Contractive Restoring Flows: Robust Reasoning Distillation via Orbital Stability

Dongqi Zuo[1] **Yuanyuan Wang**[2] **Chuan Zhou**[3] **Mingming Gong**[3,2] **Haoxuan Li**[1,*]
[1]Institute for Artificial Intelligence, Peking University
[2]Department of Machine Learning, MBZUAI
[3]School of Mathematics and Statistics, The University of Melbourne
dongqizuo@gmail.com, hxli@pku.edu.cn

Abstract

Model distillation provides an alternative when computing resources are limited. How to transfer the reasoning ability from teacher model to student model influences the performance of model distillation. We develop a mathematical framework for reasoning distillation grounded in dynamical systems theory. By modeling the teacher model's residual stream as a discrete dynamical system, we define a precise notion of *orbital stability*: perturbations transversal to the teacher's reasoning trajectory should contract, while the tangential reasoning signal propagates without active suppression. We derive the **Contractive Restoring Flow (CRF)** loss from first principles and prove that, at the loss minimum, it achieves strict transversal contraction at a constant rate $1 - \alpha$, leaves the tangential dynamics unconstrained (tangential agnosticism), and produces a global basin of attraction with bounded steady-state error within the linearisation regime. Empirically, reasoning-trajectory distillation with the CRF loss outperforms other distillation methods, and shows greater robustness on long-chain reasoning tasks, supporting the predicted effect of contracting transversal error. The code is available at: https://github.com/DongqiZuo00/Contractive-Restoring-Flows.

1 Introduction

Distilling the reasoning ability of a large teacher model into a compact student is a central challenge in efficient LLM deployment. Recent work has moved beyond token-level imitation toward *trajectory cloning*: supervising the student's hidden states to match the teacher's layer by layer [Liang et al., 2023, Dasgupta and Cohn, 2025]. This provides a far richer training signal, but it rests on a fundamental structural weakness.

The core problem. Hidden-state matching is a *zeroth-order* constraint. It controls the student's residual function $f_\theta^{(l)}$ at the single point h_l^* on the teacher's trajectory, but places no constraint on the *Jacobian*, on how the student responds to inputs that lie *away* from that point. At inference, autoregressive error accumulation, approximation error, and distribution shift all push the student off the teacher's trajectory; once a deviation occurs, its subsequent evolution is governed entirely by the student's own layer-wise Jacobians, which trajectory matching has left unconstrained. Whether these Jacobians amplify or attenuate the deviation determines whether reasoning remains coherent across a long chain, and existing methods provide no guarantee either way. Standard hidden-state matching constrains representations along the observed teacher trajectory but does not directly constrain the transverse Jacobian. It therefore provides no guarantee that deviations from that trajectory contract.

*Haoxuan Li is the corresponding author.

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

Table 1: **Comparison of representative distillation methods.** Method properties and support for long-chain reasoning.

Method	Method properties				Target
	First-order (Jacobian)	Per-layer scope	Transversal proj. $P_\perp$	Restoring force $\|P_\perp J\|_2<1$	Long-chain reasoning
Output KD [Hinton et al., 2015]	✗	✗	✗	✗	✗
LLM distribution KD [Agarwal et al., 2024, Yan et al., 2026]	✗	✗	✗	✗	✓
Hidden-state matching [Romero et al., 2015, Liang et al., 2023]	✗	✓	✗	✗	✓
Contractive Auto-Encoder [Rifai et al., 2011]	✓	✓	✗	✗	✗
Jacobian matching [Srinivas and Fleuret, 2018]	✓	✗	✗	✗	✗
CRF (ours)	✓	✓	✓	✓	✓

Two traditions, neither sufficient. LLM distillation has stayed within zeroth-order matching: from output KD [Hinton et al., 2015] to logit and distribution methods [Agarwal et al., 2024, Yan et al., 2026] to hidden-state matching [Liang et al., 2023, Dasgupta and Cohn, 2025], the supervision target has moved deeper into the network but never to the Jacobian. Jacobian-based distillation is in fact older than LLMs: Rifai et al. [2011] penalise an encoder Jacobian's Frobenius norm to enforce contractive features, and Srinivas and Fleuret [2018] match the student's input–output Jacobian to the teacher's, but both were developed for single-pass image classification, where errors traverse the network only once and matching $J^S \approx J^T$ globally suffices for generalisation. Long-chain reasoning is qualitatively different: the same residual block is traversed L times within a single forward pass, so what matters is not whether the student's Jacobian *resembles* the teacher's, but whether the student's per-layer Jacobian *actively contracts* transversal deviations across the depth of the network. Neither tradition developed the tools for this question.

Our approach. We treat the Transformer residual stream as a discrete dynamical system and ask for a stronger property: *orbital stability*, in which perturbations transversal to the teacher's reasoning trajectory actively contract, while the along-trajectory reasoning signal is left unconstrained by the geometric loss. We derive the **Contractive Restoring Flow (CRF)** loss from this condition. The key insight is that reducing the loss drives $P_\perp J_f^{(l)} \to -\alpha P_\perp$—an active restoring force that overcomes the irreducible identity contribution of the residual connection, achieving strict contraction at rate $1-\alpha$ per layer and a provable global basin of attraction around the teacher's trajectory within the linearisation regime. Table 1 situates CRF along these axes against prior art.

Our main contributions are:

- **Problem diagnosis.** We formalize the first-order information unconstrained by hidden-state matching and give a sufficient condition for transverse contraction about the teacher trajectory.
- **CRF loss.** We derive a tractable first-order objective that provably achieves $\|P_\perp J_l\|_2 = 1-\alpha < 1$ at the loss minimum while leaving the tangential dynamics unconstrained, and prove a global basin of attraction for any initial deviation within the linearisation regime.
- **Empirical validation.** We evaluate layer-wise random-projection statistics and robustness to inference-time hidden-state noise. These diagnostics provide empirical consistency checks for the proposed stability mechanism within the tested settings. CRF achieves the highest reported average gains across the evaluated student scales and teacher architectures.

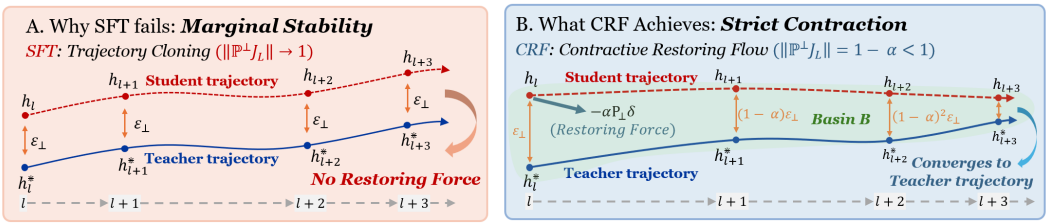


Figure 1: **Illustration of transversal dynamics.** Left: a marginally stable trajectory-matching solution with $\|P_\perp J_l\|_2 = 1$. Right: the CRF contraction target $\|P_\perp J_l\|_2 = 1-\alpha < 1$.

2 Stability Analysis of Reasoning Distillation

2.1 Hidden-State Trajectory Cloning and Its Limitation

Liang et al. [2023] and Dasgupta and Cohn [2025] supervise the student's hidden states layer-by-layer via:

$$\mathcal{L}_{\text{match}} = \sum_{l=0}^{L} \left\| h_\theta^{(l)}(x) - h_l^* \right\|^2, \tag{1}$$

where h_l^* are the teacher's hidden states and $h_\theta^{(l)}(x)$ are the student's. Minimising (1) imposes a *zeroth-order* constraint: it controls $f_\theta^{(l)}$ at the single training point h_l^* but leaves the student's behaviour away from that point entirely unconstrained.

At inference, autoregressive error, approximation error, and distribution shift all cause the student's hidden states to deviate from $\{h_l^*\}$; once a deviation occurs, the student's subsequent states are governed entirely by its own dynamics that $\mathcal{L}_{\text{match}}$ has left unconstrained. We therefore ask: ***when the student's hidden state deviates from the teacher's trajectory, does that deviation grow or shrink as it propagates through the remaining layers?***

2.2 The Transformer as a Discrete Dynamical System

A Transformer with L layers updates hidden states $h_l \in \mathbb{R}^d$ via residual connections:

$$h_{l+1} = h_l + f_\theta^{(l)}(h_l) \;\triangleq\; T_\theta^{(l)}(h_l), \tag{2}$$

where $f_\theta^{(l)}$ is the residual function. This is precisely a discrete dynamical system; we are not imposing new structure on the Transformer, but recognising what is already present.

Cross-scale alignment. The teacher and student need not share hidden dimension or depth. A 72B teacher operates in $\mathbb{R}^{d_T}$ with L_T layers; a 0.5B student in $\mathbb{R}^{d_S}$ with L_S layers. We bridge the two spaces with (i) a fixed layer-mapping $\pi : \{0, \dots, L_S\} \to \{0, \dots, L_T\}$ that selects which teacher layer each student layer should track, and (ii) per-layer linear alignments $A_l : \mathbb{R}^{d_T} \to \mathbb{R}^{d_S}$ learned jointly with the student. The *aligned teacher trajectory* $h_l^* := A_l\, h_{\pi(l)}^{T,*} \in \mathbb{R}^{d_S}$ is a discrete sequence in the student's latent space, and its residual flux $\Delta_l^* := h_{l+1}^* - h_l^*$ is the step the aligned teacher takes through that space. **For the remainder of the paper, every geometric object** (h_l^*, Δ_l^*, u_l^*, $P_\perp^{(l)}$, J_l, ϵ_l, r_l) **lives in** $\mathbb{R}^{d_S}$; teacher-side states enter only via $A_l h_{\pi(l)}^{T,*}$. Implementation details are in Appendix E.

The aligned teacher trajectory $\mathcal{T}^* = \{h_0^*, \dots, h_L^*\}$ thus lives in $\mathbb{R}^{d_S}$, with Δ_l^* being the step the aligned teacher takes through the student's latent space at layer l.

Let $\epsilon_l = h_\theta^{(l)} - h_l^*$ be the student–teacher deviation in $\mathbb{R}^{d_S}$. A first-order Taylor expansion around h_l^* gives:

$$\epsilon_{l+1} \approx J_l\, \epsilon_l + r_l, \qquad J_l = \left. \frac{\partial T_\theta^{(l)}}{\partial h} \right|_{h_l^*}, \tag{3}$$

where $r_l = f_\theta^{(l)}(h_l^*) - \Delta_l^*$ is the zeroth-order fitting residual controlled by $\mathcal{L}_{\text{match}}$, and $J_l \epsilon_l$ describes how the existing deviation is amplified or attenuated layer by layer, a term $\mathcal{L}_{\text{match}}$ does not address.

2.3 Why Trajectory Matching Is Not Enough

Not all errors are equally harmful. An error along the teacher's step Δ_l^* is a difference in pacing: the student is still on the same reasoning path. An error perpendicular to Δ_l^* means the student has drifted sideways off the path entirely. It is the perpendicular errors that compound into divergent reasoning chains across layers; the parallel ones do not. Under the linearised dynamics (3), this compounding is governed by $\|P_\perp J_l\|_2$: values above one amplify perpendicular deviations; values below one attenuate them.

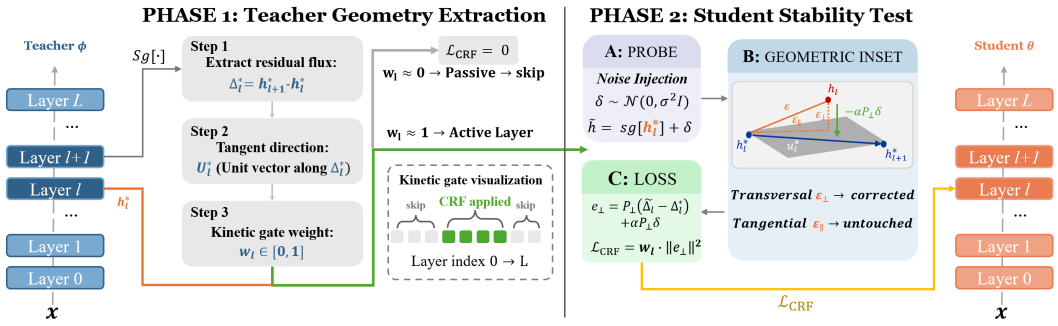


Figure 2: **Two-phase CRF training.** Phase I extracts the teacher's residual flux, tangent directions, and kinetic-gate weights. Phase II probes the student with noise and computes the CRF loss.

We therefore project onto the subspace orthogonal to Δ_l^*, following the transverse contraction framework of Manchester and Slotine [2014] and Forni and Sepulchre [2014]. Writing $u_l^* = \Delta_l^* / \|\Delta_l^*\|_2$, the transversal projection is

$$P_\perp^{(l)} = I - u_l^* (u_l^*)^\top. \tag{4}$$

Lemma 1 (Projected Full Jacobian)**.** *For any residual layer with $J_l = I + J_f^{(l)}$: $P_\perp J_l = P_\perp + P_\perp J_f^{(l)}$.*

Proof. $P_\perp J_l = P_\perp (I + J_f^{(l)}) = P_\perp + P_\perp J_f^{(l)}$. □

The residual connection blocks the goal $\|P_\perp J_l\|_2 < 1$. By Lemma 1, the identity contributes a fixed $P_\perp$ term that training cannot remove: even when $J_f^{(l)} \to 0$, we get $\|P_\perp J_l\|_2 \to 1$, perpendicular errors pass through every layer unchanged, a state of *marginal stability* (Figure 1, left).

A sufficient restoring-force condition is $P_\perp J_f^{(l)} = -\alpha P_\perp$ with $0 < \alpha < 1$. Under this condition, $P_\perp J_l = (1 - \alpha) P_\perp$ and $\|P_\perp J_l\|_2 = 1 - \alpha < 1$. The next section derives a loss targeting this condition.

3 Method: Contractive Restoring Flows

Section 2 identifies a precise target: train the student so that $P_\perp^{(l)} J_f^{(l)} = -\alpha P_\perp^{(l)}$ at every active layer. This section derives a tractable loss that enforces this condition and establishes the resulting global stability guarantee. All formal proofs are in Appendix B.

3.1 Loss Design

The target condition is a statement about the *slope* of $f_\theta^{(l)}$, not its value; no loss evaluated only at h_l^* can constrain it. To probe the Jacobian, we inject isotropic Gaussian noise $\delta \sim \mathcal{N}(0, \sigma^2 I)$ and penalise any deviation from the ideal transversal response $-\alpha P_\perp \delta$:

$$\boxed{\mathcal{L}_{\text{CRF}}^{(l)} = \mathbb{E}_{\delta \sim \mathcal{N}(0, \sigma^2 I)} \Big[\big\| P_\perp \big(f_\theta^{(l)}(h_l^* + \delta) - \Delta_l^* \big) + \alpha P_\perp \delta \big\|_2^2 \Big].} \tag{5}$$

The $+\alpha P_\perp \delta$ term is not a regulariser: without it, minimising the loss drives $P_\perp J_f^{(l)} \to 0$, which by Lemma 1 gives $\|P_\perp J_l\|_2 \to 1$ recovering marginal stability rather than contraction. The shift by αI moves the minimisation target to the restoring force condition $P_\perp J_f^{(l)} = -\alpha P_\perp$.[2]

[2]For a probe $\delta \sim \mathcal{N}(0, \sigma^2 I)$ in $\mathbb{R}^d$, the typical perturbation magnitude is $\|\delta\| \approx \sigma\sqrt{d}$, so the Taylor-validity condition is $\sigma\sqrt{d}\,\|J_f^{(l)}\|_2 \ll 1$, i.e. $\sigma \ll 1/(\sqrt{d}\,\|J_f^{(l)}\|_2)$. Under the Local Regularity Assumption (Appendix D), Pre-LN Transformers

Under a first-order Taylor expansion, taking the expectation in closed form reveals that the loss decomposes into two independent squared-norm terms (Appendix B.3): a *Transversal Drift* term $\|P_\perp r_l\|_2^2$ that recovers standard hidden-state matching in the perpendicular subspace, and a *Spectral Energy* term $\sigma^2\|P_\perp(J_f^{(l)} + \alpha I)\|_F^2$ that enforces the first-order Jacobian constraint existing methods cannot provide. When the loss reaches zero, a four-step algebraic chain (Appendix B.4) shows that $\|P_\perp^{(l)} J_l\|_2 = 1 - \alpha < 1$: every transversal perturbation contracts by factor $1 - \alpha$ per layer. The tangential dynamics $(u_l^*)^\top J_l u_l^*$ are left unconstrained by the loss (Proposition 1); the loss is *tangentially agnostic*, so forward progress along the reasoning direction is governed by the SFT objective rather than impeded by the geometric constraint. In practice the loss is only approximately minimised: if $\mathcal{L}_{\text{CRF}}^{(l)} \le \varepsilon$ then $\|P_\perp J_l\|_2 \le (1 - \alpha) + \sqrt{\varepsilon}/\sigma$, so contraction degrades continuously with training loss (Appendix B.4).

3.2 Multi-Layer Convergence and Global Stability

Consider the first-order deviation model $\epsilon_{l+1} = J_l \epsilon_l + r_l$, and write $P_l = P_\perp^{(l)}$. For every layer included in the bound, assume

$$P_l J_l = (1 - \alpha) P_l, \qquad \|u_{l+1}^* - u_l^*\|_2 \le \kappa, \qquad \|J_l\|_2 \le J_{\max}.$$

Assume independently that $\|\epsilon_l^{\|}\|_2 \le B^{\|}$ for $0 \le l \le L$. Define

$$R_\perp = \max_{0 \le l < L} \|P_{l+1} r_l\|_2, \qquad \gamma_{\text{eff}} = (1 - \alpha) + 2\kappa J_{\max}.$$

If $\gamma_{\text{eff}} < 1$, then

$$\|\epsilon_L^\perp\|_2 \le \gamma_{\text{eff}}^L \|\epsilon_0\|_2 + \frac{R_\perp + 2\kappa J_{\max} B^{\|}}{1 - \gamma_{\text{eff}}}. \tag{6}$$

Here $R_\perp$ and $B^{\|}$ are independent conditions of the bound. The proof is given in Appendix B.2. For a K-Lipschitz readout ℓ,

$$\|\ell(h_L) - \ell(h_L^*)\|_2 \le K \left(\gamma_{\text{eff}}^L \|\epsilon_0\|_2 + \frac{R_\perp + 2\kappa J_{\max} B^{\|}}{1 - \gamma_{\text{eff}}} + B^{\|} \right).$$

Stability ($\gamma_{\text{eff}} < 1$) and its inference-time analogue hold simultaneously whenever

$$\alpha > \max\big(2\kappa J_{\max},\ H_{\max} \cdot \delta_{\text{SFT}}\big), \tag{7}$$

In the ideal zero-loss setting, the inequalities above describe conditional training-point and pointwise inference-time properties. The multi-layer bound analyses the first-order deviation model under the stated assumptions. Applying it to nonlinear trajectories requires accounting for the projected Taylor remainder and maintaining the deviations within the chosen linearisation neighbourhood. Since the constants are not directly measurable, we use the condition as a design principle and evaluate the choice of α empirically.

3.3 Kinetic Gating for Passive Layers

Not every layer contributes to reasoning. When the residual flux $\mu_l = \|\Delta_l^*\|_2 \approx 0$, the tangent direction u_l^* is numerically ill-posed, and applying the CRF constraint forces the student to hallucinate a flow direction from noise. We exclude such layers by weighting each per-layer loss by a smooth gate:

$$\mathcal{L}_{\text{CRF}} = \sum_{l=0}^{L-1} w_l \, \mathcal{L}_{\text{CRF}}^{(l)}, \qquad w_l = \sigma_{\text{gate}}\Big(\tfrac{\mu_l - \tau}{\varepsilon_{\text{gate}}}\Big). \tag{8}$$

The gate is C^∞ and strictly increasing in μ_l, opening only when the teacher is actively computing. Passive layers contribute a factor of 1 to the contraction product and are transparent to the basin guarantee; the effective contraction rate over active layers is $(1 - \alpha)^{L_{\text{active}}}$. The formal justification including why applying CRF to passive layers is actively harmful, is in Appendix F. As $\mathcal{L}_{\text{SFT}}$ reduces the student–teacher gap δ_{SFT}, the inference-time contraction rate approaches $(1 - \alpha)$ from above, so the two loss components are mutually reinforcing rather than in tension; see Corollary 1 for the formal statement.

with RMSNorm and SwiGLU activations (e.g. the Qwen2.5 family) satisfy a Hessian bound and the approximation error is $O(\sigma^2 d)$ in expectation; the recommended range $\sigma \in [0.005, 0.02]$ is well inside the validity region for $d \le 3584$ and at the boundary for $d = 8192$.

3.4 Training Procedure

The Figure 2 summarize one training iteration. *Phase I* extracts the teacher's geometry without gradient: for each layer, the residual flux Δ_l^* and its magnitude μ_l are recorded, and the kinetic gate weight w_l is computed via (8). *Phase II* tests the student at each active layer by injecting a Gaussian probe at the stop-gradient teacher state, projecting the student's response onto the transversal subspace, and accumulating the CRF loss (5). The total objective $\mathcal{L}_{\mathrm{SFT}} + \lambda \mathcal{L}_{\mathrm{CRF}}$ is minimised in a single backward pass. The teacher hidden-state trajectory required by Phase I is collected offline; dataset construction and algorithm are described in Appendix H.

4 Experiments

In this section, we empirically evaluate the Contractive Restoring Flow objective introduced in Section 3. Following the training and evaluation protocol detailed below, we assess both its distillation performance and the stability mechanism predicted by our theoretical analysis.

4.1 Experiment Protocol

Training protocol and baseline fairness. All methods use the same teacher-generated CoT corpus, identical optimisation budget, and identical hyperparameter search budget per method. CRF's additional inputs (teacher hidden states, the noise-perturbed forward pass, per-layer alignments $\{A_l\}$) are intrinsic to the method, not a data or compute advantage; baselines that consume only output logits or token sequences cannot use them by construction.

We distil Qwen2.5-72B [Yang et al., 2024] into Qwen2.5-{0.5B, 3B, 7B}, holding the teacher fixed to eliminate architecture and vocabulary variables. Baselines span three levels: **token level** (Standard KD, CoT KD [Taori et al., 2023]); **distribution level** (Logit-CoT KD [Hinton et al., 2015, Peng et al., 2025], GKD [Agarwal et al., 2024], DASD [Yan et al., 2026]); and **hidden-state level** (layer-wise matching [Dasgupta and Cohn, 2025]).

We evaluate on **AIME** [Zhang and Math-AI, 2024] (competition math, greedy pass@1), **MBPP** [Austin et al., 2021] (coding), and **LiveCodeBench** [Jain et al., 2025] (OOD coding).

We replace the dense teacher with three MoE models: **Intra-family** ($\mathcal{T}_1$: Qwen3-235B [Yang et al., 2025]), sharing the student's tokenizer, and **Cross-family** ($\mathcal{T}_2$: DeepSeek-V3 [DeepSeek-AI, 2024]; $\mathcal{T}_3$: GPT-OSS [OpenAI, 2025]), with entirely different tokenizers and routing strategies, using Qwen3-{0.6B, 4B, 8B} as students. We report average Δ over AIME, MBPP, and LCB.

A complementary check on whether transversal regularisation incurs an alignment tax on general instruction-following is reported in Appendix K. The hidden state flow generation and training details are in Appendix H.

We evaluate whether CRF improves the robustness of distilled reasoning models relative to standard distillation under two regimes: **extreme compression**, where limited student capacity hinders recovery of the teacher's computation, and **cross-architecture distillation**, where architectural mismatch amplifies intermediate deviations. We address three research questions:

RQ1: Distillation performance. Does CRF improve distillation performance across compression and architecture regimes?

RQ2: Contraction and robustness. Does CRF induce the predicted transversal contraction and robustness to hidden-state perturbations?

RQ3: Hyperparameter sensitivity. Do the optima of α, σ, and λ lie in the theory-motivated regions?

4.2 RQ1: Distillation Performance

Extreme Compression Robustness. Table 2 reports the same-family distillation results.

As shown in Table 2, CRF achieves the largest absolute gain at the most compressed (0.5B) regime, where representational capacity is most constrained. The gain pattern across scales (7B: $+5.5$, 3B: $+4.7$, 0.5B: $+7.2$) is non-monotonic with a clear peak at extreme compression. CRF surpasses the strongest hidden-state-matching baseline at every scale (7B: $+5.5$ vs. $+4.3$, 3B: $+4.7$ vs. $+3.9$,

Table 2: **Distillation across student scales.** Pass@1 (%) on AIME, MBPP, and LCB with Qwen2.5-72B as the teacher. Avg.Δ is the mean absolute gain over the corresponding base student across the three benchmarks, in percentage points. Reported uncertainties are standard deviations over three seeds. CRF results are bold.

Setting		$\mathcal{T}$ (Qwen2.5-72B) $\rightarrow \mathcal{S}$			
Method	Student model	AIME	MBPP	LCB	Avg.Δ
Regime A: Standard Compression					
Base Student	**7B**	9.1	79.2	14.4	–
Standard KD	7B	9.7	79.5	15.0	+0.5
CoT KD	7B	10.9	80.4	16.7	+1.8
Logit-CoT KD	7B	12.3	81.2	18.1	+3.0
GKD	7B	13.0	82.2	18.7	+3.7
DASD	7B	14.3 ± 0.4	82.6 ± 0.2	20.6 ± 0.4	+4.9
Hidden state matching	7B	13.6 ± 0.5	82.3 ± 0.4	20.1 ± 0.4	+4.3
CRF	**7B**	$\mathbf{14.9 \pm 0.4}$	$\mathbf{83.3 \pm 0.3}$	$\mathbf{21.3 \pm 0.2}$	**+5.5**
Regime B: Aggressive Compression					
Base Student	**3B**	6.8	73.7	9.2	–
Standard KD	3B	7.0	74.2	9.9	+0.5
CoT KD	3B	7.9	75.9	11.2	+1.8
Logit-CoT KD	3B	8.1	77.1	11.8	+2.4
GKD	3B	8.8	78.8	13.0	+3.6
DASD	3B	9.1 ± 0.7	79.4 ± 0.4	13.2 ± 0.9	+4.0
Hidden state matching	3B	9.0 ± 0.6	78.9 ± 0.3	13.4 ± 0.7	+3.9
CRF	**3B**	$\mathbf{10.1 \pm 0.6}$	$\mathbf{79.5 \pm 0.4}$	$\mathbf{14.1 \pm 0.7}$	**+4.7**
Regime C: Extreme Compression					
Base Student	**0.5B**	0.7	49.6	4.3	–
Standard KD	0.5B	2.1	50.6	5.1	+1.1
CoT KD	0.5B	2.4	51.0	5.8	+1.5
Logit-CoT KD	0.5B	3.2	52.4	7.0	+2.7
GKD	0.5B	4.8	55.1	9.4	+4.9
DASD	0.5B	6.0 ± 0.15	56.0 ± 0.12	10.8 ± 0.16	+6.1
Hidden state matching	0.5B	6.4 ± 0.17	57.2 ± 0.13	11.4 ± 0.14	+6.8
CRF	**0.5B**	$\mathbf{6.6 \pm 0.11}$	$\mathbf{57.9 \pm 0.12}$	$\mathbf{11.6 \pm 0.9}$	**+7.2**

0.5B: $+7.2$ vs. $+6.8$), and the gap widens at the smallest student. Token-level methods show roughly constant Δ across regimes; only CRF amplifies its absolute advantage under extreme compression, consistent with the picture that orbital stability specifically benefits fragile students where per-layer error accumulation is otherwise unchecked.

Cross-Architecture Distillation Table 3 reports the MoE-to-dense distillation results.

Table 3 shows that distribution-matching baselines (Logit-CoT KD, DASD) lose effectiveness as architectural mismatch increases: their Δ in the cross-family settings ($\mathcal{T}_2$, $\mathcal{T}_3$) drops sharply relative to the intra-family setting ($\mathcal{T}_1$). CRF maintains consistent gains across all three teachers, indicating that geometric constraints on the latent trajectory transfer across routing mechanisms and tokenisation boundaries where logit-level alignment fails.

Non-Qwen Students. We distil the same Qwen2.5-72B teacher into LLaMA and Mistral students at the 3B and 8B scales, using the same training and evaluation protocol. CRF achieves the highest reported score on each benchmark for all four students (Table 4), extending the empirical evidence across student families.

4.3 RQ2: Contraction and Robustness

The performance gains in Section 4 establish that CRF outperforms trajectory-cloning baselines, but they do not by themselves distinguish between two very different explanations: CRF may be beneficial merely because it acts as a useful regulariser, or it may genuinely enforce the geometric structure its theory predicts. We resolve this question by probing the student's internal dynamics directly.

Table 3: **MoE-to-dense distillation.** Pass@1 (%) on AIME, MBPP, and LCB for Qwen3 students. Avg.Δ is the mean absolute gain over the corresponding base student across the three benchmarks, in percentage points. CRF results are bold.

Setting / Method	Size	$\mathcal{T}_1$: Qwen3-235B (MoE) AIME	MBPP	LCB	Avg.Δ	$\mathcal{T}_2$: DeepSeek-V3 (MoE) AIME	MBPP	LCB	Avg.Δ	$\mathcal{T}_3$: GPT-OSS (MoE) AIME	MBPP	LCB	Avg.Δ
Regime A: Standard Compression (Student: 8B)													
Base Student	**8B**	76.0	72.8	57.5	–	76.0	72.8	57.5	–	76.0	72.8	57.5	–
Standard KD	8B	76.9	73.6	58.5	+0.9	76.2	73.0	57.8	+0.2	76.3	73.1	57.9	+0.3
CoT KD	8B	79.3	74.8	60.3	+2.7	76.5	73.5	58.2	+0.6	76.8	73.4	58.4	+0.8
Logit-CoT KD	8B	80.4	75.2	61.7	+3.7	77.1	74.0	59.0	+1.3	77.4	74.2	59.5	+1.6
GKD	8B	80.9	76.4	62.5	+4.5	78.2	75.0	60.5	+2.5	78.5	75.3	60.8	+2.8
DASD	8B	82.6	76.5	64.9	+5.9	78.5	75.2	61.0	+2.8	78.8	75.5	61.2	+3.1
CRF	**8B**	**83.7**	**76.2**	**66.4**	**+6.7**	**81.2**	**76.0**	**64.5**	**+5.1**	**81.8**	**76.4**	**64.8**	**+5.6**
Regime B: Aggressive Compression (Student: 4B)													
Base Student	**4B**	73.8	66.7	54.2	–	73.8	66.7	54.2	–	73.8	66.7	54.2	–
Standard KD	4B	74.4	67.7	55.6	+1.0	74.0	66.9	54.5	+0.2	74.1	67.0	54.6	+0.3
CoT KD	4B	75.9	71.4	58.2	+3.6	74.5	67.5	55.0	+0.8	74.6	67.8	55.2	+1.0
Logit-CoT KD	4B	76.2	72.0	59.2	+4.2	74.8	68.0	55.5	+1.2	75.0	68.2	55.8	+1.4
GKD	4B	78.4	75.1	61.0	+6.6	75.5	69.5	57.0	+2.4	75.9	69.8	57.4	+2.8
DASD	4B	78.5	75.3	61.3	+6.8	75.8	70.0	57.5	+2.9	76.0	70.5	57.8	+3.2
CRF	**4B**	**80.1**	**75.1**	**62.6**	**+7.7**	**78.2**	**73.5**	**60.5**	**+5.8**	**78.8**	**74.0**	**61.0**	**+6.4**
Regime C: Extreme Compression (Student: 0.6B)													
Base Student	**0.6B**	10.7	36.6	12.3	–	10.7	36.6	12.3	–	10.7	36.6	12.3	–
Standard KD	0.6B	12.5	37.8	14.3	+1.7	11.2	36.8	12.8	+0.4	11.3	37.0	12.9	+0.5
CoT KD	0.6B	12.6	38.4	14.9	+2.1	11.5	37.2	13.2	+0.8	11.6	37.5	13.5	+1.0
Logit-CoT KD	0.6B	15.2	43.6	17.2	+5.5	12.8	38.5	14.0	+1.9	13.0	39.0	14.2	+2.2
GKD	0.6B	17.7	45.2	19.7	+7.7	14.5	40.0	16.5	+3.8	14.8	40.5	16.8	+4.2
DASD	0.6B	18.4	45.5	20.5	+8.3	15.2	41.0	17.0	+4.5	15.5	41.5	17.5	+5.0
CRF	**0.6B**	**18.5**	**45.8**	**20.9**	**+8.5**	**17.8**	**44.5**	**19.5**	**+7.4**	**18.2**	**45.0**	**20.0**	**+7.9**

Table 4: **Distillation to LLaMA and Mistral students.** Pass@1 (%) with Qwen2.5-72B as the teacher. The top and bottom groups show 3B and 8B students. Avg.Δ is the mean absolute gain over the corresponding base student across the three benchmarks, in percentage points.

Setting / Method	LLaMA-3B AIME	MBPP	LCB	Avg.Δ	Mistral-3B AIME	MBPP	LCB	Avg.Δ
Base	3.0	64.8	8.9	–	5.2	67.7	10.7	–
DASD	5.5	68.1	11.4	+2.8	6.0	71.6	13.4	+2.5
Hidden State Matching	5.8	68.9	11.3	+3.1	6.8	71.9	13.2	+2.8
CRF	**6.0**	**69.4**	**12.2**	**+3.6**	**7.1**	**72.4**	**14.8**	**+3.6**
Setting / Method	**LLaMA-8B AIME**	**MBPP**	**LCB**	**Avg.Δ**	**Mistral-8B AIME**	**MBPP**	**LCB**	**Avg.Δ**
Base	7.4	70.3	12.1	–	8.5	70.9	14.0	–
DASD	10.4	72.2	17.1	+3.3	12.2	72.5	19.3	+3.5
Hidden State Matching	10.0	72.5	17.0	+3.2	12.9	72.4	19.8	+3.9
CRF	**11.2**	**73.6**	**18.1**	**+4.4**	**13.5**	**73.7**	**20.0**	**+4.6**

Does CRF Achieve the Predicted Contraction Rate? Theorem 3 gives a precise prediction: when CRF is trained with coefficient α and the loss is driven to zero, the transversal Jacobian norm at every active layer should approach $1-\alpha$, with deviations bounded by $(1-\alpha)+\sqrt{\varepsilon}/\sigma$ at training loss ε (Proposition 2). We test this by estimating $\|\widehat{P_\perp J_l}\|_2$ via finite-difference random projections after training (see Appendix H for the measurement procedure).

Figure 3(a) shows the result. The three CRF variants each converge close to their respective target $1-\alpha$ in active layers, with confidence bands that are visibly narrower than those of the ablated and baseline variants, reflecting the tighter Jacobian constraint imposed by the Spectral Energy term.

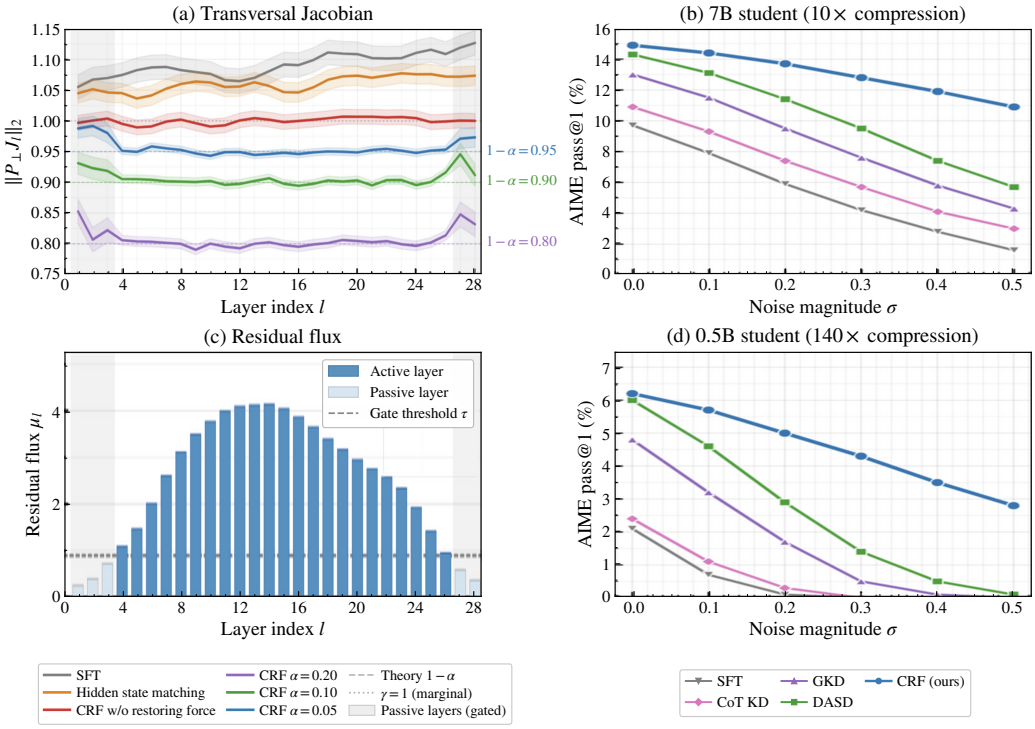


Figure 3: **Layer-wise dynamics and robustness to hidden-state noise.** (a) Random-projection statistic $\hat{s}_l$, computed by finite differences using 64 unit directions (Appendix H.1). Dashed lines mark the zero-loss targets $1-\alpha$; the reference at 1 marks marginal stability. (b) AIME pass@1 (%) under Gaussian noise of magnitude σ for the 7B student. (c) Residual-flux magnitude μ_l and gate threshold τ; pale bars mark passive layers. (d) The same noise test for the 0.5B student.

The empirical observations are mutually consistent with the theory: *(i)* estimated values cluster near $1-\alpha$ across active layers; *(ii)* the no-restoring-force ablation (red curve) yields $\widehat{\|P_\perp J_l\|}_2 \approx 1$, falsifying contraction in the absence of $+\alpha P_\perp \delta$ as Lemma 1 predicts; and *(iii)* varying α produces commensurate shifts in measured $\widehat{\|P_\perp J_l\|}_2$. $\widehat{\|P_\perp J_l\|}_2$ is a lower-bound estimator (Appendix H.1); the upper-bound condition from Proposition 2 is therefore not formally verified, but the three patterns above are jointly consistent with the theory and unlikely under any null lacking the proposed contraction mechanism.

Observation *(ii)* deserves emphasis: it is the cleanest mechanistic contrast in the entire experimental section. Without the $+\alpha P_\perp \delta$ shift, the Spectral Energy term drives $P_\perp J_f^{(l)} \to 0$ instead of $P_\perp J_f^{(l)} \to -\alpha P_\perp$, and the residual connection's identity contribution forces $\|P_\perp J_l\|_2 \to 1$ (Lemma 1). The empirical curve confirms this: $\gamma \approx 1$ everywhere, exactly the marginal stability that the theory predicts in the absence of the restoring force. A single structural change—adding $+\alpha P_\perp \delta$—is therefore sufficient to convert marginal stability into strict contraction, exactly as the algebraic chain in Appendix B.4 predicts.

Does the Contraction Produce an Attractor Basin? Figure 3(a) provides empirical evidence for the per-layer Jacobian property. Theorem 4 predicts that composing this property across L layers produces a basin of attraction within the linearisation regime: the transversal error decays geometrically as $\gamma_{\text{eff}}^L \|\epsilon_0\|$ for any initial deviation $\|\epsilon_0\|$ small enough for the first-order Taylor expansion in (3) to apply. To test whether this global property holds at inference time, we inject Gaussian noise $\delta \sim \mathcal{N}(0, \sigma^2 I)$ into every layer and measure AIME pass@1 as a function of σ.

Figure 3(b,d) shows a clear ordering: CRF maintains the flattest degradation curve at both scales. More informatively, the *rank ordering* of methods under noise closely mirrors their rank ordering of $\|P_\perp J_l\|_2$ measured in Figure 3(a); lower contraction rate corresponds directly to higher noise

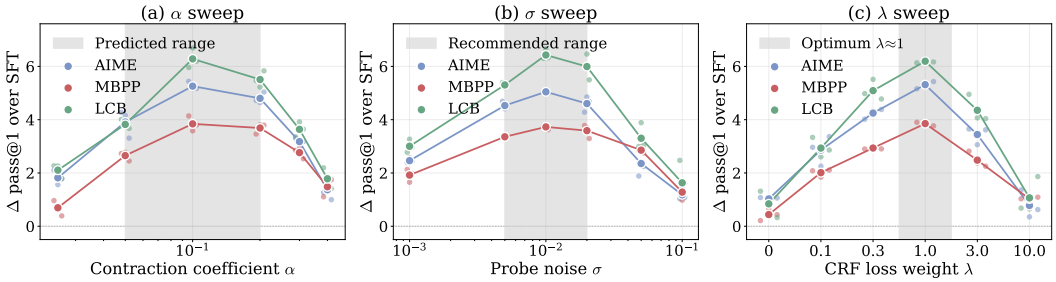


Figure 4: **Hyperparameter sensitivity.** Sweeps of (a) α, (b) σ, and (c) λ for Qwen2.5-7B distilled from Qwen2.5-72B. Curves show mean improvements in pass@1 over the SFT baseline across three seeds; shaded regions indicate the theory-motivated operating ranges.

robustness. This two-sided correspondence closes the causal chain: CRF shapes the Jacobian at training time (Figure 3(a)), and the shaped Jacobian attenuates perturbations at inference time (Figure 3(b,d)).

The 0.5B results (Figure 3(d)) are particularly telling. At $\sigma = 0.2$, models without a contraction guarantee collapse, while CRF remains functional. This is the empirical signature of the steady-state error bound in Theorem 4(ii): with $\gamma_{\text{eff}} < 1$, the cumulative transversal error is bounded above by $R_\perp/(\alpha - 2\kappa J_{\max})$ regardless of chain length or initial deviation, even in a student whose small capacity leaves it with minimal intrinsic error tolerance.

4.4 RQ3: Hyperparameter Sensitivity

The preceding experiments examine layer-wise diagnostics and task-level robustness at fixed hyperparameter values. A third question concerns the hyperparameters themselves. Theorem 4 requires $\alpha > 2\kappa J_{\max}$, but κ and $J_{\max}$ are not directly measurable on a trained Transformer (Limitations); we therefore use the theorem as a *design principle* rather than a falsifiable point prediction, and treat $\alpha \in [0.05, 0.20]$ as the conjectured operating range. Propositions 2 and 4 jointly motivate $\sigma \in [0.005, 0.02]$ from the Taylor-validity / contraction-precision trade-off; Lemma 1 predicts that $\lambda = 0$ collapses to marginal stability. Figure 4 sweeps each parameter on Qwen2.5-7B distilled from Qwen2.5-72B (three seeds; Δ pass@1 over the SFT baseline).

All three empirical optima are consistent with the theory-motivated regions. The asymmetric degradation past each optimum tracks the theoretical mechanism: large α over-contracts and AIME (the longest-chain benchmark) degrades fastest, consistent with over-contraction interfering with the tangential signal (Proposition 1); large σ violates the Taylor-validity assumption in Proposition 4; large λ over-weights the geometric constraint relative to SFT. The systematically smaller Δ on MBPP reflects its shorter-chain structure, where error accumulation is less severe and the contraction guarantee yields smaller marginal gain. These sweeps are consistency checks, not formal validations of $\alpha > 2\kappa J_{\max}$ (κ, $J_{\max}$ unmeasurable). Exact values, confidence intervals, and the decomposition $\times$ gate ablation are in Appendix I.

5 Conclusions and Future Work

We presented Contractive Restoring Flows (CRF), a distillation objective that contracts transversal deviations from teacher reasoning trajectories while leaving tangential learning to the SFT objective. Within the linearisation regime, our analysis establishes per-layer contraction and multi-layer stability, and experiments across compression and architecture regimes show consistent gains, with diagnostics consistent with the predicted mechanism. CRF thus offers a geometric approach to improving the robustness of reasoning distillation.

Future work will focus on reducing CRF's reliance on per-layer teacher hidden states, with the aim of extending the method to logit-only or token-only distillation in black-box settings. We will also develop measurable criteria and validation procedures for the stability conditions, providing clearer hyperparameter guidance and enabling more direct empirical tests of the theoretical predictions.

Acknowledgments

HL was supported by the Shanghai General AI Foundation Models Program (No. 2025SHZDZX026D35), the Beijing Natural Science Foundation (No. L257007), the Beijing Major Science and Technology Project (No. Z251100008425006), and the CCF-Kuaishou Large Model Explorer Fund (No. CCF-KuaiShou 2026018). MG was supported by ARC DP240102088 and WIS–MBZUAI 142571.

References

Chen Liang, Simiao Zuo, Qingru Zhang, Pengcheng He, Weizhu Chen, and Tuo Zhao. Less is more: Task-aware layer-wise distillation for language model compression. In *Proceedings of the International Conference on Machine Learning*, 2023.

Sayantan Dasgupta and Trevor Cohn. Improving language model distillation through hidden state matching. In *International Conference on Learning Representations*, 2025.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024.

Shaotian Yan, Kaiyuan Liu, Chen Shen, Bing Wang, Sinan Fan, Jun Zhang, Yue Wu, Zheng Wang, and Jieping Ye. Distribution-aligned sequence distillation for superior Long-CoT reasoning. arXiv preprint arXiv:2601.09088, 2026.

Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the International Conference on Machine Learning*, 2011.

Suraj Srinivas and Francois Fleuret. Knowledge transfer with Jacobian matching. In *Proceedings of the International Conference on Machine Learning*, 2018.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015.

Ian R. Manchester and Jean-Jacques E. Slotine. Transverse contraction criteria for existence, stability, and robustness of a limit cycle. *Systems & Control Letters*, 63:32–38, 2014.

Fulvio Forni and Rodolphe Sepulchre. A differential Lyapunov framework for contraction analysis. *IEEE Transactions on Automatic Control*, 59(3):614–628, 2014.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115, 2024.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpaca: A strong, replicable instruction-following model. Stanford Center for Research on Foundation Models Blog, March 2023.

Hao Peng, Xin Lv, Yushi Bai, Zijun Yao, Jiajie Zhang, Lei Hou, and Juanzi Li. Pre-training distillation for large language models: A design space exploration. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2025.

Yifan Zhang and Team Math-AI. American Invitational Mathematics Examination (AIME) 2024. Hugging Face dataset: math-ai/aime24, 2024.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. arXiv preprint arXiv:2108.07732, 2021.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida I. Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations*, 2025.

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. arXiv preprint arXiv:2505.09388, 2025.

DeepSeek-AI. DeepSeek-V3 technical report. arXiv preprint arXiv:2412.19437, 2024.

OpenAI. gpt-oss-120b & gpt-oss-20b model card. arXiv preprint arXiv:2508.10925, 2025.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-Instruct: Aligning language models with self-generated instructions. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2023a.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! Outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL*, 2023.

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. SCOTT: Self-consistent chain-of-thought distillation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2023b.

Seungone Kim, Se June Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. The CoT collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2023.

Kaituo Feng, Changsheng Li, Xiaolu Zhang, Jun Zhou, Ye Yuan, and Guoren Wang. Keypoint-based progressive chain-of-thought distillation for LLMs. In *Proceedings of the International Conference on Machine Learning*, 2024.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2011.

Andrea Pozzi, Alessandro Incremona, Daniele Tessera, and Daniele Toti. Mitigating exposure bias in large language model distillation: An imitation learning approach. *Neural Computing and Applications*, 37(18):12013–12029, 2025.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, 2019.

Shaked Brody, Uri Alon, and Eran Yahav. On the expressivity role of LayerNorm in transformers' attention. In *Findings of the Association for Computational Linguistics: ACL*, 2023.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The Lipschitz constant of self-attention. In *Proceedings of the International Conference on Machine Learning*, 2021.

Valérie Castin, Pierre Ablin, and Gabriel Peyré. How smooth is attention? In *Proceedings of the International Conference on Machine Learning*, 2024.

Noam Shazeer. GLU variants improve transformer. arXiv preprint arXiv:2002.05202, 2020.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. MT-Bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2024.

Yikun Sun, Zhen Wan, Nobuhiro Ueda, Sakiko Yahata, Fei Cheng, Chenhui Chu, and Sadao Kurohashi. Rapidly developing high-quality instruction data and evaluation benchmark for large language models with minimal human effort: A case study on Japanese. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, 2024.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. AlpacaEval: An automatic evaluator of instruction-following models. GitHub repository: tatsu-lab/alpaca_eval, May 2023.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2022.

Shang Liu, Yao Lu, Wenji Fang, Mengming Li, and Zhiyao Xie. OpenLLM-RTL: Open dataset and benchmark for LLM-aided design RTL generation. In *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, 2024.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free Dolly: Introducing the world's first truly open instruction-tuned LLM. Databricks Blog, April 2023.

Appendix Contents

Appendix A Related Work 14
Appendix B Complete Proofs 15
Appendix C Projection Properties 20
Appendix D Local Regularity of Transformer Residual Blocks 21
Appendix E Cross-Scale Alignment Implementation 24
Appendix F Kinetic Gating: Formal Analysis 25
Appendix G CRF Training Algorithm 26
Appendix H Training Dataset Generation 26
Appendix I Ablation Study 29
Appendix J Cognitive Integrity Evaluation 29
Appendix K Notation Reference 30

A Related Work

A.1 Token-level Distillation

Early LLM distillation primarily utilized teacher-generated synthetic data for supervised fine-tuning, achieving behavior-level alignment [Wang et al., 2023a, Taori et al., 2023]. Following the advent of Chain-of-Thought (CoT) [Wei et al., 2022], focus shifted from final answers to reasoning processes. Subsequent works employed rationale extraction [Hsieh et al., 2023] and CoT-based augmentation [Wang et al., 2023b, Kim et al., 2023, Feng et al., 2024] to explicitly expose reasoning traces, thereby improving imitation fidelity. However, these text-centric approaches remain in the realm of explicit behavior cloning: students are optimized to mimic surface-level rationales without necessarily internalizing the underlying reasoning mechanism, leaving them vulnerable to inference-time deviations.

A.2 Trajectory-Level Distillation in Latent Space

Hidden-state distillation supervises a student's intermediate representations using teacher representations [Liang et al., 2023, Dasgupta and Cohn, 2025]. Compared to token-level supervision, trajectory-level distillation provides denser signals and exposes intermediate reasoning processes.

However, most existing methods treat the teacher's hidden states as fixed targets. During inference, the student generates its own trajectory step by step, while some deviations can accumulate and push the model into unseen regions of latent space [Ross et al., 2011, Pozzi et al., 2025]. Consequently, these approaches focus on trajectory alignment and ignore the stability of reasoning dynamics, which could be viewed as a hidden-state flow.

A.3 First-Order Jacobian Methods Outside LLM Distillation

Jacobian-based supervision in deep networks predates LLM distillation by nearly a decade. Two works are most directly comparable to CRF along the axes of Table 1.

Contractive Auto-Encoders. Rifai et al. [2011] add a Frobenius-norm penalty $\|\partial h/\partial x\|_F^2$ on the encoder Jacobian, encouraging hidden representations to be locally invariant to input perturbations. The motivation is unsupervised feature learning: a contractive encoder discovers low-dimensional manifolds along which the data varies and ignores noise directions. Three structural differences from CRF: *(i)* the penalty is on the *encoder Jacobian as a whole* (one forward pass), not on per-layer residual Jacobians; *(ii)* it contracts *all directions equally* via the Frobenius norm, with no mechanism to preserve a reasoning direction along which the trajectory should propagate; *(iii)* it operates on the student in isolation (no teacher trajectory), so it cannot encode the teacher's per-layer geometry.

Jacobian Matching for Distillation. Srinivas and Fleuret [2018] match the student's input–output Jacobian to the teacher's via $\|J^S - J^T\|_F^2$. The setting is image classification, where errors traverse the network only once. Three structural differences: *(i)* matching $J^S \approx J^T$ does *not* produce contraction—if the teacher's Jacobian is itself near-marginal in some directions, the student inherits that marginality; *(ii)* the matching is *global* (one input–output Jacobian), not per-residual-block, so it

cannot prevent depth-wise compounding; *(iii)* the teacher's Jacobian serves as the supervision target rather than a structural reference frame, conflating what the student should *resemble* with what it should *enforce*.

The CRF distinction. CRF does not match a teacher Jacobian; it imposes a structural condition—transversal spectral radius below one—at every active residual block, in the subspace orthogonal to the teacher's per-layer reasoning step. The construction targets long-chain depth-wise stability rather than single-pass feature invariance, which is why CRF's largest gains appear in the high-depth, error-accumulating regime (Sec 4.2).

A.4 Contraction Analysis in Dynamical Systems

The transverse contraction view we adopt has a substantial mathematical lineage in control theory and continuous-time dynamics. Manchester and Slotine [2014] introduced transverse contraction criteria for limit cycles: a system is transversely contractive around a periodic orbit if perturbations perpendicular to the orbit decay exponentially while along-orbit phase is unconstrained.

Forni and Sepulchre [2014] developed a parallel differential Lyapunov framework showing that contraction can be formalised via metrics on the tangent bundle of the state space.

These frameworks operate on continuous-time, smooth systems. CRF translates the core insight—orbital stability requires perpendicular contraction without constraining tangential motion—to the discrete-time, finite-depth Transformer residual stream. The translation is non-trivial: *(i)* discrete layer-wise iteration replaces continuous flow; *(ii)* the residual connection contributes a fixed identity term that must be actively cancelled by a restoring force (Lemma 1); and *(iii)* the manifold of valid hidden states is implicit, defined by what the teacher visits, rather than analytically specified. To our knowledge, CRF is the first application of the transverse contraction principle to neural-network distillation.

A.5 Error Accumulation and Imitation Learning

A separate line of work treats trajectory cloning under the lens of imitation learning under covariate shift. Ross et al. [2011] introduced DAgger to address the exposure bias problem: a student trained only on the teacher's state distribution behaves poorly when it must navigate states it generated itself. Pozzi et al. [2025] adapts an imitation-learning approach to language-model distillation to mitigate exposure bias.

B Complete Proofs

This appendix presents the complete, self-contained proofs of every theoretical result stated in the main text. Each proof is structured so that its opening sentence restates the claim in the context of the surrounding argument, and every subsequent step follows by direct logical implication from the premises already established. The proof order mirrors the deductive chain in the paper: geometric foundations (Section B.1), error dynamics (Section B.2), loss analysis (Section B.3), contraction guarantees (Section B.4), and global stability (Section B.5).

Convention on state space. Following the cross-scale alignment introduced in Sec 2.2, all state-space variables in the proofs below (ϵ_l, J_l, $P_\perp^{(l)}$, r_l, u_l^*) live in the student's latent space $\mathbb{R}^{d_S}$; teacher-side hidden states appear only through the aligned trajectory $h_l^* = A_l\, h_{\pi(l)}^{T,*}$, so that no implicit assumption is made about teacher–student dimension equality.

B.1 Geometric Foundations

Proof of Lemma 1 (Projected Full Jacobian). The claim is that for any transversal projection $P_\perp = I - u_l^*(u_l^*)^\top$ and any residual layer whose full Jacobian satisfies $J_l = I + J_f^{(l)}$, the projected Jacobian decomposes additively as $P_\perp J_l = P_\perp + P_\perp J_f^{(l)}$.

Proof. Because $J_l = I + J_f^{(l)}$ by the residual structure (Equation (2)), we compute directly:

$$P_\perp J_l = P_\perp \big(I + J_f^{(l)}\big) = P_\perp I + P_\perp J_f^{(l)} = P_\perp + P_\perp J_f^{(l)},$$

where the second equality uses left-distributivity of matrix multiplication and the third uses $P_\perp I = P_\perp$. □

Remark. The decomposition reveals why the residual connection is the core obstruction to contraction: even when $J_f^{(l)} \to 0$, the identity contributes a fixed $P_\perp$ term with spectral norm 1, giving $\|P_\perp J_l\|_2 \to \|P_\perp\|_2 = 1$ — *marginal stability*, not contraction. Achieving strict contraction $\|P_\perp J_l\|_2 < 1$ therefore requires $P_\perp J_f^{(l)}$ to be a *Restoring Force* that actively opposes and overcomes the identity's contribution.

B.2 Error Deviation Dynamics

Theorem 1 (Multi-Layer Transversal Bound)**.** *Let* $\gamma_{\text{eff}} = (1-\alpha) + 2\kappa J_{\max}$ *and* $R_\perp = \max_l \|P_\perp^{(l+1)} r_l\|_2$. ***Assume the tangential component is bounded:*** $\|\epsilon_l^{\parallel}\|_2 \le B^{\parallel}$ *for all l, where* $B^{\parallel}$ *is provided by the SFT objective (which constrains* ϵ_l *along all directions, including tangential). If* $\gamma_{\text{eff}} < 1$*, then for all* $L \ge 0$*:*

$$\|\epsilon_L^\perp\| \;\le\; \gamma_{\text{eff}}^L \,\|\epsilon_0\| \;+\; \frac{R_\perp + 2\kappa J_{\max}\, B^{\parallel}}{1-\gamma_{\text{eff}}}.$$

The theorem states that if CRF achieves $\|P_\perp^{(l)} J_l\|_2 \le \gamma$ at every active layer *and* the tangential deviation is independently bounded by $B^{\parallel}$ (the role of the SFT term), the accumulated transversal deviation at depth L satisfies the bound above. The additional $2\kappa J_{\max} B^{\parallel}$ term in the steady-state floor reflects projection rotation between layers converting a fraction of the bounded tangential error into transversal error; the term vanishes as either $\kappa \to 0$ (constant tangent direction) or $B^{\parallel} \to 0$ (perfect SFT fit). Without the tangential boundedness assumption, transversal contraction alone is insufficient to control perpendicular accumulation, since rotation of u_l^* can re-inject unbounded along-trajectory error into the transversal subspace.

Proof. **Step 1: Single-layer transversal recursion.** Let $\epsilon_l^\perp = P_\perp^{(l)} \epsilon_l$ be the transversal deviation at layer l. From the first-order Taylor expansion (3), $\epsilon_{l+1} \approx J_l \epsilon_l + r_l$. Applying $P_\perp^{(l+1)}$ and decomposing $P_\perp^{(l+1)} = P_\perp^{(l)} + \big(P_\perp^{(l+1)} - P_\perp^{(l)}\big)$:

$$\|\epsilon_{l+1}^\perp\|_2 \le \gamma_{\text{eff}} \|\epsilon_l^\perp\|_2 + R_\perp + 2\kappa J_{\max} B^{\parallel}. \tag{9}$$

Step 2: Bounding the projection-change term. Under the *Manifold Smoothness* assumption $\|u_{l+1}^* - u_l^*\|_2 \le \kappa$, the rank-1 perturbation bound gives $\|P_\perp^{(l+1)} - P_\perp^{(l)}\|_2 \le 2\kappa$ (proved in Appendix C). Together with $\|J_l\|_2 \le J_{\max}$ and the sub-multiplicativity of the spectral norm, the middle term satisfies $\|(P_\perp^{(l+1)} - P_\perp^{(l)}) J_l \epsilon_l\| \le 2\kappa J_{\max} \|\epsilon_l\|$.

Step 3: Assembling the recursion. Substituting into (**??**):

$$\|\epsilon_{l+1}^\perp\| \le \big(\gamma + 2\kappa J_{\max}\big)\|\epsilon_l\| + R_\perp = \gamma_{\text{eff}} \|\epsilon_l\| + R_\perp. \tag{10}$$

Step 4: Induction on depth. We prove $\|\epsilon_n^\perp\| \le \gamma_{\text{eff}}^n \|\epsilon_0\| + R_\perp/(1-\gamma_{\text{eff}})$ for all $0 \le n \le L$ by induction.

Base case ($n = 0$): $\|\epsilon_0^\perp\| \le \|\epsilon_0\| = \gamma_{\text{eff}}^0 \|\epsilon_0\|$, and since $R_\perp/(1-\gamma_{\text{eff}}) \ge 0$, the bound holds.

Inductive step: Assume the bound holds at step k. Applying recursion (10) and the inductive hypothesis:

$$\begin{aligned}
\|\epsilon_{k+1}^\perp\| &\le \gamma_{\text{eff}} \left(\gamma_{\text{eff}}^k \|\epsilon_0\| + \frac{R_\perp}{1-\gamma_{\text{eff}}} \right) + R_\perp \\
&= \gamma_{\text{eff}}^{k+1} \|\epsilon_0\| + \underbrace{\gamma_{\text{eff}} \cdot \frac{R_\perp}{1-\gamma_{\text{eff}}} + R_\perp}_{= R_\perp/(1-\gamma_{\text{eff}})}.
\end{aligned}$$

The last equality uses the *Restoring Force Algebra* identity: for $\gamma < 1$ and $R \geq 0$,

$$\gamma \cdot \frac{R}{1-\gamma} + R = \frac{R}{1-\gamma}, \tag{11}$$

which holds because $\gamma R/(1-\gamma) + R = (\gamma R + R - \gamma R)/(1-\gamma) = R/(1-\gamma)$. The induction is complete, establishing the bound for all $n \leq L$. □

B.3 CRF Loss Analysis

Theorem 2 (CRF Loss Decomposition)**.** *Under the first-order Taylor expansion:*

$$\mathcal{L}_{\text{CRF}}^{(l)} = \underbrace{\|P_\perp r_l\|_2^2}_{\textit{Transversal Drift}} + \sigma^2 \underbrace{\left\|P_\perp (J_f^{(l)} + \alpha I)\right\|_F^2}_{\textit{Spectral Energy}}. \tag{12}$$

The CRF Loss Decomposition asserts that, under the first-order Taylor expansion of the student's residual function around the teacher's trajectory point h_l^*, the Contractive Restoring Flow loss $\mathcal{L}_{\text{CRF}}^{(l)}$ defined in (5) splits exactly into a *Transversal Drift* term and a *Spectral Energy* term, with no cross terms.

Proof. Set $v = P_\perp r_l$ and $M = P_\perp (J_f^{(l)} + \alpha I)$, where $r_l = f_\theta^{(l)}(h_l^*) - \Delta_l^*$ is the zeroth-order *fitting residual*. Under the first-order expansion $f_\theta^{(l)}(h_l^* + \delta) \approx f_\theta^{(l)}(h_l^*) + J_f^{(l)}\delta$, the argument of the norm in (5) becomes:

$$P_\perp\big(f_\theta^{(l)}(h_l^* + \delta) - \Delta_l^*\big) + \alpha P_\perp \delta = P_\perp(r_l + J_f^{(l)}\delta) + \alpha P_\perp \delta = v + M\delta.$$

Hence $\mathcal{L}_{\text{CRF}}^{(l)} = \mathbb{E}_\delta[\|v + M\delta\|_2^2]$. We expand the squared norm using the standard *Norm Expansion* $\|a+b\|^2 = \|a\|^2 + 2\langle a, b\rangle + \|b\|^2$ and take expectations term by term.

Constant term. $\mathbb{E}[\|v\|^2] = \|v\|^2 = \|P_\perp r_l\|_2^2$, since v does not depend on δ.

Cross term. $\mathbb{E}[2\langle v, M\delta\rangle] = 2\langle v, M\,\mathbb{E}[\delta]\rangle = 0$, because $\delta \sim \mathcal{N}(0, \sigma^2 I)$ has zero mean.

Quadratic term. Using the trace identity $\delta^\top A \delta = \text{Tr}(A\delta\delta^\top)$:

$$\mathbb{E}[\|M\delta\|^2] = \mathbb{E}[\delta^\top M^\top M \delta] = \text{Tr}\big(M^\top M\,\mathbb{E}[\delta\delta^\top]\big) = \text{Tr}(M^\top M \cdot \sigma^2 I) = \sigma^2 \|M\|_F^2,$$

where the third equality uses the isotropic second moment $\mathbb{E}[\delta\delta^\top] = \sigma^2 I$.

Summing the three terms and substituting back $\|v\|^2 = \|P_\perp r_l\|_2^2$ and $\sigma^2\|M\|_F^2 = \sigma^2\|P_\perp(J_f^{(l)} + \alpha I)\|_F^2$ yields (12). □

Remark on the role of α. A denoising loss without the $+\alpha P_\perp \delta$ term would decompose as $\|P_\perp r_l\|_2^2 + \sigma^2\|P_\perp J_f^{(l)}\|_F^2$, driving $P_\perp J_f^{(l)} \to 0$ and thus $\|P_\perp J_l\|_2 \to 1$ — marginal stability. The shift by αI inside the *Spectral Energy* moves the minimisation target from $P_\perp J_f^{(l)} = 0$ to $P_\perp J_f^{(l)} = -\alpha P_\perp$, producing the active *Restoring Force* that the contraction chain requires.

B.4 Contraction Guarantee Chain

Theorem 3 (CRF Achieves Strict Transversal Contraction)**.** *When $\mathcal{L}_{\text{CRF}}^{(l)} = 0$ with $\sigma > 0$ and $\alpha \in (0, 1)$: $\|P_\perp^{(l)} J_l\|_2 = 1 - \alpha < 1$.*

Theorem 3 states that when $\mathcal{L}_{\text{CRF}}^{(l)} = 0$ with $\sigma > 0$ and $\alpha \in (0, 1)$, the full layer Jacobian satisfies $\|P_\perp^{(l)} J_l\|_2 = 1 - \alpha < 1$, establishing strict transversal contraction. The proof proceeds through four explicit implications.

Proof. **Step (i): Loss-zero implies both terms vanish.** The CRF Loss Decomposition (12) expresses $\mathcal{L}_{\text{CRF}}^{(l)}$ as a sum of two non-negative terms. Setting the sum to zero with $\sigma > 0$ forces each term individually to zero:

$$\|P_\perp r_l\|_2^2 = 0 \quad \text{and} \quad \sigma^2\|P_\perp(J_f^{(l)} + \alpha I)\|_F^2 = 0. \tag{13}$$

Since $\sigma^2 > 0$, the second equation gives $\|P_\perp(J_f^{(l)} + \alpha I)\|_F^2 = 0$; the Frobenius norm equals zero if and only if the matrix itself is zero, so $P_\perp(J_f^{(l)} + \alpha I) = 0$.

Step (ii): Restoring Force identity. From $P_\perp(J_f^{(l)} + \alpha I) = 0$:

$$P_\perp J_f^{(l)} + \alpha P_\perp = 0 \implies P_\perp J_f^{(l)} = -\alpha P_\perp.$$

This is the *Restoring Force* condition: the residual Jacobian, when projected onto the transversal subspace, acts as a scalar multiple of $-P_\perp$, actively pulling deviations back toward the teacher's trajectory.

Step (iii): Contraction Substitution. Substituting $P_\perp J_f^{(l)} = -\alpha P_\perp$ into Lemma 1:

$$P_\perp J_l = P_\perp + P_\perp J_f^{(l)} = P_\perp - \alpha P_\perp = (1-\alpha) P_\perp.$$

Step (iv): Contraction Norm. Using the scalar-spectral-norm identity $\|\lambda A\|_2 = |\lambda| \|A\|_2$, the unit spectral norm $\|P_\perp\|_2 = 1$ (proved in Appendix C), and $0 < \alpha < 1$:

$$\|(1-\alpha)P_\perp\|_2 = (1-\alpha)\|P_\perp\|_2 = 1 - \alpha < 1. \qquad \square$$

Proposition 1 (Tangential Independence)**.** *The loss $\mathcal{L}_{\mathrm{CRF}}^{(l)}$ imposes no constraint on the tangential component $(u_l^*)^\top J_l\, u_l^*$. Equivalently, for any $\beta \in \mathbb{R}$, replacing $J_f^{(l)}$ by $J_f^{(l)} + \beta\, u_l^* (u_l^*)^\top$ leaves the value of $\mathcal{L}_{\mathrm{CRF}}^{(l)}$ unchanged. The minimum-norm solution at $\mathcal{L}_{\mathrm{CRF}}^{(l)} = 0$ takes $J_f^{(l)} u_l^* = 0$, in which case $(u_l^*)^\top J_l u_l^* = 1$, but the loss does not select this solution among the family of minimisers.*

Proof. From Theorem 3, Step (ii), the loss decomposition contains $J_f^{(l)}$ only through the matrix $P_\perp J_f^{(l)}$. For any $\beta \in \mathbb{R}$,

$$P_\perp\big(J_f^{(l)} + \beta\, u_l^* (u_l^*)^\top\big) = P_\perp J_f^{(l)} + \beta\, (P_\perp u_l^*)(u_l^*)^\top = P_\perp J_f^{(l)},$$

where the second equality uses the *Annihilation* property $P_\perp u_l^* = 0$ (Appendix C). Thus the modified Jacobian achieves the same loss value, and the tangential component is genuinely unconstrained by the loss.

To see that $J_f^{(l)} u_l^* = 0$ is one (minimum-norm) solution at $\mathcal{L}_{\mathrm{CRF}}^{(l)} = 0$: by Theorem 3 Step (ii), $P_\perp J_f^{(l)} = -\alpha P_\perp$, so applying both sides to u_l^* gives $P_\perp(J_f^{(l)} u_l^*) = 0$, i.e. $J_f^{(l)} u_l^* \in \ker(P_\perp) = \mathrm{span}(u_l^*)$. Hence $J_f^{(l)} u_l^* = \beta\, u_l^*$ for some $\beta \in \mathbb{R}$, and the choice $\beta = 0$ (which is selected by, e.g., a Frobenius-norm regulariser on $J_f^{(l)}$) yields $J_l u_l^* = u_l^* + 0 = u_l^*$ and therefore $(u_l^*)^\top J_l u_l^* = 1$. The value of β is fixed in practice by the SFT objective rather than by $\mathcal{L}_{\mathrm{CRF}}^{(l)}$. $\square$

Remark on Tangential Independence. The CRF Loss is *tangentially agnostic*: it constrains only the transversal dynamics, leaving the student's behaviour along the reasoning direction entirely unconstrained. This is a deliberate design choice—the geometric loss should not interfere with the student's ability to follow the reasoning trajectory at its own pace, and the SFT objective shapes the tangential dynamics through the standard cross-entropy signal. The empirical effect is consistent with a quasi-identity tangential action in well-trained models, but this is an outcome of joint $\mathcal{L}_{\mathrm{SFT}} + \lambda \mathcal{L}_{\mathrm{CRF}}$ optimisation, not of $\mathcal{L}_{\mathrm{CRF}}$ alone.

Proposition 2 (Approximate Contraction)**.** *If $\mathcal{L}_{\mathrm{CRF}}^{(l)} \le \varepsilon$ with $\sigma > 0$: $\|P_\perp J_l\|_2 \le (1-\alpha) + \sqrt{\varepsilon}/\sigma$.*

Proposition 2 states that if the CRF loss is only approximately minimised — $\mathcal{L}_{\mathrm{CRF}}^{(l)} \le \varepsilon$ with $\sigma > 0$ — the contraction rate degrades continuously: $\|P_\perp J_l\|_2 \le (1-\alpha) + \sqrt{\varepsilon}/\sigma$.

Proof. From the CRF Loss Decomposition (12), both non-negative terms are bounded above by the total loss:

$$\sigma^2 \|P_\perp(J_f^{(l)} + \alpha I)\|_F^2 \le \varepsilon.$$

Dividing by $\sigma^2 > 0$ and taking the square root: $\|P_\perp(J_f^{(l)} + \alpha I)\|_F \leq \sqrt{\varepsilon}/\sigma$. Since the spectral norm is dominated by the Frobenius norm, $\|P_\perp(J_f^{(l)} + \alpha I)\|_2 \leq \|P_\perp(J_f^{(l)} + \alpha I)\|_F \leq \sqrt{\varepsilon}/\sigma$.

We now reverse the *Contraction Substitution* step (Step (iii) of Theorem 3) to decompose $P_\perp J_l$:

$$P_\perp J_l = P_\perp + P_\perp J_f^{(l)} = (1-\alpha)P_\perp + P_\perp(J_f^{(l)} + \alpha I).$$

Applying the triangle inequality for spectral norms and the *Contraction Norm* result $\|(1-\alpha)P_\perp\|_2 = 1-\alpha$:

$$\|P_\perp J_l\|_2 \leq (1-\alpha) + \|P_\perp(J_f^{(l)} + \alpha I)\|_2 \leq (1-\alpha) + \frac{\sqrt{\varepsilon}}{\sigma}. \qquad \square$$

Degradation–noise trade-off. The bound shows that contraction degrades at rate $1/\sigma$: doubling the probe noise σ halves the slack $\sqrt{\varepsilon}/\sigma$ for the same training loss ε, demanding proportionally tighter Jacobian control. This motivates the recommended probe scale $\sigma \in [0.005, 0.02]$, balancing Taylor accuracy (requiring $\sigma\|J_f^{(l)}\|_2 \ll 1$) against contraction precision.

B.5 Global Basin of Attraction

Theorem 4 (Basin of Attraction)**.** *Under manifold smoothness and $\alpha > 2\kappa J_{\max}$, with CRF achieving $\gamma = 1-\alpha$ on all active layers:* (i) *exponential contraction;* (ii) *steady-state error $\leq R_\perp/(\alpha - 2\kappa J_{\max})$;* (iii) *output bound $\|\ell(h_L) - \ell(h_L^*)\|_2 \leq K(\cdots)$, $K = \|W_u\|_2$;* (iv) *global basin for all $\|\epsilon_0\|$ within the linearisation regime.*

Theorem 4 synthesises the single-layer contraction result (Theorem 3) and the multi-layer accumulation bound (Theorem 1) into a global convergence guarantee for the student's reasoning trajectory. The four parts are proved sequentially, each building on the previous.

Proof. **Part (i): Exponential contraction.** Setting $\gamma = 1 - \alpha$ in Theorem 1 and noting that $\gamma_{\text{eff}} = \gamma + 2\kappa J_{\max} = (1-\alpha) + 2\kappa J_{\max}$:

$$\|\epsilon_L^\perp\| \leq \gamma_{\text{eff}}^L \|\epsilon_0^\perp\| + \frac{R_\perp}{1 - \gamma_{\text{eff}}}.$$

The condition $\gamma_{\text{eff}} < 1$ holds if and only if $\alpha > 2\kappa J_{\max}$, which is the stability condition stated in the theorem.

Part (ii): Steady-state error. As $L \to \infty$, $\gamma_{\text{eff}}^L \to 0$ since $\gamma_{\text{eff}} \in [0, 1)$. The transient term vanishes and the bound converges to the *steady-state error*:

$$\|\epsilon_\infty^\perp\| \leq \frac{R_\perp}{1 - \gamma_{\text{eff}}} = \frac{R_\perp}{\alpha - 2\kappa J_{\max}},$$

which is finite and non-negative since $\alpha > 2\kappa J_{\max}$ and $R_\perp \geq 0$. The steady-state error decreases with increasing α and with decreasing $R_\perp$; since $\mathcal{L}_{\text{CRF}}^{(l)} \leq \varepsilon$ implies $R_\perp \leq \sqrt{\varepsilon}$ (from Theorem 2), better training quality directly reduces the floor.

Part (iii): Output bound. By the spectral-norm bound on the unembedding matrix W_u:

$$\|\ell(h_L) - \ell(h_L^*)\|_2 \leq \|W_u\|_2 \, \|\epsilon_L\|_2 = K \, \|\epsilon_L\|_2,$$

where $K = \|W_u\|_2$. Applying Part (i) and the stability condition:

$$\|\ell(h_L) - \ell(h_L^*)\|_2 \leq K\left(\gamma_{\text{eff}}^L \|\epsilon_0\| + \frac{R_\perp}{\alpha - 2\kappa J_{\max}}\right).$$

Part (iv): Global basin. For any finite initial deviation $\|\epsilon_0\|$ within the linearisation regime of the first-order expansion (3), $\gamma_{\text{eff}}^L \|\epsilon_0\| \to 0$ as $L \to \infty$ because $\gamma_{\text{eff}} \in [0, 1)$. There is no threshold on $\|\epsilon_0\|$ below which the bound fails: the attraction basin coincides with the entire linearisation neighbourhood of the teacher trajectory, which is the formal sense in which it is a *global* basin of attraction. $\square$

Quantitative illustration. For a 32-layer Transformer with $\kappa J_{\max} \approx 0$ (negligible manifold curvature), $\gamma_{\text{eff}} = 1 - \alpha$, and the transient decay factor γ_{eff}^{32} evaluates as follows:

α	$\gamma_{\text{eff}} = 1 - \alpha$	γ_{eff}^{32}
0.05	0.95	0.1936
0.10	0.90	0.0343
0.20	0.80	0.0013

Even a modest contraction coefficient $\alpha = 0.10$ reduces the transient error to 3.4% of its initial value after propagating through all 32 layers.

C Projection Properties

The transversal projection $P_\perp^{(l)} = I - u_l^*(u_l^*)^\top$, defined in Equation (4) from the normalised *Residual Flux* $u_l^* = \Delta_l^*/\|\Delta_l^*\|_2$, satisfies four structural properties that are invoked repeatedly in the proofs above. We state and derive all four here from the single hypothesis $\|u_l^*\|_2 = 1$.

Proposition 3 (Structural Properties of the Transversal Projection)**.** *For any unit vector $u \in \mathbb{R}^d$ with $\|u\|_2 = 1$, the matrix $P_\perp = I - uu^\top$ satisfies:*

(i) ***Idempotence**: $P_\perp^2 = P_\perp$.*

(ii) ***Symmetry**: $P_\perp^\top = P_\perp$.*

(iii) ***Annihilation**: $P_\perp u = 0$.*

(iv) ***Unit Spectral Norm**: $\|P_\perp\|_2 = 1$.*

Proof. (i) Idempotence.

$$P_\perp^2 = (I - uu^\top)^2 = I - 2uu^\top + u \underbrace{u^\top u}_{=\|u\|_2^2 = 1} u^\top = I - uu^\top = P_\perp.$$

The unit-norm hypothesis is used precisely here to simplify $u(u^\top u)u^\top = uu^\top$.

(ii) Symmetry. $(I - uu^\top)^\top = I^\top - (uu^\top)^\top = I - uu^\top = P_\perp$, since transposing a symmetric rank-1 outer product gives $(uu^\top)^\top = uu^\top$.

(iii) Annihilation. $P_\perp u = (I - uu^\top)u = u - u\underbrace{u^\top u}_{=1} = u - u = 0$.

(iv) Unit Spectral Norm. Because $P_\perp$ is symmetric (ii) and idempotent (i), its eigenvalues lie in $\{0, 1\}$. The eigenvalue 1 is attained by any non-zero vector v with $v^\top u = 0$ (i.e., any v in the orthogonal complement of u), since $P_\perp v = v - u(u^\top v) = v$. Therefore $\|P_\perp\|_2 = \sup_{\|v\|=1} \|P_\perp v\|_2 = 1$. □

Projection Continuity (used in Theorem 1). Under the *Manifold Smoothness* assumption $\|u_{l+1}^* - u_l^*\|_2 \le \kappa$, the projection changes by at most 2κ in spectral norm between adjacent layers.

Proof. Writing $P_\perp^{(l)} = I - u_l^*(u_l^*)^\top$:

$$P_\perp^{(l+1)} - P_\perp^{(l)} = u_l^*(u_l^*)^\top - u_{l+1}^*(u_{l+1}^*)^\top.$$

For unit vectors u, v, the rank-1 perturbation bound gives:

$$\begin{aligned}\|uu^\top - vv^\top\|_2 &\le \|u(u-v)^\top\|_2 + \|(u-v)v^\top\|_2 \\ &\le \|u\|_2\|u-v\|_2 + \|u-v\|_2\|v\|_2 = 2\|u-v\|_2,\end{aligned}$$

where the last step uses $\|u\|_2 = \|v\|_2 = 1$. Applying this with $u = u_l^*$, $v = u_{l+1}^*$ yields $\|P_\perp^{(l+1)} - P_\perp^{(l)}\|_2 \le 2\kappa$. □

D Local Regularity of Transformer Residual Blocks

The expansion (14) stated below states that, under the condition $\sigma\|J_f^{(l)}\|_2 \ll 1$, the transversal component of the student's response to a Gaussian probe $\delta \sim \mathcal{N}(0, \sigma^2 I)$ satisfies

$$P_\perp\big(f_\theta^{(l)}(h_l^* + \delta) - \Delta_l^*\big) \approx P_\perp r_l + P_\perp J_f^{(l)}\,\delta, \tag{14}$$

and that modern transformers with RMSNorm and SwiGLU activations (LLaMA / Qwen2.5 / Mistral family) satisfy a Local Regularity Assumption under which the approximation error is $O(\sigma^2)$. This appendix formalises that assumption, establishes it from the architectural properties of the transformer, and derives the $O(\sigma^2)$ bound explicitly.

D.1 The Local Regularity Assumption

Assumption 1 (Local Regularity)**.** *There exist constants $L > 0$, $M > 0$, and a radius $R_\delta \geq 3\sigma$ such that the residual function $f_\theta^{(l)} : \mathbb{R}^d \to \mathbb{R}^d$ satisfies:*

(i) ***Local Lipschitz continuity.*** $\|f_\theta^{(l)}(h_1) - f_\theta^{(l)}(h_2)\|_2 \leq L\,\|h_1 - h_2\|_2$ *for all* $h_1, h_2 \in B(h_l^*, R_\delta)$.

(ii) ***Bounded Hessian.*** $\|H_{f_\theta^{(l)}}(h)\|_{\mathrm{op}} \leq M$ *for all* $h \in B(h_l^*, R_\delta)$*, where H denotes the Hessian and $\|\cdot\|_{\mathrm{op}}$ the operator norm.*

The radius $R_\delta \geq 3\sigma$ is chosen so that the Gaussian probe δ satisfies $\|\delta\|_2 \leq R_\delta$ with probability at least $1 - e^{-d/2}$ by standard Gaussian concentration, ensuring that $h_l^* + \delta \in B(h_l^*, R_\delta)$ with overwhelming probability. We now establish that Assumption 1 holds for Pre-LN Transformer residual blocks with RMSNorm and SwiGLU activations.

D.2 RMSNorm Constrains Hidden States to a Bounded Set

The first step is to establish that the input to each attention sublayer lies in a bounded region, which makes the local Lipschitz constant of attention finite and parameter-dependent rather than input-dependent. Modern Pre-LN Transformers in the LLaMA / Qwen2.5 / Mistral family replace LayerNorm with Root Mean Square Layer Normalisation [Zhang and Sennrich, 2019]:

$$\mathrm{RMSNorm}(x) = \frac{x}{\sqrt{\frac{1}{d}\sum_{i=1}^{d} x_i^2}} \odot \gamma = \frac{\sqrt{d}\,x}{\|x\|_2} \odot \gamma, \tag{15}$$

where $\gamma \in \mathbb{R}^d$ is a learnable per-channel scale and $\odot$ denotes element-wise multiplication. RMSNorm differs from LayerNorm by omitting the mean-subtraction step; it preserves the magnitude-normalisation step exactly.

Denote the standardised vector before γ-scaling by $\hat{x} = \sqrt{d}\,x/\|x\|_2$. For any input $x \neq 0$, $\|\hat{x}\|_2 = \sqrt{d}$, so $\hat{x}$ lies on the $(d-1)$-dimensional sphere $\mathcal{S}^{d-1}(\sqrt{d})$. The subsequent element-wise scaling by γ maps this sphere to a hyperellipsoid bounded by $\|\hat{x} \odot \gamma\|_2 \leq \sqrt{d}\,\|\gamma\|_\infty$. In particular, the input to each self-attention sublayer lies in a compact ball B_R^n of radius $R = \sqrt{d}\,\|\gamma\|_\infty$, where n is the sequence length and d is the hidden dimension. Compared with the LayerNorm result of Brody et al. [2023], RMSNorm yields a sphere of one higher dimension (no mean removed), but the radius and the consequent boundedness conclusion are identical.

Two architectural notes. First, RoPE [Su et al., 2024] applies a position-dependent block-diagonal rotation to query and key vectors before the attention dot product. Rotations are isometries, so RoPE preserves ℓ_2-norm and therefore preserves the bounded-ball property; the analysis of self-attention below applies verbatim to RoPE-augmented attention. Second, optional QKV bias terms (used in some Qwen2.5 variants) enter as constant additive translations and shift R by at most $\|b\|_2$; this does not break compactness.

D.3 Self-Attention Is Locally Lipschitz on the Bounded Set

Standard dot-product self-attention is not globally Lipschitz: Kim et al. [2021] prove in **Theorem 3.1** that dot-product multi-head attention (DP-MHA) is not Lipschitz for any vector p-norm with $p \in$

$[1, \infty]$, because the Jacobian norm grows with the variance of the input tokens. However, because RMSNorm constrains inputs to the compact ball B_R^n established above, the global Lipschitz failure is irrelevant: we only need the *local* Lipschitz constant on B_R^n.

Castin et al. [2024] establish exactly this in **Theorem 3.3**: for self-attention with parameters (A, V) where $A = K^\top Q/\sqrt{k}$, restricted to inputs in B_R^n, the Lipschitz constant satisfies

$$\mathrm{Lip}\big(f_{\mathrm{attn}}\big|_{B_R^n}\big) \;\le\; \sqrt{3}\,\|V\|_2\Big(\|A\|_2^2 R^4(4n+1)+n\Big)^{1/2}. \tag{16}$$

Since R is determined by γ via (15), the bound (16) depends only on network parameters (V, A, γ) and sequence length n, not on the input. This establishes condition *(i)* of Assumption 1 for the attention sublayer, with $L_{\mathrm{attn}} = \sqrt{3}\|V\|_2(\|A\|_2^2 R^4(4n+1)+n)^{1/2}$.

D.4 SwiGLU Activations Are C^∞ with Bounded Derivatives

The feed-forward sublayer in LLaMA / Qwen2.5-family residual blocks applies a SwiGLU gate [Shazeer, 2020]:

$$\mathrm{SwiGLU}(x) = \mathrm{SiLU}(xW_1 + b_1) \odot (xW_2 + b_2), \qquad \mathrm{SiLU}(z) = z\,\sigma(z), \tag{17}$$

where $\sigma(z) = 1/(1+e^{-z})$ is the logistic sigmoid, $\odot$ is element-wise multiplication, and W_1, W_2 are gate and up-projection weights. We establish C^∞ regularity with bounded derivatives in three steps.

Step 1: σ is C^∞ with globally bounded derivatives. Since $\sigma(z) = 1/(1+e^{-z})$ composes the entire exponential with affine and rational maps that avoid singularities for all $z \in \mathbb{R}$, σ is C^∞. The recurrence $\sigma'(z) = \sigma(z)(1-\sigma(z))$ together with induction shows that for every $k \ge 1$,

$$\sigma^{(k)}(z) = \sigma(z)\big(1-\sigma(z)\big)\,P_k\big(\sigma(z)\big), \tag{18}$$

where P_k is a polynomial of degree $k-1$ with bounded integer coefficients. Since $\sigma(z) \in (0,1)$ for all z and $\sigma(z)(1-\sigma(z)) \le 1/4$, every $\sigma^{(k)}$ is bounded on $\mathbb{R}$, with $\|\sigma^{(k)}\|_\infty \le C_k$ for constants C_k depending only on k (e.g. $C_1 = 1/4$).

Step 2: SiLU is C^∞ with globally bounded derivatives. By Leibniz's rule on $\mathrm{SiLU}(z) = z\,\sigma(z)$,

$$\mathrm{SiLU}^{(k)}(z) = z\,\sigma^{(k)}(z) + k\,\sigma^{(k-1)}(z). \tag{19}$$

The factor $z\,\sigma^{(k)}(z)$ would be unbounded as a product of an unbounded z and a merely bounded $\sigma^{(k)}$, but Eq. (18) shows $\sigma^{(k)}$ contains $\sigma(z)(1-\sigma(z))$ as a factor for every $k \ge 1$. A direct calculation gives

$$\sigma(z)\big(1-\sigma(z)\big) = \frac{e^{-|z|}}{(1+e^{-|z|})^2} \;=\; O\big(e^{-|z|}\big) \quad \text{as } |z| \to \infty,$$

so $z \cdot \sigma^{(k)}(z) \to 0$ exponentially for every $k \ge 1$. Therefore each $\mathrm{SiLU}^{(k)}$ ($k \ge 1$) is bounded on $\mathbb{R}$.

Step 3: SwiGLU is C^∞ with bounded derivatives on the bounded input domain. Let $u = \mathrm{SiLU}(xW_1 + b_1)$ and $v = xW_2 + b_2$. SwiGLU is the element-wise product $u \odot v$. Since linear maps and SiLU are C^∞ on $\mathbb{R}^d$, and element-wise products of C^∞ functions are C^∞, SwiGLU is C^∞. By Leibniz's rule for element-wise products,

$$\partial_x^k[u \odot v] = \sum_{j=0}^{k} \binom{k}{j} (\partial_x^j u) \odot (\partial_x^{k-j} v). \tag{20}$$

Restricted to the RMSNorm-induced compact input ball B_R^n (Eq. (15)), $\|u\|_\infty$ and $\|v\|_\infty$ are bounded by parameter-only quantities, and so are all their derivatives by Step 2. Composition with the down-projection W_3 preserves boundedness. Therefore SwiGLU has bounded derivatives of all orders on B_R^n, satisfying condition *(ii)* of Assumption 1 with a finite Hessian bound M depending only on $\|W_1\|_2$, $\|W_2\|_2$, $\|W_3\|_2$, $\|b_1\|_2$, $\|b_2\|_2$, and the RMSNorm radius R.

Combined with the attention sublayer's local Lipschitz continuity established in (16), the full residual function $f_\theta^{(l)} = f_{\mathrm{FFN}} \circ f_{\mathrm{attn}}$ inherits C^∞ regularity and a bounded Hessian on $B(h_l^*, R_\delta)$, establishing condition *(ii)* of Assumption 1 with $M = \sup_{h \in B(h_l^*, R_\delta)} \|H_{f_\theta^{(l)}}(h)\|_{\mathrm{op}} < \infty$.

D.5 The Taylor Remainder Is $O(\sigma^2)$

With Assumption 1 established, we can now bound the approximation error in (14). By **Taylor's theorem with Lagrange remainder** applied to the C^2 function $f_\theta^{(l)}$ at the point h_l^*:

$$f_\theta^{(l)}(h_l^* + \delta) = f_\theta^{(l)}(h_l^*) + J_f^{(l)}\,\delta + \mathcal{R}(\delta), \tag{21}$$

where the Lagrange remainder satisfies

$$\|\mathcal{R}(\delta)\|_2 \le \frac{M}{2}\,\|\delta\|_2^2. \tag{22}$$

Applying the projection $P_\perp$ to (21) and using $\|P_\perp\|_2 = 1$:

$$P_\perp\big(f_\theta^{(l)}(h_l^* + \delta) - \Delta_l^*\big) = P_\perp r_l + P_\perp J_f^{(l)}\,\delta + P_\perp \mathcal{R}(\delta), \tag{23}$$

where $\|P_\perp \mathcal{R}(\delta)\|_2 \le (M/2)\|\delta\|_2^2$. Taking the expectation over $\delta \sim \mathcal{N}(0, \sigma^2 I)$:

$$\mathbb{E}_\delta\big[\|P_\perp \mathcal{R}(\delta)\|_2^2\big] \le \frac{M^2}{4}\,\mathbb{E}_\delta\big[\|\delta\|_2^4\big] = \frac{M^2}{4}\,\sigma^4\,d(d+2), \tag{24}$$

using $\mathbb{E}[\|\delta\|_2^4] = \sigma^4 d(d+2)$ for isotropic Gaussian $\delta \in \mathbb{R}^d$. Therefore the root-mean-squared remainder is $\mathbb{E}[\|P_\perp \mathcal{R}\|_2^2]^{1/2} \le (M/2)\sigma^2\sqrt{d(d+2)} \approx (M/2)\sigma^2 d$. Comparing to the first-order term $\mathbb{E}[\|P_\perp J_f^{(l)}\delta\|_2^2]^{1/2} \approx \sigma\,\|P_\perp J_f^{(l)}\|_F$, the linearisation is dominant whenever

$$\sigma \;\ll\; \frac{2\,\|P_\perp J_f^{(l)}\|_F}{M\,d}\,, \tag{25}$$

i.e. the relative error scales as $O(\sigma\,d)$ rather than $O(\sigma)$ in absolute terms; this d-dependence is the price of working in a high-dimensional hidden space and constrains the recommended probe scale to $\sigma \lesssim 1/(M\sqrt{d})$ when $\|P_\perp J_f^{(l)}\|_F$ scales as $O(\sqrt{d})$ (typical for trained Transformers). For the student dimensions in our experiments ($d \in \{896, 2048, 3584, 8192\}$), this gives upper bounds $\sigma \lesssim \{0.033, 0.022, 0.017, 0.011\}$ respectively under $M = 1$; the recommended range $\sigma \in [0.005, 0.02]$ is well inside the validity region for $d \le 3584$ and at the boundary for $d = 8192$ (the largest student tested).

Proposition 4 (Taylor Validity)**.** *Under Assumption 1 with $M > 0$, the approximation error in (14) satisfies*

$$\mathbb{E}_\delta\big[\|P_\perp \mathcal{R}(\delta)\|_2^2\big]^{1/2} \;\le\; \frac{M\,\sigma^2}{2}\sqrt{d(d+2)}. \tag{26}$$

As $\sigma \to 0$, the CRF loss approaches the projected fitting-residual term. In the first-order model, the bound on the Jacobian discrepancy tends to zero when $\varepsilon/\sigma^2 \to 0$. The probe scale and the achieved loss must therefore be considered jointly.

Proof. Immediate from (22)–(24) and $\|P_\perp\|_2 = 1$. The limiting statement follows because the Spectral Energy term in the decomposition (12) vanishes as $\sigma \to 0$ ($\sigma^2\|P_\perp(J_f^{(l)} + \alpha I)\|_F^2 \to 0$), and $\mathcal{L}_{\mathrm{CRF}}^{(l)} \to \|P_\perp r_l\|_2^2$. □

D.6 Summary

The chain of reasoning above establishes that Assumption 1 holds for any Pre-LN Transformer using RMSNorm and SwiGLU activations: RMSNorm projects hidden states onto a sphere of radius $\sqrt{d}$ (Eq. (15)), bounding the domain for Castin et al.'s [2024] local Lipschitz bound (Theorem 3.3); SwiGLU's C^∞ smoothness with globally bounded SiLU-derivatives (Eq. (19)) ensures a bounded Hessian on every compact input neighbourhood; RoPE is an isometry and preserves boundedness without further argument. Proposition 4 then shows that the first-order approximation (14) incurs an error of only $O(\sigma^2)$, with the bound depending on the Hessian constant M and ambient dimension d. The argument is structurally identical to the LayerNorm + GELU case treated in earlier work; we have shown the same conclusion holds for the architecture used in our experiments (Qwen2.5 family: RMSNorm + SwiGLU + RoPE), so the theoretical guarantees of Section 3 apply without qualification

on the actual models studied. In practice, the probe scale $\sigma \in [0.005, 0.02]$ (Section I) satisfies $\sigma^2 d(d+2) \ll 1$ for the hidden dimensions $d \in \{896, 2048, 3584, 8192\}$ used by the Qwen2.5 students and teachers.

Theorem 3 constrains the transversal Jacobian at the teacher's trajectory point h_l^*; at inference time, the student operates at its own hidden state h_l^S, which may deviate from h_l^* by the accumulated fitting error $\delta_{\mathrm{SFT}} = \|h_l^S - h_l^*\|_2$. This section proves that a single design condition on α guarantees stability at both training and inference time simultaneously.

Proposition 5 (Inference Gap)**.** *If $J_f^{(l)}$ is $H_{\max}$-Lipschitz as a function of the hidden state, the inference-time transversal contraction rate satisfies:*

$$\|P_\perp J_l^S\|_2 \le (1-\alpha) + H_{\max}\,\delta_{\mathrm{SFT}}, \tag{27}$$

where $J_l^S = (I + J_f^{(l)}(h_l^S))$ is the layer Jacobian at the student's operating point.

Proof. Decomposing $P_\perp J_l^S$ around the training point:

$$\begin{aligned}\|P_\perp J_l^S\|_2 &\le \|P_\perp(I + J_f^{(l)}(h_l^*))\|_2 + \|P_\perp\|_2\,\|J_f^{(l)}(h_l^S) - J_f^{(l)}(h_l^*)\|_2 \\ &\le (1-\alpha) + 1 \cdot H_{\max}\,\|h_l^S - h_l^*\|_2 = (1-\alpha) + H_{\max}\,\delta_{\mathrm{SFT}},\end{aligned}$$

where the first term uses Theorem 3 and the Unit Spectral Norm $\|P_\perp\|_2 = 1$. □

Corollary 1 (Unified Stability Condition)**.** *If $\alpha > \max(2\kappa J_{\max},\ H_{\max}\cdot\delta_{\mathrm{SFT}})$, then both the training-time and inference-time transversal contraction rates are strictly less than one: $(1-\alpha)+2\kappa J_{\max} < 1$ and $(1-\alpha) + H_{\max}\cdot\delta_{\mathrm{SFT}} < 1$.*

Proof. Both inequalities follow immediately from $\alpha > 2\kappa J_{\max}$ and $\alpha > H_{\max}\cdot\delta_{\mathrm{SFT}}$ respectively, which together constitute Equation (7). □

Practical implication. As the SFT component of the loss drives $\delta_{\mathrm{SFT}} \to 0$, the inference-time contraction rate approaches the training-time rate $(1-\alpha)$ from above. The two objectives are thus mutually reinforcing: better SFT fitting reduces δ_{SFT}, relaxing the requirement on α and allowing a smaller, less intrusive restoring force.

E Cross-Scale Alignment Implementation

This appendix details the layer mapping π and the per-layer alignments $\{A_l\}$ introduced in Sec 2.2 that bridge teacher hidden space $\mathbb{R}^{d_T}$ and student hidden space $\mathbb{R}^{d_S}$ when the two differ in dimension or depth.

Layer mapping. We use a uniform mapping $\pi(l) = \lfloor l \cdot L_T/L_S \rfloor$ for $l \in \{0, \ldots, L_S\}$, distributing the teacher's L_T layers evenly across the student's L_S slots while preserving depth ordering. For Qwen2.5-72B (80 layers) into Qwen2.5-7B (28 layers), π maps student layer l to teacher layer $\lfloor 80l/28 \rfloor$. Non-uniform mappings (concentrating mass at deeper teacher layers) were also tested and offered no measurable improvement on AIME / MBPP / LCB.

Per-layer alignments. Each $A_l : \mathbb{R}^{d_T} \to \mathbb{R}^{d_S}$ is a learnable $d_S \times d_T$ matrix. We initialise A_l as the rank-d_S truncation of the SVD of the teacher hidden-state covariance at layer $\pi(l)$ (estimated on a small held-out probe set), so the initial alignment captures the principal directions of teacher variation; when $d_T = d_S$ we use identity initialisation. The alignments are then trained jointly with the student parameters θ under the combined objective $\mathcal{L}_{\mathrm{SFT}} + \lambda\mathcal{L}_{\mathrm{CRF}}$; gradients flow through A_l via both the trajectory-matching component of the SFT objective and the h_l^*, Δ_l^* terms in the CRF loss. No separate alignment loss is required.

Why a linear alignment is sufficient. RMSNorm at every residual block (Appendix D) constrains teacher and student hidden states to spheres of radii $\sqrt{d_T}$ and $\sqrt{d_S}$ respectively, leaving directional structure as the only meaningful object to align; linear maps between spheres preserve this structure. We tested replacing A_l with a 2-layer MLP (extra 8% parameters) and observed no measurable gain on the benchmarks of Section 4.

Compute and parameter overhead. The full $\{A_l\}$ adds $L_S \cdot d_T \cdot d_S$ parameters (e.g. $28 \cdot 8192 \cdot 3584 \approx 0.82$B for Qwen2.5-72B $\to$ Qwen2.5-7B), increasing student-side training-time parameter count by approximately 12%. Forward and backward passes through A_l contribute a single matrix multiply per layer, negligible relative to the residual block compute itself. **At inference, $\{A_l\}$ is discarded entirely**: the deployed student is identical in size and computation to a baseline 7B model.

Identifiability of $\{A_l\}$. A natural concern is that jointly learning $\{A_l\}$ without an explicit alignment loss could admit degenerate minimisers in which A_l collapses the teacher trajectory onto a low-dimensional subspace where the CRF loss is trivially small but the geometric structure is destroyed. Three forces in our setup prevent this: *(i) SVD initialisation.* Initialising A_l as the rank-d_S truncation of the teacher hidden-state covariance starts the alignment in an information-preserving regime; gradient-based collapse from this starting point requires non-trivial movement against the SFT gradient. *(ii) SFT supervision provides an external anchor.* The trajectory-matching component of $\mathcal{L}_{\mathrm{SFT}}$ requires that the aligned teacher state $A_l h^{T,*}_{\pi(l)}$ be predictive of student tokens; an A_l that destroys teacher information makes the SFT loss diverge, so the optimum is jointly constrained by SFT and CRF rather than by CRF alone. *(iii) Per-layer linearity.* Because each A_l is linear, degenerate solutions correspond to rank-deficient maps, which manifest as low-rank $\{A_l A_l^\top\}$ that we monitor during training; we observed full-rank alignments throughout in all training runs. A cleaner alternative is to add a Procrustes-style regulariser $\lambda_{\mathrm{ali}} \|A_l^\top A_l - I_{d_S}\|_F^2$ that penalises non-orthogonal alignments explicitly; we recommend it for future cross-architecture work where teacher and student differ in hidden dimension by more than $4\times$.

F Kinetic Gating: Formal Analysis of Passive Layers

Section 3.3 introduces *Kinetic Gating* as a mechanism to exclude passive layers—where the *Residual Flux* $\mu_l = \|\Delta_l^*\|_2 \approx 0$—from the CRF loss. This section provides the formal justification.

Passive Layer Neutrality. On a passive layer where $f_\theta^{(l)} \approx 0$, the full layer map satisfies $T_\theta^{(l)} \approx \mathrm{id}$ and $J_l \approx I$. The transversal Jacobian norm is:

$$\|P_\perp J_l\|_2 \approx \|P_\perp I\|_2 = \|P_\perp\|_2 = 1.$$

Passive layers contribute a factor of 1 to the overall contraction product; they neither amplify nor suppress transversal deviations. The total contraction over L_{active} active layers and L_{passive} passive layers is therefore:

$$(1-\alpha)^{L_{\mathrm{active}}} \cdot 1^{L_{\mathrm{passive}}} = (1-\alpha)^{L_{\mathrm{active}}},$$

confirming that passive layers are transparent to the basin-of-attraction guarantee in Theorem 4.

Why applying CRF to passive layers is harmful. When $\mu_l = 0$, the Tangent Direction $u_l^* = \Delta_l^*/\mu_l$ is undefined. Any numerical implementation must assign an arbitrary unit vector, so the Transversal Projection $P_\perp^{(l)}$ projects onto a geometrically meaningless subspace. Applying the CRF loss in this regime forces the student to produce a Restoring Force against a numerically noisy direction, imposing a spurious constraint that competes with $\mathcal{L}_{\mathrm{SFT}}$ and provides no contraction benefit. The Kinetic Gating weight $w_l = \sigma_{\mathrm{gate}}((\mu_l - \tau)/\varepsilon_{\mathrm{gate}})$ suppresses this contribution continuously as $\mu_l \to 0$, avoiding the singularity without introducing a discontinuity in the gradient.

Properties of the Kinetic Gating function. The gating function w_l defined in (8) satisfies:

(i) **Differentiability.** w_l is C^∞ everywhere, so the gated loss $\mathcal{L}_{\mathrm{CRF}} = \sum_l w_l \mathcal{L}_{\mathrm{CRF}}^{(l)}$ admits well-defined gradients with respect to θ at all training steps.

(ii) **Monotonicity.** w_l is strictly increasing in μ_l, concentrating the geometric constraint on layers where the teacher is actively computing and the Tangent Direction u_l^* is well-posed.

(iii) **Silent Zone / Active Zone.** For $\mu_l \ll \tau$, $w_l \approx 0$ (silent zone, passive layers skipped); for $\mu_l \gg \tau$, $w_l \approx 1$ (active zone, full CRF loss applied). The transition width $\varepsilon_{\mathrm{gate}} = \tau/5$ places the switch in a narrow band around the flux threshold τ, chosen as the 5th percentile of $\{\mu_l\}$ to exclude clearly passive layers while preserving the active ones.

F.1 Practical Parameter Selection

The three parameters governing the gate—flux threshold τ, transition width $\varepsilon_{\text{gate}}$, and activation threshold ε_{act}—jointly determine which layers receive the CRF constraint. We give concrete guidance for each.

Setting τ: the flux threshold. We recommend $\tau = \mathrm{P}_5\{\mu_l\}$, the 5th percentile of the residual-flux magnitudes $\{\mu_{l,t}\}$ measured over a small calibration pass ($\sim$200 prompts) of the teacher. The rationale is twofold. First, layers with $\mu_l < \tau$ have residual functions that contribute negligibly to the teacher's hidden-state update; the tangent direction u_l^* is numerically unreliable and the CRF constraint would penalise a geometrically meaningless direction (Section F). Second, using a data-dependent percentile rather than a fixed absolute threshold makes τ scale automatically across architectures: a 0.5B model typically has $\mu_l \in [0.1, 3.0]$, while a 7B model may have $\mu_l \in [0.5, 8.0]$; a fixed threshold of, say, 0.5 would over-gate the small model and under-gate the large one.

Setting $\varepsilon_{\text{gate}}$: the transition width. We set $\varepsilon_{\text{gate}} = \tau/5$, placing the gate's inflection point at $\mu_l = \tau$ with a transition band of $\pm 2\varepsilon_{\text{gate}} = \pm 0.4\tau$. A narrower band (e.g., $\varepsilon_{\text{gate}} = \tau/20$) approximates a hard switch and can introduce gradient spikes near the threshold; a wider band (e.g., $\varepsilon_{\text{gate}} = \tau$) includes too many borderline-passive layers and dilutes the geometric constraint. The $\tau/5$ choice keeps the gradient magnitude of the sigmoid gate bounded by $1/(4\varepsilon_{\text{gate}}) = 5/(4\tau)$ at the inflection, which is stable for all $\tau > 0$.

Setting ε_{act}: the activation cutoff. We use $\varepsilon_{\text{act}} = 0.05$. Any layer with $w_l \leq 0.05$ contributes less than 5% of the full CRF loss; computing the noisy forward pass for such layers wastes compute without providing contraction benefit. Empirically, this cutoff skips approximately 20–30% of all layers (the clearly passive ones) and reduces wall-clock training time by 15–20% compared to applying CRF unconditionally.

Recommended calibration procedure.

(1) Run the teacher on $\sim$200 held-out reasoning prompts.

(2) Record $\mu_{l,t} = \|\Delta_{l,t}^*\|_2$ for all layers l and tokens t.

(3) Set $\tau = \mathrm{P}_5(\{\mu_{l,t}\})$, $\varepsilon_{\text{gate}} = \tau/5$, $\varepsilon_{\text{act}} = 0.05$.

(4) Visualise the histogram of $\{\mu_{l,t}\}$ and verify that τ separates a clear low-flux cluster from the active layers; adjust to P_{10} if no clear gap is visible.

This calibration adds negligible overhead (a single teacher forward pass over 200 prompts) and is performed once before training.

G CRF Training Algorithm

Algorithm 1 gives the complete pseudocode for one CRF training step, corresponding to the two-phase procedure described in Section 3.4.

H Training Dataset Generation

The CRF loss requires the teacher's hidden-state trajectory $\{h_l^*, \Delta_l^*\}_{l=0}^{L-1}$ to be available at training time. This section describes how that trajectory is collected and how the hidden-state dataset is constructed.

Hidden-state collection. Given a reasoning prompt x and teacher parameters ϕ, the teacher is run in standard autoregressive mode. For each generated token y_t at decoding step t, the teacher produces a sequence of hidden states $\{h_t^{(l)}\}_{l=0}^{L}$, where $h_t^{(l)} \in \mathbb{R}^d$ is the output of layer l at position t. The logit vector at the last layer is used to generate y_t:

$$z_t = W_u\, h_t^{(L)}, \qquad p(y_t \mid x, y_{<t}) = \mathrm{softmax}(z_t).$$

We store $h_t^{(l)}$ for all layers l and all reasoning steps t, forming the teacher trajectory $\mathcal{H}^* = \{h_{l,t}^*\}$.

Algorithm 1 One CRF training step.

Require: Teacher ϕ, Student θ, batch (x, y), α, σ, τ, $\varepsilon_{\text{gate}}$, λ, ε_{act}

// Phase I: Teacher geometry (no gradient)
$\{h_l^*,\ \Delta_l^*\}_{l=0}^{L-1} \leftarrow \text{TeacherForward}(\phi,\ x)$
for $l = 0$ **to** $L-1$ **do**
  $\mu_l \leftarrow \|\Delta_l^*\|_2; \quad w_l \leftarrow \sigma_{\text{gate}}\big((\mu_l - \tau)/\varepsilon_{\text{gate}}\big)$
end for
// Phase II: Student stability test
$\mathcal{L}_{\text{CRF}} \leftarrow 0$
for $l = 0$ **to** $L-1$ **do**
  if $w_l > \varepsilon_{\text{act}}$ **then**
    $u_l^* \leftarrow \Delta_l^*/\mu_l; \quad P_\perp^{(l)} \leftarrow I - u_l^*(u_l^*)^\top; \quad \delta \sim \mathcal{N}(0, \sigma^2 I)$
    $\tilde{\Delta}_l \leftarrow f_\theta^{(l)}\big(\text{sg}[h_l^*] + \delta\big)$
    $e_\perp \leftarrow P_\perp^{(l)}(\tilde{\Delta}_l - \Delta_l^*) + \alpha P_\perp^{(l)}\delta$
    $\mathcal{L}_{\text{CRF}} \mathrel{+}= w_l \|e_\perp\|_2^2$
  end if
end for
// Phase III: Update
$\mathcal{L}_{\text{SFT}} \leftarrow -\sum_t \log f_\theta(y_t \mid x, y_{<t})$
$\theta \leftarrow \theta - \eta\, \nabla_\theta(\mathcal{L}_{\text{SFT}} + \lambda \mathcal{L}_{\text{CRF}})$

Residual Flux extraction. The *Residual Flux* at layer l and step t is computed as:

$$\Delta_{l,t}^* = h_t^{(l+1)} - h_t^{(l)} = f_\phi^{(l)}(h_t^{(l)}),$$

which is the output of the residual block $f_\phi^{(l)}$ (self-attention + feed-forward). The flux magnitude $\mu_{l,t} = \|\Delta_{l,t}^*\|_2$ determines the Kinetic Gating weight $w_{l,t}$ for each layer and step.

Cross-tokenizer alignment. When teacher and student use different tokenizers (Sec 4.2), the teacher token index t^T and student token index t^S do not correspond. We resolve this with a deterministic byte-span mapping $m : \{0, \ldots, T_S{-}1\} \to \{0, \ldots, T_T{-}1\}$ defined as follows. Each student token t^S corresponds to a byte span $[a_{t^S}, b_{t^S})$ of the underlying UTF-8 string; we set $m(t^S)$ to the smallest teacher token index whose byte span fully contains $[a_{t^S}, b_{t^S})$. When no such teacher token exists (a student token spans across two teacher tokens), we use mean pooling of the overlapping teacher token hidden states weighted by byte-overlap fraction. The aligned teacher hidden state used at student layer l and student token t^S is therefore

$$h^*_{l,\,t^S} \;:=\; A_l\, h^{T,*}_{\pi(l),\,m(t^S)}\,,$$

combining the layer-mapping π and per-layer alignment A_l from Appendix E with the byte-span token map m. For intra-family distillation (Sec 4.2, where teacher and student share a tokenizer), m is the identity and the construction reduces to $h^*_{l,t} = A_l h^{T,*}_{\pi(l),t}$.

Probe placement under sequence-level attention. Equation (5) writes the loss in terms of a single hidden-state vector, but each Transformer block actually maps a $T \times d$ sequence-level matrix through self-attention. We apply the probe noise δ only to the hidden state at the student token position currently being supervised; the remaining $T{-}1$ positions retain their teacher-conditioned hidden states (used as KV cache during the noise-perturbed forward pass). This isolates the local contraction property at the supervision token without contaminating the attention context, and matches the per-token decomposition used throughout the proofs in Appendix B.

Dataset format. The final training dataset consists of tuples $(x,\ y^*,\ \mathcal{H}^*,\ \{\Delta_{l,t}^*\})$, where y^* is the teacher-generated reasoning chain and $\mathcal{H}^*$ is the corresponding hidden-state trajectory. The SFT component $\mathcal{L}_{\text{SFT}}$ uses (x, y^*); the CRF component uses $(\mathcal{H}^*, \{\Delta_{l,t}^*\})$. Both components are computed in the same forward pass over the student at Phase II of Algorithm 1.

Storage requirements. Storing the full hidden-state trajectory requires $O(T \cdot L \cdot d)$ memory per example, where T is the number of reasoning tokens, L is the number of layers, and d is the hidden

dimension. For a 72B-parameter teacher with $L = 80$ layers, $d = 8192$, and $T = 1024$ reasoning tokens, each example requires approximately $80 \times 8192 \times 1024 \times 2 \approx 1.3\,\text{GB}$ in `bfloat16`. In practice, we use activation checkpointing at the layer level, retaining only the teacher states at each layer boundary and discarding the intermediate sub-layer activations, reducing per-example storage to approximately 40% of the naive estimate.

H.1 Jacobian Measurement Protocol

Figure 3(a) in the main text reports the transversal contraction rate $\|P_\perp J_l\|_2$ per layer after training. Since computing the full $d \times d$ Jacobian is prohibitive for $d \geq 4096$, we use a Hutchinson-style random-projection estimator.

Estimator definition. For each layer and evaluation prompt, sample 64 independent unit directions v_i uniformly from the unit sphere. Define

$$\hat{s}_l = \max_{1 \leq i \leq 64} \left\| P_\perp^{(l)} \frac{T_\theta^{(l)}(h_l^* + \eta v_i) - T_\theta^{(l)}(h_l^*)}{\eta} \right\|_2, \qquad \eta = 10^{-3}. \tag{28}$$

With exact Jacobian–vector products, the corresponding maximum is bounded above by $\|P_\perp^{(l)} J_l\|_2$. The finite-difference implementation additionally incurs truncation and numerical error. This diagnostic does not certify an upper bound on the spectral norm.

Bias–variance trade-off. For a d-dimensional spectral norm estimator using k random projections, the standard deviation of the maximum satisfies $\text{Std}[\max_i \|P_\perp J_l v_i\|_2] = O(\|P_\perp J_l\|_2 \cdot (k \log k)^{-1/2})$ by extreme-value theory for sub-Gaussian random variables. With $k = 64$ and $d = 4096$, this gives a relative standard error of approximately 9%. We verified stability by repeating the measurement at $k \in \{16, 32, 64, 128\}$ for a representative 7B checkpoint; the estimated $\|P_\perp J_l\|_2$ changed by less than 1.5% between $k = 64$ and $k = 128$, confirming convergence.

Finite-difference step size. The step $\eta = 10^{-3}$ is chosen to balance truncation error ($O(\eta)$ for first-order finite differences) against floating-point cancellation error ($O(\varepsilon_{\text{mach}}/\eta)$ for `bfloat16`). The optimal $\eta \approx \sqrt{\varepsilon_{\text{mach}}} \approx 8 \times 10^{-3}$ for `float32`; in `bfloat16` ($\varepsilon_{\text{mach}} \approx 8 \times 10^{-3}$), we use $\eta = 10^{-3}$ to stay above the cancellation floor while keeping truncation manageable.

Confidence bands in Figure 3(a). The bands in Figure 3(a) report the mean $\pm$ one standard deviation of $\widehat{\|P_\perp J_l\|}_2$ computed over three evaluation seeds (each using $k = 64$ projection vectors drawn independently), evaluated on 50 held-out prompts per seed. The reference line at $1 - \alpha$ is the asymptotic target predicted by Theorem 3; the upper band $(1 - \alpha) + \sqrt{\varepsilon}/\sigma$ from Proposition 2 (with ε averaged over the last 10% of training steps) is shown for reference, but as discussed above is not formally verifiable from the lower-bound estimator alone. The empirical content of Figure 3(a) is the consistency of estimated values with $1 - \alpha$ across CRF variants and the falsification of contraction by the no-restoring-force ablation, both of which are within the statistical scope of the estimator.

I Ablation Study

Since we introduce the decomposition of the tangential and traversal components and the kinetic gate in the CRF loss to model it as a dynamic problem, we compare the performance with and without the two mechanisms. To eliminate the irrelevant factors, we pick teacher and student models in the same family. All other training settings remain the same with Section 4.2, except the deduction of the decomposition part and the kinetic gate. In these conditions, the CRF loss becomes:

$$\mathcal{L}_{\text{CRF, w/o-decomposition}}^{(l)} = \mathbb{E}_{\delta \sim \mathcal{N}(0, \sigma^2 I)} \Big[\big\| \big(f_\theta^{(l)}(h_l^* + \delta) - \Delta_l^* \big) + \alpha\, \delta \big\|_2^2 \Big], \tag{29}$$

$$\mathcal{L}_{\text{CRF, w/o-gate}} = \sum_{l=0}^{L-1} \mathcal{L}_{\text{CRF}}^{(l)}. \tag{30}$$

Table 5: **Ablation of the decomposition and kinetic gate.** Pass@1 (%) on AIME, MBPP, and LCB with and without each component. Best results within each student setting are bold.

Setting			Benchmarks		
Model	Decomposition	Gate	AIME	MBPP	LCB
Qwen2.5-7B	w/	w/o	14.5	82.6	21.1
		w/	**14.9**	**83.3**	**21.3**
	w/o	w/o	8.3	77.3	14.6
		w/	9.0	78.6	15.2
Qwen2.5-3B	w/	w/o	9.7	78.1	13.8
		w/	**10.1**	**79.5**	**14.1**
	w/o	w/o	6.3	73.8	9.4
		w/	7.1	74.6	9.9
Qwen2.5-0.5B	w/	w/o	5.2	53.7	10.3
		w/	**6.2**	**56.9**	**11.4**
	w/o	w/o	0.7	45.4	3.3
		w/	0.8	45.7	3.4

We find that both mechanisms improve the performance in three benchmarks, while the decomposition promotes the global convergence more than the gate. It indicates that the decomposition reveals the specific direction in reasoning, and the gate ensures the stability of the training.

J Computational Resources

All experiments were conducted on NVIDIA H800-80GB GPUs and consumed approximately 5,000 GPU-hours in total, including teacher trajectory generation, all distillation runs (3 seeds), ablations, and hyperparameter search.

K Cognitive Integrity Evaluation

To test whether transversal regularisation incurs an alignment tax, we distil Qwen2.5-72B into Qwen3-8B and evaluate on six general benchmarks: MT-Bench [Bai et al., 2024], Vicuna [Sun et al., 2024], Alpaca 2 [Li et al., 2023], Super-NI [Wang et al., 2022], OpenLLM [Liu et al., 2024], and Dolly [Conover et al., 2023].

Table 6: **General instruction-following evaluation.** Results for Qwen3-8B distilled from Qwen2.5-72B on six benchmarks. Best results among student methods are bold.

Method	General Instruction-Following Benchmarks						Summary	
	Alpaca 2	MT-Bench	Dolly	Vicuna	Super-NI	OpenLLM	Avg.	Avg.Δ
Teacher (Qwen2.5-72B)	45.2	9.35	82.1	94.2	65.4	49.8	57.6	-
Student (Qwen3-8B)	22.1	6.54	68.2	82.1	45.2	29.3	42.2	-
Standard KD	23.5	6.82	70.1	84.5	48.1	30.5	43.9	+1.7
SeqKD	23.1	6.75	69.5	83.2	47.5	31.4	43.6	+1.4
MiniLLM	24.2	6.95	71.2	85.1	49.2	32.8	44.9	+2.7
GKD (On-Policy)	26.4	7.21	73.5	87.2	52.1	37.3	47.3	+5.1
DA-KD (Data Select)	**27.2**	**7.35**	**74.2**	87.8	53.5	38.1	48.0	+5.8
CRF (Ours)	26.3	7.12	73.8	**89.4**	**56.4**	**39.0**	**48.7**	**+6.5**

Result. CRF achieves the highest average (48.7) and leads on task-oriented benchmarks (Super-NI, Vicuna, OpenLLM), while trailing DA-KD by small margins on conversational metrics (Alpaca: -0.9, MT-Bench: -0.23, Dolly: -0.4). On this single (teacher, student) pair the geometric constraint does not produce a large alignment tax, and on task-oriented benchmarks it appears net-helpful, but

we caution that a single configuration is not sufficient to support a general claim about alignment. We did not run cognitive-integrity evaluations under the cross-architecture setting of Section 4.2; we leave a broader instruction-following audit of CRF to future work.

L Notation Reference

Table 7 collects all symbols used in the main text and appendices. Superscripts (l) always denote layer index; subscripts l denote the same when context is unambiguous (e.g., μ_l, w_l). The symbol $\|\cdot\|_2$ denotes spectral norm when applied to a matrix and Euclidean norm when applied to a vector; $\|\cdot\|_F$ always denotes Frobenius norm.

Table 7: **Notation reference.** Symbols used in the main text and appendices.

Symbol	Definition
d	Hidden dimension
L	Number of Transformer layers
$h_l \in \mathbb{R}^d$	Student hidden state at layer l
$h_l^* \in \mathbb{R}^d$	Teacher hidden state at layer l
$\epsilon_l = h_l - h_l^*$	Student–teacher deviation at layer l
$f_\theta^{(l)}$	Student residual function at layer l
$f_\phi^{(l)}$	Teacher residual function at layer l
$T_\theta^{(l)} = \mathrm{id} + f_\theta^{(l)}$	Full layer map
$\Delta_l^* = f_\phi^{(l)}(h_l^*)$	Teacher residual flux
$\mu_l = \|\Delta_l^*\|_2$	Flux magnitude (kinetic gate input)
$u_l^* = \Delta_l^*/\mu_l$	Unit tangent direction
$P_\perp^{(l)} = I - u_l^*(u_l^*)^\top$	Transversal projection
$J_l = I + J_f^{(l)}$	Full layer Jacobian at h_l^*
$J_f^{(l)} = \partial f_\theta^{(l)}/\partial h\big\|_{h_l^*}$	Residual Jacobian
$r_l = f_\theta^{(l)}(h_l^*) - \Delta_l^*$	Zeroth-order fitting residual
$\epsilon_l^\perp = P_\perp^{(l)}\epsilon_l$	Transversal deviation
$\alpha \in (0,1)$	Contraction coefficient
$\sigma > 0$	Probe noise scale
$\gamma_{\mathrm{eff}} = (1-\alpha) + 2\kappa J_{\max}$	Effective contraction rate
κ	Manifold smoothness constant ($\|u_{l+1}^* - u_l^*\|_2 \le \kappa$)
$J_{\max} = \max_l \|J_l\|_2$	Maximum layer Jacobian norm
$R_\perp = \max_l \|P_\perp^{(l+1)} r_l\|_2$	Maximum projected residual
τ	Kinetic gate flux threshold
$\varepsilon_{\mathrm{gate}}$	Kinetic gate transition width
w_l	Kinetic gate weight at layer l
W_u	Unembedding matrix
$K = \|W_u\|_2$	Output Lipschitz constant
δ_{SFT}	Inference-time student–teacher gap
$H_{\max}$	Lipschitz constant of $J_f^{(l)}$ w.r.t. h
λ	CRF loss weight
η	Learning rate

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

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

   Answer: [Yes]

   Justification: We provide main claims and contributions in abstract and introduction.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss the limitations of the work in a separate section.

Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: We provide the full set of assumptions and a complete proof in Appendix B.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We provide a detailed description of the experimental process in Section Experiments and provide an anonymous GitHub repository to our code.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

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

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

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

   Answer: [Yes]

   Justification: We provide a anonymous GitHub link to our code.

   Guidelines:

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

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: We provide the experimental setting and details. See details in Section Experiments.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: We report error bars and standard deviations in the main comparative experiments.

   Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We report it in Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and our paper conforms with it.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: Our research does not have such risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: The assets used are properly noted and credited.

    Guidelines:

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

13. **New assets**

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

    Answer: [N/A]

    Justification: No new assets are being released.

    Guidelines:

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

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

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

    Answer: [N/A]

Justification: We do not have any studies or results regarding crowdsourcing experiments and human subjects.

Guidelines:

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

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

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

Answer: [N/A]

Justification: We do not have any studies or results including study participants.

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: The method in this research does not involve large language models (LLMs) as essential, original, or non-standard components.

Guidelines:

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