Figure 4: E5 budgeted mitigation, viewed through worst-case robustness. Small points show individual seeds and diamonds show policy means. Under the same budget as the random and risk-only controllers, BGSS substantially reduces max mismatch and yields the strongest robust budget-matched policy.

Summary. Taken together, E1 validates the local surrogates, E2 validates the end-to-end transport-aware predictor on GPT-2, E3 validates the resulting layer-wise localization, and E5 shows that the same score can drive a useful budgeted controller. The experiments therefore support the full pipeline of the paper: from first-order layerwise theory, to a practical online estimator, to a selective mitigation algorithm.

7 LIMITATIONS

Our analysis is intentionally first-order. The theory isolates the dominant finite-precision mechanisms arising from attention, LayerNorm, and residual transport, and E1–E3 validate precisely this regime. The practical estimator, however, still depends on causal surrogates for quantities such as softmax sensitivity, downstream transport, and effective remainder magnitude. These surrogates preserve the structure of the unified bound and work well empirically, but they are not online certificates of exact mismatch. Likewise, the implementation-dependent constants in the theory are treated as fixed for a given hardware/runtime pair rather than derived analytically for every kernel.

The empirical scope is also deliberate. Our large-model experiments use GPT-2 and LayerNorm-based stabilization, so we do not claim immediate quantitative transfer to architectures with substantially different normalization, routing, or attention implementations. In addition, BGSS studies a narrow but controlled intervention class—selective LayerNorm ε updates under an explicit budget—rather than the full space of numerical mitigation methods. Accordingly, the paper supports the claim that layer-wise fragility structure can drive useful selective stabilization, but it does not claim that BGSS is a globally optimal controller across architectures, kernels, and runtimes.

Future work. Several extensions are especially natural. On the theory side, an important next step is to derive sharper implementation-aware constants and to extend the analysis to RMSNorm, grouped-query or multi-query attention, mixture-of-experts routing, and more aggressive quantization regimes. On the algorithmic side, it would be valuable to enlarge the intervention space beyond LayerNorm ε bumps to include selective recomputation, mixed-precision routing, or budgeted precision promotion for specific kernels. On the empirical side, scaling the evaluation from GPT-2 to larger contemporary language models would test how the same layer-wise decomposition behaves under newer architectural choices and runtime stacks.

8 CONCLUSION

We presented a layer-wise theory of numerical fragility in Transformers that connects attention-side sensitivity, LayerNorm instability, and residual transport through a unified first-order forward-error decomposition. From this decomposition, we derived a practical causal risk estimator and a budgeted selective stabilization rule, BGSS. The experiments support the full pipeline: controlled sweeps verify the local mechanism statements, GPT-2 evaluation shows that the transport-aware combined predictor tracks FP32-reference mismatch, reference-patch attribution shows that the same score carries useful layer-localization information, and budget-matched intervention experiments show that it can support selective mitigation. Taken together, these results argue that finite-precision instability in Transformers is not merely a global numerical artifact, but a structured layer-wise phenomenon that can be analyzed, estimated, localized, and selectively acted upon using causal information available during execution.

References

Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bondarenko, Y., Nagel, M., and Blankevoort, T. (2023). Quantizable transformers: Removing outliers by helping attention heads do nothing. In *Advances in Neural Information Processing Systems*, volume 36.

Dettmers, T., Lewis, M., Belkada, Y., and Zettlemoyer, L. (2022). LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale. In *Advances in Neural Information Processing Systems*, volume 35.

Frantar, E., Ashkboos, S., Hoefler, T., and Alistarh, D. (2022). GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers. *arXiv preprint arXiv:2210.17323*.

Goldberg, D. (1991). What every computer scientist should know about floating-point arithmetic. *ACM Computing Surveys*, 23(1).

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Higham, N. J. (2002). *Accuracy and Stability of Numerical Algorithms*. SIAM, 2 edition.

IEEE Computer Society (2019). Ieee standard for floating-point arithmetic. IEEE Std 754-2019. DOI: 10.1109/IEEESTD.2019.8766229.

Kalamkar, D., Mudigere, D., Mellempudi, N., Das, D., Banerjee, K., Avancha, S., Vooturi, D. T., Jammalamadaka, N., Huang, J., Yuen, H., Yang, J., Park, J., Heinecke, A., Georganas, E., Srinivasan, S., Kundu, A., Smelyanskiy, M., Kaul, B., and Dubey, P. (2019). A study of BFLOAT16 for deep learning training. *arXiv preprint arXiv:1905.12322*.

Lin, J., Tang, J., Tang, H., Yang, S., Chen, W.-M., Wang, W.-C., Xiao, G., Dang, X., Gan, C., and Han, S. (2024). AWQ: Activation-aware Weight Quantization for On-Device LLM Compression and Acceleration. In *Proceedings of Machine Learning and Systems*, volume 6.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. (2022). Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. (2016). Pointer sentinel mixture models.

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., and Wu, H. (2018). Mixed precision training. In *International Conference on Learning Representations*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI technical report.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., and Shieber, S. (2020). Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33.

Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., and Han, S. (2023). SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning*.

Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., and Liu, T.-Y. (2020). On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning*.

Yao, Z., Aminabadi, R. Y., Zhang, M., Wu, X., Li, C., and He, Y. (2022). ZeroQuant: Efficient and Affordable Post-Training Quantization for Large-Scale Transformers. In *Advances in Neural Information Processing Systems*, volume 35.

Zafrir, O., Boudoukh, G., Izsak, P., and Wasserblat, M. (2019). Q8BERT: Quantized 8Bit BERT. In *Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing at NeurIPS*.

Zhang, B. and Sennrich, R. (2019). Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, volume 32.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers or curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A PROOFS AND ADDITIONAL DERIVATIONS

A.1 Floating-Point Kernel Model and Basic Lemmas

We adopt the scalar FP model $\mathrm{fl}(a \circ b) = (a \circ b)(1+\delta)$ with $|\delta| \le \epsilon_{\mathrm{mach}}$ and $\circ \in \{+, -, \times, \div\}$. For structured kernels (reductions, GEMMs, softmax, layer normalization), we use standard first-order bounds that collect rounding into implementation-dependent constants (independent of the specific model instance). We record the basic lemmas used in the proofs.

Lemma 1 (GEMM first-order score bound)**.** *Let $\widetilde{S} = \mathrm{fl}(QK^\top/\sqrt{d})$. Then*

$$\|\widetilde{S} - S\| \le \epsilon_{\mathrm{mach}} \left(\|S\| + c_{\mathrm{gemm}} \frac{\|Q\|\,\|K\|}{\sqrt{d}} \right) + O(\epsilon_{\mathrm{mach}}^2).$$

Proof. Write $\widetilde{S} = \frac{1}{\sqrt{d}}\mathrm{fl}(QK^\top) = \frac{1}{\sqrt{d}}(QK^\top + E)$ with $\|E\| \le c_{\mathrm{gemm}}\epsilon_{\mathrm{mach}}\|Q\|\,\|K\|$. The subsequent scalar division contributes a multiplicative factor $(1+\delta)$ with $|\delta| \le \epsilon_{\mathrm{mach}}$, so

$$\widetilde{S} = \frac{QK^\top + E}{\sqrt{d}}(1+\delta) = S + S\delta + \frac{E}{\sqrt{d}} + O(\epsilon_{\mathrm{mach}}^2).$$

Taking norms and using triangle inequality gives the claim. □

Lemma 2 (GEMM output forward bound)**.** *Let $\widetilde{A} = \mathrm{fl}(MV)$ for compatible matrices M, V. Then*

$$\widetilde{A} = MV + E_{MV},$$

with

$$\|E_{MV}\| \le c'_{\mathrm{gemm}}\epsilon_{\mathrm{mach}}\|M\|\,\|V\|_2 + O(\epsilon_{\mathrm{mach}}^2).$$

Proof. Write the rows of M as $m_i^\top$. Standard first-order backward-error analysis for each row-times-matrix product gives row perturbations Δm_i such that the computed ith output row satisfies

$$\widetilde{a}_i^\top = (m_i + \Delta m_i)^\top V, \qquad \|\Delta m_i\|_2 \le \gamma_{\mathrm{gemm}}\|m_i\|_2,$$

with

$$\gamma_{\mathrm{gemm}} = c'_{\mathrm{gemm}}\epsilon_{\mathrm{mach}} + O(\epsilon_{\mathrm{mach}}^2),$$

where c'_{gemm} depends only on the GEMM kernel/runtime. Stacking the rows defines a perturbation matrix ΔM satisfying

$$\widetilde{A} = (M + \Delta M)V = MV + E_{MV}, \qquad E_{MV} := \Delta M\, V.$$

Moreover,

$$\|\Delta M\|_F^2 = \sum_i \|\Delta m_i\|_2^2 \le \gamma_{\mathrm{gemm}}^2 \sum_i \|m_i\|_2^2 = \gamma_{\mathrm{gemm}}^2\|M\|_F^2.$$

Therefore,

$$\|E_{MV}\|_F \le \|\Delta M\|_F\|V\|_2 \le \gamma_{\mathrm{gemm}}\|M\|_F\|V\|_2 = c'_{\mathrm{gemm}}\epsilon_{\mathrm{mach}}\|M\|\,\|V\|_2 + O(\epsilon_{\mathrm{mach}}^2),$$

which is the claimed bound. □

Lemma 3 (Softmax kernel first-order forward bound)**.** *Let $\mathcal{S}$ denote the row-wise softmax map, let*

$$\widehat{P} := \mathcal{S}(S + \Delta S),$$

and let $\widetilde{P} = \mathrm{fl}_{\mathrm{smx}}(S + \Delta S)$ be the finite-precision softmax output produced from the perturbed scores $S + \Delta S$, where $\|\Delta S\| = O(\epsilon_{\mathrm{mach}})$. Then

$$\widetilde{P} = \widehat{P} + E_{\mathrm{smx}},$$

with

$$\|E_{\mathrm{smx}}\| \le c_{\mathrm{smx}}\epsilon_{\mathrm{mach}}\|\widehat{P}\| + O(\epsilon_{\mathrm{mach}}^2).$$

Equivalently,

$$\|E_{\mathrm{smx}}\| \le c_{\mathrm{smx}}\epsilon_{\mathrm{mach}}\|P\| + O(\epsilon_{\mathrm{mach}}^2),$$

where $P = \mathcal{S}(S)$.

Proof. The first bound is the standard first-order forward model for the implementation of the row-wise softmax kernel, with c_{smx} depending only on the runtime/kernel. Since $\mathcal{S}$ is smooth and $\|\Delta S\| = O(\epsilon_{\mathrm{mach}})$,

$$\widehat{P} = \mathcal{S}(S + \Delta S) = P + D\mathcal{S}_S[\Delta S] + O(\|\Delta S\|^2) = P + O(\epsilon_{\mathrm{mach}}).$$

Hence

$$\|\widehat{P}\| \le \|P\| + O(\epsilon_{\mathrm{mach}}),$$

and multiplying by ϵ_{mach} gives

$$\epsilon_{\mathrm{mach}}\|\widehat{P}\| = \epsilon_{\mathrm{mach}}\|P\| + O(\epsilon_{\mathrm{mach}}^2),$$

which yields the second form. □

Lemma 4 (Row-wise softmax Jacobian norm)**.** *For $p = \mathrm{softmax}(s) \in \mathbb{R}^n$, $J(p) = \mathrm{Diag}(p) - pp^\top$ satisfies $0 \le \|J(p)\|_2 \le \frac{1}{2}$, with the maximum approached near two-way ties.*

Proof. $J(p)$ is a covariance matrix of a categorical distribution with probabilities p, hence PSD with operator norm bounded by the largest variance along any direction in the probability simplex. The extremum occurs when mass is split between two coordinates, giving $\|J(p)\|_2 = \frac{1}{2}$. □

Lemma 5 (Closure of local first-order coefficients)**.** *Let $\mathcal{R}_\ell$ be a finite index set and suppose that, for each $r \in \mathcal{R}_\ell$,*

$$\|\Delta_{\ell,r}\| \le \epsilon_{\mathrm{mach}}\, b_{\ell,r} + O(\epsilon_{\mathrm{mach}}^2).$$

Define

$$\Delta_\ell^{\mathrm{eff}} := \sum_{r\in\mathcal{R}_\ell} \Delta_{\ell,r}, \qquad M_\ell^{\mathrm{eff}} := \sum_{r\in\mathcal{R}_\ell} b_{\ell,r}.$$

Then

$$\|\Delta_\ell^{\mathrm{eff}}\| \le \epsilon_{\mathrm{mach}}\, M_\ell^{\mathrm{eff}} + O(\epsilon_{\mathrm{mach}}^2).$$

Proof. By triangle inequality,

$$\|\Delta_\ell^{\mathrm{eff}}\| \le \sum_{r\in\mathcal{R}_\ell} \|\Delta_{\ell,r}\|.$$

Applying the assumed bound termwise gives

$$\|\Delta_\ell^{\mathrm{eff}}\| \le \epsilon_{\mathrm{mach}} \sum_{r\in\mathcal{R}_\ell} b_{\ell,r} + \sum_{r\in\mathcal{R}_\ell} O(\epsilon_{\mathrm{mach}}^2).$$

Because $\mathcal{R}_\ell$ is finite, the last sum remains $O(\epsilon_{\mathrm{mach}}^2)$, yielding the claim. □

A.2 Proof of Theorem 1 (Self-Attention Forward Error)

Proof.

1. **Pipeline decomposition.**
   The self-attention pipeline is $S = \frac{1}{\sqrt{d}}QK^\top$, $P = \mathrm{softmax}(S)$ (row-wise), and $A = PV$. First-order floating-point (FP) rounding is injected at each stage and then propagated; we bound each stage and compose.

2. **Stage 1 — score computation (GEMM + scaling).**
   Let $\widetilde{S} = \mathrm{fl}(QK^\top/\sqrt{d})$ and write $\Delta S := \widetilde{S} - S$. By Lemma 1,

$$\|\Delta S\| \le \epsilon_{\mathrm{mach}} \left(\|S\| + c_{\mathrm{gemm}} \frac{\|Q\|\,\|K\|}{\sqrt{d}} \right) + O(\epsilon_{\mathrm{mach}}^2).$$

3. **Stage 2 — softmax (Fréchet sensitivity + in-kernel rounding).**
Let $\mathcal{S}$ denote the row-wise softmax map. Since $\mathcal{S}$ is smooth,

$$\widetilde{P} = \mathcal{S}(S + \Delta S) + E_{\text{smx}} = P + D\mathcal{S}_S[\Delta S] + E_{\text{smx}} + O(\|\Delta S\|^2),$$

where E_{smx} is the kernel-rounding term from Lemma 3 and satisfies

$$\frac{\|E_{\text{smx}}\|}{\|P\|} \le c_{\text{smx}}\epsilon_{\text{mach}} + O(\epsilon_{\text{mach}}^2).$$

Because $D\mathcal{S}_S$ is block diagonal across rows,

$$\|D\mathcal{S}_S\|_{F\to F} = \max_{1\le i\le n} \|J(P_{i:})\|_2.$$

By Lemma 4, this quantity is always finite and at most $\frac{1}{2}$. Therefore

$$\begin{aligned}\frac{\|\widetilde{P} - P\|}{\|P\|} &\le \frac{\|D\mathcal{S}_S\|_{F\to F}\,\|\Delta S\|}{\|P\|} + c_{\text{smx}}\epsilon_{\text{mach}} + O(\epsilon_{\text{mach}}^2)\\ &\le \Big[c_{\text{smx}} + \kappa_{\text{softmax}} + c_{\text{gemm}}\chi_{\text{score}}\Big]\epsilon_{\text{mach}} + O(\epsilon_{\text{mach}}^2).\end{aligned}$$

4. **Stage 3 — value projection (GEMM).**
Let $\widehat{A} := \widetilde{P}V$. By Lemma 2, the finite-precision GEMM output satisfies

$$\widetilde{A} = \widehat{A} + E_A = \widetilde{P}V + E_A,$$

with

$$\|E_A\| \le c'_{\text{gemm}}\epsilon_{\text{mach}}\|\widetilde{P}\|\,\|V\|_2 + O(\epsilon_{\text{mach}}^2).$$

Therefore

$$\widetilde{A} - A = (\widetilde{P} - P)V + E_A.$$

By Stage 2, $\|\widetilde{P} - P\|/\|P\| = O(\epsilon_{\text{mach}})$, hence

$$\|\widetilde{P}\| \le \|P\| + \|\widetilde{P} - P\| = \|P\| + O(\epsilon_{\text{mach}})\|P\|.$$

Substituting this into the bound for E_A gives

$$\|E_A\| \le c'_{\text{gemm}}\epsilon_{\text{mach}}\|P\|\,\|V\|_2 + O(\epsilon_{\text{mach}}^2).$$

Hence

$$\|\widetilde{A} - A\| \le \Big(\|\widetilde{P} - P\| + c'_{\text{gemm}}\epsilon_{\text{mach}}\|P\|\Big)\|V\|_2 + O(\epsilon_{\text{mach}}^2).$$

Dividing by $\|A\| = \|PV\| > 0$ and multiplying/dividing the first term by $\|P\|$ yields

$$\begin{aligned}\frac{\|\widetilde{A} - A\|}{\|A\|} &\le \left(\frac{\|\widetilde{P} - P\|}{\|P\|} + c'_{\text{gemm}}\epsilon_{\text{mach}}\right)\frac{\|P\|\,\|V\|_2}{\|A\|}\\ &+O(\epsilon_{\text{mach}}^2) = \left(\frac{\|\widetilde{P} - P\|}{\|P\|} + c'_{\text{gemm}}\epsilon_{\text{mach}}\right)\kappa_{\text{val}}\\ &+O(\epsilon_{\text{mach}}^2).\end{aligned}$$

5. **Final combination.**
Substitute the Stage 2 bound into Stage 3:

$$\begin{aligned}\|\widetilde{A} - A\| \;\le\; &\Big[c_{\text{smx}} + \kappa_{\text{softmax}} + c_{\text{gemm}}\chi_{\text{score}}\\ &+ c'_{\text{gemm}}\Big]\epsilon_{\text{mach}}\;\|P\|\,\|V\|_2 + O(\epsilon_{\text{mach}}^2).\end{aligned}$$

If additionally $\|A\| > 0$, dividing by $\|A\|$ gives

$$\frac{\|\widetilde{A} - A\|}{\|A\|} \leq \Big[c_{\mathrm{smx}} + \kappa_{\mathrm{softmax}} + c_{\mathrm{gemm}}\chi_{\mathrm{score}} + c'_{\mathrm{gemm}}\Big]\epsilon_{\mathrm{mach}}\ \kappa_{\mathrm{val}} + O(\epsilon_{\mathrm{mach}}^2).$$

This is exactly the main-text statement of Theorem 1.

□

A.3 Proof of Theorem 2 and Corollary 1

Proof.

(1) **Theorem 2.** Let $F := J_f$ with $\|F\|_2 < 1$ and $T := I + F$.

(a) *Forward norm.* By subadditivity, $\|T\|_2 = \|I + F\|_2 \leq \|I\|_2 + \|F\|_2 = 1 + \|F\|_2$.

(b) *Inverse norm.* Since $\|F\|_2 < 1$, T is invertible and $(I + F)^{-1} = \sum_{k=0}^{\infty}(-F)^k$; thus $\|T^{-1}\|_2 \leq \sum_{k=0}^{\infty}\|F\|_2^k = (1 - \|F\|_2)^{-1}$.

(c) *Combine.* Therefore

$$\kappa(T) = \|T\|_2\,\|T^{-1}\|_2 \ \leq\ \frac{1 + \|F\|_2}{1 - \|F\|_2}.$$

(2) **Corollary 1 (depth-wise relaxation).** Consider a stack of residual blocks with Jacobians $\{F_\ell := J_{f_\ell}\}_{\ell=1}^{L}$ and $\rho_\ell := \|F_\ell\|_2 < 1$. Let $T_\ell := I + F_\ell$. For the linearized composition $T := \prod_{\ell=1}^{L} T_\ell$ we have

(a) *Submultiplicativity of κ.* For any compatible A, B, $\kappa(AB) \leq \kappa(A)\,\kappa(B)$ since $\|AB\|_2 \leq \|A\|_2\|B\|_2$ and $\|(AB)^{-1}\|_2 = \|B^{-1}A^{-1}\|_2 \leq \|B^{-1}\|_2\|A^{-1}\|_2$.

(b) *Product bound.* Applying (a) iteratively and Theorem 2 to each T_ℓ,

$$\kappa\Big(\prod_{\ell=1}^{L} T_\ell\Big) \ \leq\ \prod_{\ell=1}^{L}\kappa(T_\ell) \ \leq\ \prod_{\ell=1}^{L}\frac{1 + \rho_\ell}{1 - \rho_\ell}.$$

(c) *First-order relaxation (used in main text).* Since $\frac{1+\rho}{1-\rho} = 1 + 2\rho + O(\rho^2)$ for $\rho \in [0, 1)$, a first-order *relaxation model* replaces $\frac{1+\rho_\ell}{1-\rho_\ell}$ by $(1 + \rho_\ell)$. This captures the attenuation effect of residuals while avoiding an overly pessimistic multiplicative growth; note this replacement is an approximation (not an upper bound).

□

A.4 Proof of Proposition 1 (Normalization-path forward error and monotonicity)

Proof.

1. **Setup and notation.** Let $d = d_{\mathrm{model}}$, $m = \mu(x) = \frac{1}{d}\mathbf{1}^\top x$, $v = \sigma^2(x) = \frac{1}{d}\|x - m\mathbf{1}\|_2^2$, $C := I - \frac{1}{d}\mathbf{1}\mathbf{1}^\top$, and

$$c := x - m\mathbf{1} = Cx, \qquad \alpha := (v + \varepsilon)^{-1/2}, \qquad g(x) := \frac{x - m\mathbf{1}}{\sqrt{v + \varepsilon}} = \alpha c, \qquad Z_\varepsilon^{\mathrm{LN}}(x) := \mathrm{Diag}(\gamma)\,g(x).$$

The full LayerNorm output is $\mathrm{LN}(x) = Z_\varepsilon^{\mathrm{LN}}(x) + \beta$, but Proposition 1 isolates the ε-dependent normalization path $Z_\varepsilon^{\mathrm{LN}}$. The final bias addition is ε-independent and is absorbed into the effective remainder term in Theorem 3.

2. **Centering stage.** Let $\widetilde{m} = \mathrm{fl}(\frac{1}{d}\mathbf{1}^\top x) = m + \delta_m$. Standard first-order reduction bounds give

$$|\delta_m| \le c_m \epsilon_{\mathrm{mach}} \frac{\|x\|_2}{\sqrt{d}} + O(\epsilon_{\mathrm{mach}}^2)$$

for a kernel- and dimension-dependent constant $c_m > 0$. Let

$$\widetilde{c} := \mathrm{fl}(x - \widetilde{m}\mathbf{1}) = c + e_c.$$

The subtraction stage satisfies

$$e_c = -\delta_m \mathbf{1} + e_{\mathrm{sub}},$$

with

$$\|e_{\mathrm{sub}}\|_2 \le c_{\mathrm{sub}} \epsilon_{\mathrm{mach}} \big(\|x\|_2 + \sqrt{d}\,|\widetilde{m}|\big) + O(\epsilon_{\mathrm{mach}}^2).$$

Since $|m| \le \|x\|_2/\sqrt{d}$ and $|\widetilde{m}| = |m| + O(\epsilon_{\mathrm{mach}})\|x\|_2/\sqrt{d}$, there exists $c_c > 0$ such that

$$\|e_c\|_2 \le c_c \epsilon_{\mathrm{mach}} \|x\|_2 + O(\epsilon_{\mathrm{mach}}^2).$$

3. **Variance and reciprocal-square-root stage.** Let

$$\widetilde{v} := \mathrm{fl}\left(\frac{1}{d}\|\widetilde{c}\|_2^2\right) = v + \delta_v.$$

Decompose

$$\delta_v = \frac{\|\widetilde{c}\|_2^2 - \|c\|_2^2}{d} + e_v,$$

where e_v collects first-order reduction/multiplication rounding in the squared-norm computation. Using $\widetilde{c} = c + e_c$ and $\|c\|_2 = \sqrt{d}\,\sigma(x)$,

$$\left|\frac{\|\widetilde{c}\|_2^2 - \|c\|_2^2}{d}\right| \le \frac{2}{d}\|c\|_2\|e_c\|_2 + O(\epsilon_{\mathrm{mach}}^2) \le \frac{2c_c}{\sqrt{d}}\epsilon_{\mathrm{mach}}\sigma(x)\|x\|_2 + O(\epsilon_{\mathrm{mach}}^2).$$

Standard first-order reduction bounds also give

$$|e_v| \le c_v \epsilon_{\mathrm{mach}}\, v + O(\epsilon_{\mathrm{mach}}^2)$$

for some $c_v > 0$. Therefore there exist kernel- and dimension-dependent constants $\bar{a}_v, \bar{b}_v > 0$ such that

$$|\delta_v| \le \epsilon_{\mathrm{mach}} \big(\bar{a}_v\, \sigma(x)\|x\|_2 + \bar{b}_v\, v\big) + O(\epsilon_{\mathrm{mach}}^2).$$

Let

$$\widetilde{\alpha} := \mathrm{fl}\big((\widetilde{v} + \varepsilon)^{-1/2}\big) = \alpha + \delta_\alpha.$$

By the mean value theorem for $t \mapsto t^{-1/2}$ and first-order reciprocal-square-root rounding,

$$|\delta_\alpha| \le \frac{|\delta_v|}{2(v+\varepsilon)^{3/2}} + c_{\mathrm{rsqrt}}\epsilon_{\mathrm{mach}}\frac{1}{\sqrt{v+\varepsilon}} + O(\epsilon_{\mathrm{mach}}^2).$$

4. **Normalization core.** The computed normalized core is

$$\widetilde{g}(x) = \mathrm{fl}(\widetilde{\alpha}\widetilde{c}) = \alpha c + \alpha e_c + \delta_\alpha c + e_{\mathrm{mul}} + O(\epsilon_{\mathrm{mach}}^2),$$

where pointwise multiplication gives

$$\|e_{\mathrm{mul}}\|_2 \le c_{\mathrm{mul}}\epsilon_{\mathrm{mach}}\frac{\|x\|_2}{\sqrt{v+\varepsilon}} + O(\epsilon_{\mathrm{mach}}^2).$$

Hence

$$\|\widetilde{g}(x) - g(x)\|_2 \le \frac{\|e_c\|_2}{\sqrt{v+\varepsilon}} + \|c\|_2\,|\delta_\alpha| + c_{\mathrm{mul}}\epsilon_{\mathrm{mach}}\frac{\|x\|_2}{\sqrt{v+\varepsilon}} + O(\epsilon_{\mathrm{mach}}^2).$$

Substituting the bounds above yields a linear combination of

$$\frac{\|x\|_2}{\sqrt{v+\varepsilon}}, \qquad \frac{v\,\|x\|_2}{(v+\varepsilon)^{3/2}},$$

with kernel- and dimension-dependent coefficients. Here we used

$$\|c\|_2 = \sqrt{d}\,\sigma(x) \le \|x\|_2, \qquad \|c\|_2\sigma(x) = \sqrt{d}\,v,$$

so the δ_v terms are also controlled by the second quantity above after absorbing dimension-only factors into the constants. Since

$$\frac{1}{\sqrt{v+\varepsilon}} \le \frac{\varepsilon+2v}{(v+\varepsilon)^{3/2}}, \qquad \frac{v}{(v+\varepsilon)^{3/2}} \le \frac{\varepsilon+2v}{(v+\varepsilon)^{3/2}},$$

there exists a kernel- and dimension-dependent constant $\bar{a}_{\ln} > 0$ such that

$$\|\widetilde{g}(x) - g(x)\|_2 \le \epsilon_{\text{mach}}\,\bar{a}_{\ln}\frac{\varepsilon+2v}{(v+\varepsilon)^{3/2}}\|x\|_2 + O(\epsilon_{\text{mach}}^2).$$

5. **Finite-precision scaling by γ.** Let $\widetilde{Z}_\varepsilon^{\text{LN}}(x)$ denote the finite-precision result of the pointwise multiplication by γ. Standard pointwise multiplication bounds give

$$\widetilde{Z}_\varepsilon^{\text{LN}}(x) - Z_\varepsilon^{\text{LN}}(x) = \text{Diag}(\gamma)\big(\widetilde{g}(x) - g(x)\big) + e_\gamma + O(\epsilon_{\text{mach}}^2),$$

where

$$\|e_\gamma\|_2 \le c_\gamma\,\epsilon_{\text{mach}}\,\|\text{Diag}(\gamma)\|_2\,\|g(x)\|_2 + O(\epsilon_{\text{mach}}^2)$$

for a kernel-dependent constant $c_\gamma > 0$. Since

$$\|g(x)\|_2 = \frac{\|c\|_2}{\sqrt{v+\varepsilon}} \le \frac{\varepsilon+2v}{(v+\varepsilon)^{3/2}}\|x\|_2,$$

so

$$\|\widetilde{Z}_\varepsilon^{\text{LN}}(x) - Z_\varepsilon^{\text{LN}}(x)\|_2 \le \epsilon_{\text{mach}}\,\|\text{Diag}(\gamma)\|_2\,\frac{\varepsilon+2v}{(v+\varepsilon)^{3/2}}\big(\bar{a}_{\ln} + c_\gamma\big)\|x\|_2 + O(\epsilon_{\text{mach}}^2).$$

Define

$$a_{\ln} := \bar{a}_{\ln} + c_\gamma.$$

6. **Absolute forward error bound.** With $a_{\ln} := \bar{a}_{\ln} + c_\gamma$, Step 5 gives

$$\|\widetilde{Z}_\varepsilon^{\text{LN}}(x) - Z_\varepsilon^{\text{LN}}(x)\| \le \epsilon_{\text{mach}}\|\text{Diag}(\gamma)\|_2\frac{\varepsilon+2\sigma^2(x)}{(\sigma^2(x)+\varepsilon)^{3/2}}\big(a_{\ln}\|x\|_2\big) + O(\epsilon_{\text{mach}}^2),$$

which is exactly (3) with

$$M_{\text{LN}}(x,\varepsilon) := \|\text{Diag}(\gamma)\|_2\frac{\varepsilon+2\sigma^2(x)}{(\sigma^2(x)+\varepsilon)^{3/2}}\big(a_{\ln}\|x\|_2\big).$$

7. **Relative forward error bound.** If additionally $\|Z_\varepsilon^{\text{LN}}(x)\| > 0$, dividing the previous inequality by $\|Z_\varepsilon^{\text{LN}}(x)\|$ gives

$$\frac{\|\widetilde{Z}_\varepsilon^{\text{LN}}(x) - Z_\varepsilon^{\text{LN}}(x)\|}{\|Z_\varepsilon^{\text{LN}}(x)\|} \le \epsilon_{\text{mach}}\frac{M_{\text{LN}}(x,\varepsilon)}{\|Z_\varepsilon^{\text{LN}}(x)\|} + O(\epsilon_{\text{mach}}^2),$$

which is exactly the stated coefficient $C_{\text{LN}}(x,\varepsilon)$.

8. **Monotonicity of the ε-dependent factor.** For fixed $v = \sigma^2(x)$, define

$$f_v(\varepsilon) := \frac{\varepsilon+2v}{(v+\varepsilon)^{3/2}}.$$

Then

$$\frac{d}{d\varepsilon}f_v(\varepsilon) = -\frac{\varepsilon+4v}{2(v+\varepsilon)^{5/2}} < 0 \qquad (\varepsilon > 0),$$

so f_v is strictly decreasing in ε. This proves the monotonicity claim in Proposition 1.

□

A.5 Proof of Proposition 2 (Frozen-scale monotonicity)

Proof.

1. **Derivative of the frozen-scale LayerNorm magnitude.**
   Write
   $$\alpha_{\ell,t} := \|\mathrm{Diag}(\gamma_\ell)\|_2\big(a_{\ell,t}\,\|x_{\ell,t}\|_2\big),$$
   so that
   $$M^{\mathrm{frz}}_{\mathrm{LN},\ell,t}(\varepsilon) = \alpha_{\ell,t}\frac{\varepsilon + 2v_{\ell,t}}{(v_{\ell,t}+\varepsilon)^{3/2}}.$$
   Because $a_{\ell,t} > 0$, $\|\mathrm{Diag}(\gamma_\ell)\|_2 \geq 0$, and $\|x_{\ell,t}\|_2 \geq 0$, we have $\alpha_{\ell,t} \geq 0$. Differentiating gives
   $$\frac{d}{d\varepsilon}M^{\mathrm{frz}}_{\mathrm{LN},\ell,t}(\varepsilon) = -\frac{\alpha_{\ell,t}}{2}\frac{\varepsilon + 4v_{\ell,t}}{(v_{\ell,t}+\varepsilon)^{5/2}},$$
   which is exactly (17). Since $\alpha_{\ell,t} \geq 0$, $v_{\ell,t} \geq 0$, and $\varepsilon > 0$, the derivative is nonpositive, so $M^{\mathrm{frz}}_{\mathrm{LN},\ell,t}$ is nonincreasing. If $\|\mathrm{Diag}(\gamma_\ell)\|_2\,\|x_{\ell,t}\|_2 > 0$, then $\alpha_{\ell,t} > 0$, hence the derivative is strictly negative for every $\varepsilon > 0$, and $M^{\mathrm{frz}}_{\mathrm{LN},\ell,t}$ is strictly decreasing.

2. **Monotonicity of the frozen-scale layer contribution.**
   Define the ε-independent factor
   $$D_{\ell,t} := \frac{1}{\|X_{L,t}\|}\prod_{k=\ell+1}^{L}(1+\rho_{k,t}).$$
   Since $\|X_{L,t}\| > 0$ and $\rho_{k,t} = \|J_{f_k}(X_{k-1,t})\|_2 \geq 0$, we have $D_{\ell,t} \geq 0$. Therefore
   $$G^{\mathrm{frz}}_{\ell,t}(\varepsilon) = \Big(M^{\mathrm{eff}}_{\ell,t} + A_{\ell,t} + M^{\mathrm{frz}}_{\mathrm{LN},\ell,t}(\varepsilon)\Big)D_{\ell,t}.$$
   Because all remaining factors are fixed with respect to ε, the nonincreasingness of $M^{\mathrm{frz}}_{\mathrm{LN},\ell,t}$ implies that $G^{\mathrm{frz}}_{\ell,t}$ is nonincreasing in ε.

3. **Exact reduction formula.**
   For any $\varepsilon' \geq \varepsilon$, subtracting the two frozen-scale scores yields
   $$\begin{aligned} G^{\mathrm{frz}}_{\ell,t}(\varepsilon) - G^{\mathrm{frz}}_{\ell,t}(\varepsilon') = \frac{1}{\|X_{L,t}\|}\Big(M^{\mathrm{frz}}_{\mathrm{LN},\ell,t}(\varepsilon) - M^{\mathrm{frz}}_{\mathrm{LN},\ell,t}(\varepsilon')\Big) \\ \times \prod_{k=\ell+1}^{L}(1+\rho_{k,t}), \end{aligned}$$
   because the $M^{\mathrm{eff}}_{\ell,t}$ and $A_{\ell,t}$ terms cancel exactly. This is (19).

□

Appendix: ε-bump policy (as implemented)

Algorithm 2: LayerNorm ε-bump triggered by the practical regime indicator

Input: target $\rho_* \in (0,1)$, check interval t=5, floor/cap ($\varepsilon_{\min}$=10^{-6}, $\varepsilon_{\max} \in \{5\times10^{-3}, 10^{-2}\}$)
for *training step* $s = 1, 2, \ldots$ **do**
  if $s \bmod t = 0$ **then**
    sample a 16-example minibatch; **foreach** *LN layer with current* ε **do**
      compute $\rho_{\mathrm{LN}}(\varepsilon) = \frac{\sigma^2(x)}{\varepsilon}\, d_{\mathrm{model}}\, \epsilon_{\mathrm{mach}}$
      if $\rho_{\mathrm{LN}}(\varepsilon) < 1$ **then**
        $\varepsilon_{\mathrm{cand}} \leftarrow \dfrac{\mathrm{median}(\sigma^2(x))\, d_{\mathrm{model}}\, \epsilon_{\mathrm{mach}}}{\rho_*}$
        $\varepsilon_{\mathrm{new}} \leftarrow \mathrm{clip}(\varepsilon_{\mathrm{cand}}, \varepsilon_{\min}, \varepsilon_{\max})$
        if $\varepsilon_{\mathrm{new}} > \varepsilon$ **then**
          set $\varepsilon \leftarrow \varepsilon_{\mathrm{new}}$ `// monotone increase`

A.6 Proof of Theorem 3 (Unified Forward Stability)

Proof.

1. **Linearization and one-step recursion.**
   Let X_ℓ be the exact hidden after layer ℓ, with X_0 the model input, and let $\widetilde{X}_\ell$ be its FP counterpart, and $E_\ell = \widetilde{X}_\ell - X_\ell$. For the pre-LN residual block $x \mapsto x + f_\ell(x)$, write $F_\ell := J_{f_\ell}(X_{\ell-1})$ and $\rho_\ell = \|F_\ell\|_2$. With $E_0 = 0$, a first-order perturbation analysis gives, for each layer $\ell = 1, \ldots, L$,
   $$E_\ell = (I + F_\ell)E_{\ell-1} + \Delta_\ell + O(\epsilon_{\mathrm{mach}}^2),$$
   where Δ_ℓ is the layer-ℓ local first-order forward error (from MHSA, LN, FFN).

2. **Exact local layer coefficient.**
   Decompose the layer-local first-order error as
   $$\Delta_\ell = \Delta_\ell^{\mathrm{eff}} + \Delta_\ell^{\mathrm{attn}} + \Delta_\ell^{\mathrm{LN}} + O(\epsilon_{\mathrm{mach}}^2),$$
   where $\Delta_\ell^{\mathrm{eff}}$ collects the remaining non-attention and non-normalization-path first-order perturbations. By (8) and Lemma 5,
   $$\|\Delta_\ell^{\mathrm{eff}}\| \le \epsilon_{\mathrm{mach}}\, M_\ell^{\mathrm{eff}} + O(\epsilon_{\mathrm{mach}}^2).$$
   Let
   $$M_\ell^{\mathrm{LN}} := M_{\mathrm{LN}}(x_\ell, \varepsilon_\ell)$$
   denote the exact absolute LayerNorm normalization-path magnitude from Proposition 1. By Theorem 1,
   $$\|\Delta_\ell^{\mathrm{attn}}\| \le \epsilon_{\mathrm{mach}}\, A_\ell + O(\epsilon_{\mathrm{mach}}^2),$$
   and by Proposition 1,
   $$\|\Delta_\ell^{\mathrm{LN}}\| \le \epsilon_{\mathrm{mach}}\, M_\ell^{\mathrm{LN}} + O(\epsilon_{\mathrm{mach}}^2).$$
   Therefore triangle inequality gives
   $$\|\Delta_\ell\| \le \epsilon_{\mathrm{mach}} \left[M_\ell^{\mathrm{eff}} + A_\ell + M_\ell^{\mathrm{LN}} \right] + O(\epsilon_{\mathrm{mach}}^2)$$
   $$= \epsilon_{\mathrm{mach}} M_\ell + O(\epsilon_{\mathrm{mach}}^2).$$

3. **Residual relaxation (downstream amplification).**
   Iterating the recursion from Step 1 and using $\|I + F_k\|_2 \le 1 + \rho_k$ yields
   $$\|E_L\| \;\le\; \sum_{\ell=1}^{L} \Big(\|\Delta_\ell\| \prod_{k=\ell+1}^{L} \|I + F_k\|_2 \Big) \;+\; O(\epsilon_{\mathrm{mach}}^2)\, \|X_L\|$$

$$\leq \sum_{\ell=1}^{L} \Big(\|\Delta_\ell\| \prod_{k=\ell+1}^{L} (1+\rho_k) \Big) \; + \; O(\epsilon_{\text{mach}}^2)\, \|X_L\|.$$

4. **Normalization and collection of terms.**
   By Step 2, $\|\Delta_\ell\| \leq \epsilon_{\text{mach}} M_\ell + O(\epsilon_{\text{mach}}^2)$. Any benign inter-layer norm ratios used to rewrite the local absolute magnitudes relative to $\|X_L\|$ are absorbed into the same layer-dependent first-order constants already defining M_ℓ. Substitute this into Step 3 and divide by $\|X_L\|$:

$$\frac{\|E_L\|}{\|X_L\|} \leq \epsilon_{\text{mach}} \sum_{\ell=1}^{L} \frac{M_\ell}{\|X_L\|} \prod_{k=\ell+1}^{L} (1+\rho_k) + O(\epsilon_{\text{mach}}^2),$$

   which is exactly (10).

5. **Remainder.**
   The $O(\epsilon_{\text{mach}}^2)$ term collects mixed higher-order interactions across kernels and layers.

□

A.7 Estimation Details and Complexity Notes

Softmax sensitivity. Estimate $\|D\mathcal{S}_S\|_{F\to F} = \max_i \|J(P_{i:})\|_2$ by 2–3 steps of power iteration on sampled rows, then multiply by $\|S\|/\|P\|$ to obtain κ_{softmax}. For multi-head, aggregate by max (or sum for a conservative predictor).

Score transport and value conditioning. Estimate χ_{score} from the same sampled softmax differential together with the factor $\|Q\|\,\|K\|/(\|P\|\sqrt{d})$. When $\|S\| > 0$, one may equivalently form the auxiliary diagnostic κ_{score} and use $\chi_{\text{score}} = \kappa_{\text{softmax}}\kappa_{\text{score}}$. For the value path in the layer-wise score, estimate the absolute attention magnitude factor $\|P\|\,\|V\|_2$ using the monitored tensors $\widehat{P}$ and $\widehat{V}$, for example via direct norm computation or a truncated-SVD estimate of $\|\widehat{V}\|_2$. When $\|A\| > 0$, one may additionally record the relative diagnostic $\kappa_{\text{val}} = \|P\|\,\|V\|_2/\|A\|$.

Residual transport. Estimate each downstream factor $\widehat{\rho}_{k,t}$ causally from the monitored pass. Two practical options are: (i) 1–2 power-iteration steps using JVP/VJP access to estimate $\|J_{f_k}(X_{k-1,t})\|_2$, or (ii) a conservative blockwise operator-norm surrogate built from the constituent linear maps and pointwise nonlinearities inside the residual branch. Either choice preserves the role of the downstream transport factor in $\widehat{G}_{\ell,t}$.

Effective remainder magnitude. Construct $\widehat{M}_{\ell,t}^{\text{eff}}$ by aggregating the remaining non-attention, non-normalization-path operations in layer ℓ, such as output projections, FFN maps, and the ε-independent LayerNorm tail. In practice, we use causal norm-based surrogates of their first-order output magnitudes and sum them so that $\widehat{M}_{\ell,t}^{\text{eff}}$ preserves the additive local decomposition of the exact magnitude $M_{\ell,t}^{\text{eff}}$.

Kernel constants. $c_{\text{gemm}}, c'_{\text{gemm}}, c_{\text{smx}}, c_{\text{ln}}$ are treated as fixed per hardware/runtime; they factor into the absolute constants in Theorems 1 and 3.

Remark 1 (Dropout). With inverted dropout $\text{Drop}_p(z) = (M/p) \odot z$, $\|J_{\text{Drop}_p \circ f}(x)\|_2 \leq (1/p)\|J_f(x)\|_2$; hence per-layer factors in our bounds scale by at most $1/p$. Post-softmax attention dropout multiplies both J and P by $(1/p)$, leaving κ_{softmax} non-increasing under a fixed mask.

Lemma 6 (Attention dropout does not worsen κ_{softmax})**.** *If post-softmax dropout is applied with a fixed mask,* $\kappa'_{\text{softmax}} \leq \kappa_{\text{softmax}}$. Sketch. $J(P') = \text{Diag}(M/p)J(P)$ *and* $\|P'\| = \|(M/p) \odot P\|$ *so the* $(1/p)$ *factors cancel in the ratio.*

A.8 Additional Experimental Details

Protocols and metrics. E1 consists of synthetic controlled sweeps with 20 attention configurations, 10 LayerNorm ε values, and 8 residual-gain values. E2 uses the GPT-2 checkpoint (Radford et al., 2019) on WikiText-103 (Merity et al., 2016) validation with BF16/FP16 monitored passes, sequence lengths $\{128, 512, 1024\}$, and

3 seeds, for 18 completed runs and 96 monitored windows per run. E3 reuses these E2 runs, selects the top-8 highest-mismatch windows per run, and evaluates single-layer FP32 reference-patch effects against the same manual forward path; the mean recompute gap is 9.24×10^{-7}. E5 uses stable FP32-master / FP16-shadow training on WikiText-2 train with sequence length 256, 256 monitored steps, 3 seeds, and a maximum of 24 controller actions for the budget-matched policies.

Implementation, infrastructure, and assets. The released code is written for Python 3.10+ and uses PyTorch, Hugging Face `transformers`, Hugging Face `datasets`, `matplotlib`, and a Bash-compatible orchestration script. The public repository includes default configuration files specifying the model, dataset splits, precisions, sequence lengths, seeds, monitoring parameters, and BGSS controller hyperparameters used for the reported experiments. Experiments were run on NVIDIA H100 accelerators. The external assets are the Hugging Face `gpt2` checkpoint and the WikiText dataset accessed through Hugging Face `datasets`. The `gpt2` checkpoint is distributed under the MIT license, and WikiText is distributed under the Creative Commons Attribution-ShareAlike license (CC BY-SA 4.0). The official code release is provided at the URL in Section 1 and is released under the MIT license.

Secondary quantitative summaries. For E2, the transport-aware combined predictor improves Pearson correlation over the no-transport ablation by a mean of 0.163 and Spearman correlation by a mean of 0.139. For E3, the mean top-1 hit rate is 0.264, with mean top-3 and top-5 overlaps of 0.505 and 0.622, respectively. For E5, BGSS and the budget-matched baselines use identical average protected layer-steps (2742). Relative to the random same-budget controller, BGSS reduces mean onset events by 1.0 and reduces mean max mismatch by more than a factor of 2.7; relative to the risk-only same-budget controller, BGSS preserves the same mean onset event count while reducing mean max mismatch from 5.71×10^{-3} to 3.14×10^{-3}.

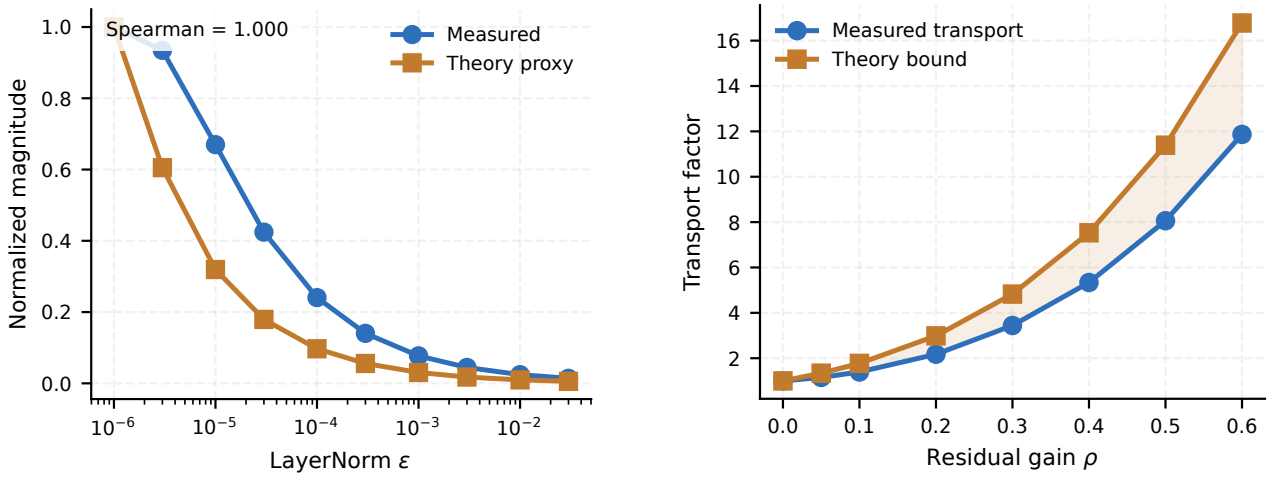

Figure 5: Additional E1 controlled sweeps. Left: the measured LayerNorm normalization-path change and the causal proxy both decrease monotonically as ε grows, matching Proposition 1. Right: the measured downstream amplification remains below the residual transport bound throughout the tested ρ-range, matching Theorem 2 and Corollary 1.

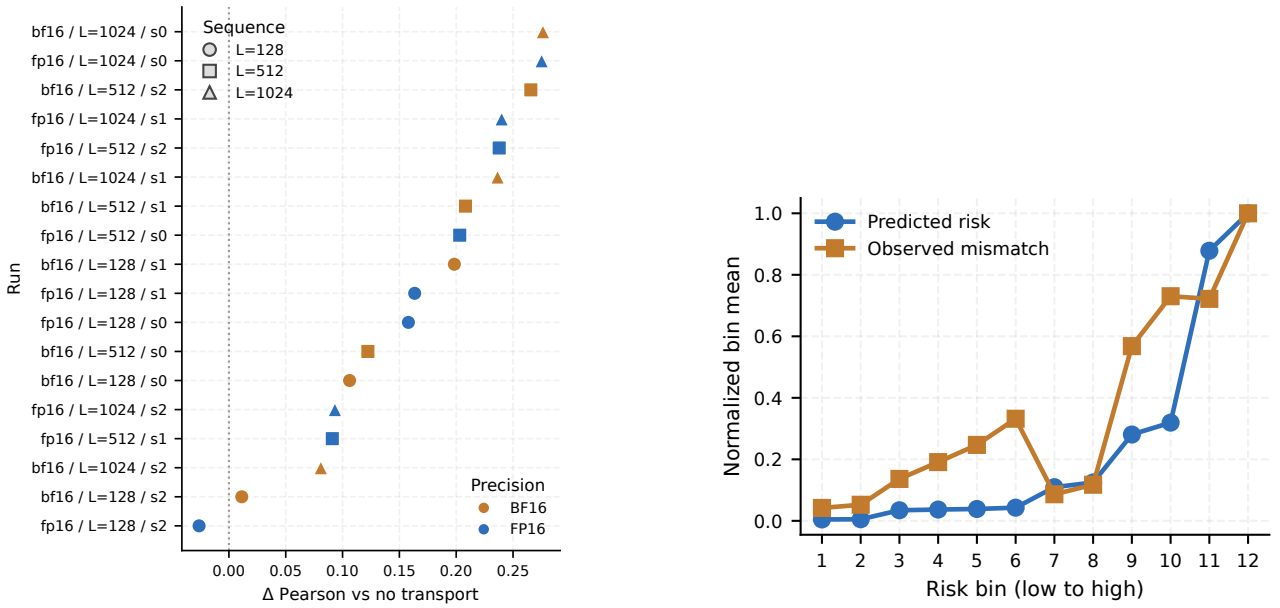

Figure 6: Additional E2 views. Left: run-wise transport gain, showing that most runs improve in Pearson correlation when downstream transport is added to the local predictor. Right: a binned trend view in which windows sorted by the predicted risk exhibit increasing mismatch in the high-risk bins.

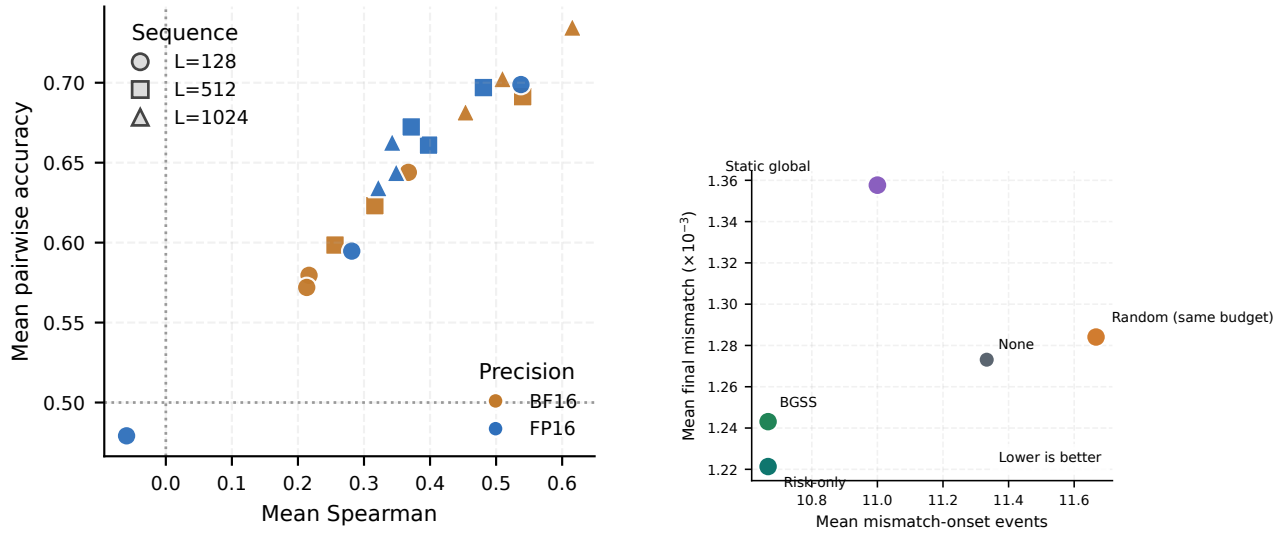

Figure 7: Additional E3 and E5 summaries. Left: run-level E3 fidelity, showing that the attribution proxy remains in the positive-fidelity regime across runs. Right: a complementary E5 tradeoff view, showing that BGSS and the risk-only same-budget controller attain similar mean onset-event counts, while BGSS improves over the random same-budget controller on both onset events and mean final mismatch.