A Control-Theoretic Account of Cognitive Effort in Language Models

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

We study how post-training reshapes the control geometry of large language models. Treating the residual stream as the state of a time-varying linear system, we fit local layer-to-layer maps, build finite-horizon controllability Gramians, and quantify (i) *geometric difficulty* via minimal end-to-end control energy $E_{\min}$ and (ii) *efficiency* $\eta = E_{\min}/E_{\text{actual}}$ along realized trajectories. Across four stages, from *Baseline* to fine-tuned models (*SFT* $\rightarrow$ *DPO* $\rightarrow$ *Instruct (RLVR)*) the Gramian spectrum compresses (fewer large-eigenvalue "easy" directions) and $E_{\min}$ rises monotonically. Principal-angle analyses show that fine-tuning rotates both "easy" and "hard" subspaces relative to Baseline, while off-manifold occupancy increases. Surprisingly, under a shared PCA, *conversational* prompts are geometrically harder than *math* prompts (higher $E_{\min}$, lower η), revealing a divergence between human-intuitive difficulty and LM (language model) control geometry. These results recast well-known post-training trade-offs as changes in controllability: steering remains possible, but "cheap" directions become scarce, implying larger control energy unless interventions target the new post-training control axes.

1 Introduction

The natural hierarchy of task difficulty is evident in everyday life. For instance, solving a complex proof demands far more cognitive effort than daydreaming or casual conversation. Why some tasks feel effortful remains an active question. A prominent account assigns the DLPFC a top-down control role and the ACC a monitoring/valuation role for conflict and the cost of control [Miller and Cohen, 2001, Shenhav et al., 2013, Botvinick et al., 2001]. In this view, hard tasks require stronger, sustained control to keep neural dynamics on track, whereas easy tasks proceed with minimal intervention [Shenhav et al., 2013]. In routine conditions, population activity lies on a low-dimensional *intrinsic manifold* that supports reliable, low-effort behavior [Sadtler et al., 2014, Gallego et al., 2017, Cunningham and Yu, 2014]. Outside this core is the *controllosphere*—hard-to-reach, weakly observed directions that demand stronger intervention [Gu et al., 2015, Holroyd, 2024].

We aimed to answer whether analogous dynamics appear in large language models (LLMs). Akin to state dynamics in the brain, we interpret the evolving activity of language models, particularly the residual stream (x_t), as the state trajectories through a high-dimensional state space [Nelson et al., 2021]. We model the residual stream as the state x_t of a time-varying discrete linear system, fit local maps between layers, and quantify *geometric difficulty* and *efficiency* along observed trajectories [Holroyd, 2024].

We found that across fine-tuned stages, the Gramian spectrum compresses (fewer very-easy directions), and end-to-end $E_{\min}$ rises. A common view is that entering the *controllosphere* region demands strong, sustained control signals and tasks like multi-step mental arithmetic push activity into this zone, which is why they feel effortful; casual conversation typically remains on the intrinsic manifold [Shenhav et al., 2013, Monsell, 2003]. Strikingly, under a shared PCA, *conversational* prompts are geometrically harder than *math* (higher $E_{\min}$, lower η), suggesting a gap with human-intuitive difficulty; future work should test whether these metrics predict accuracy and reliability via causal manipulations of controllability Gramian W_L (spectrum and hard-subspace rotations).

2 Experimental design and methodology

2.1 Data and Models

We analyzed four successive training stages of the same transformer model (Allen AI's *OLMo-2-0425-1B*): **Baseline**, **SFT**, **DPO**, and **Instruct** [OLMo et al., 2024]. To ensure comparability across conditions, we constructed a joint dataset by pooling **four** corpora: two *mathematical* sets (*Math 1* [Amini et al., 2019], *Math 2* [Ling et al., 2017]) and two *conversational* sets (*Conversational 1* [Zheng et al., 2023], *Conversational 2* [Lecorvé et al., 2022], [Bordes et al., 2015]).

To ensure comparability across conditions, we constructed a joint dataset and then split it into training, validation, and test subsets. For each model and prompt, we cached both the input embeddings and the intermediate layer activations, so that all stages were evaluated on exactly the same inputs. To characterize linear manifolds, stabilize estimation, and facilitate cross-model comparisons, we projected all cached states for each data subset into a common low-dimensional subspace. This subspace was derived by applying PCA to the concatenated activations across models, yielding shared representations of dimension $d' = 100$.

2.2 States and Local Linearization

Let $x_k \in \mathbb{R}^{d'}$ denote the residual-stream states after layer k, with x_0 corresponding to the input embedding at the model's first layer. Around a fixed context, we model step-to-step dynamics by a time-varying linear approximation

$$x_{k+1} \approx A_k\, x_k + \varepsilon_k, \qquad k = 0, \dots, L, \tag{1}$$

and collect paired sample matrices (X_k, Y_k) across N_m mathematical and N_c conversational prompts/tokens, where X_k holds x_k and Y_k holds x_{k+1}. We estimate A_k with ridge regression.

2.3 Controllability Metrics

First, we build the time-varying finite horizon controllability Gramians. Given $\{A_k\}_{k=0}^{L-1}$, we assumed inputs act in all directions ($B{=}I$) such that,

$$W_0 = 0, \qquad W_{k+1} = A_k W_k A_k^\top + I, \quad k = 0, \dots, L-1. \tag{2}$$

and the state transition matrix is $\Phi = A_{L-1} \cdots A_0$.

For each sample with endpoints z_0 and z_L, the *minimal* end–to–end control energy (geometric difficulty) is

$$E_{\min} \;=\; (z_L - \Phi z_0)^\top W(L)^{-1} \,(z_L - \Phi z_0) \;=\; \big\| \, W(L)^{-1/2} \,(z_L - \Phi z_0) \,\big\|_2^2.$$

The *actual* control energy (observed effort) accumulated along the realized trajectory is the sum of squared residual pushes

$$u_k^{\text{obs}} \;=\; z_{k+1} - A_k z_k, \qquad E_{\text{actual}} \;=\; \sum_{k=0}^{L-1} \|u_k^{\text{obs}}\|_2^2 \;=\; \sum_{k=0}^{L-1} \|\, z_{k+1} - A_k z_k \,\|_2^2,$$

By optimal–control theory $E_{\min} \le E_{\text{actual}}$ for every sample; we also report the dimensionless efficiency $\eta = E_{\min}/E_{\text{actual}} \in (0, 1]$.

Easy/Hard subspaces From the controllability Gramian W_k's eigenvectors, we define the *easy manifold* as the top-q eigenvectors (largest eigenvalues) and the *hard subspace* as the bottom-q eigenvectors (smallest). We used $q \in 10, 20, 30$.

Let $W_L = Q \Lambda Q^\top$ be the controllability Gramian at index L, with $\Lambda = \operatorname{diag}(\lambda_1 \le \cdots \le \lambda_d)$ and $Q = [q_1, \dots, q_d]$. For a hard subspace of dimension k, we take $V_h := [q_1, \dots, q_k]$, i.e., the span of the k eigenvectors associated with the smallest eigenvalues of W_L.

Off-Manifold Occupancy (Geometry). Given activations $Z \in \mathbb{R}^{N \times d}$ and a reference hard basis $V_h \in \mathbb{R}^{d \times k}$, for each sample z we define the hard-occupancy fraction

$$\mathrm{align}(z; V_h) \; = \; \sqrt{\frac{\|V_h V_h^\top z\|^2}{\|z\|^2}} \; = \; \cos\big(\angle(z, \mathrm{span}(V_h))\big) \in [0, 1].$$

We use (i) the baseline's hard subspace V_h^{base} to test a shift toward baseline-hard, and (ii) each model's own hard subspace to compare its off-manifold usage.

3 Results and Discussion

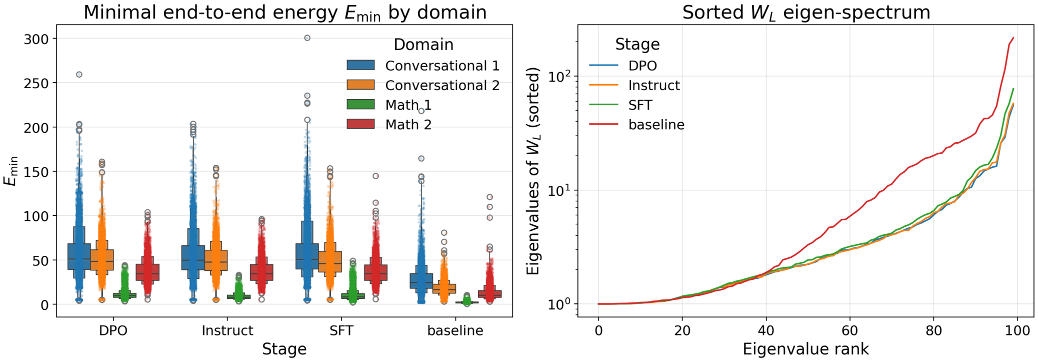

Figure 1: Minimum end-to-end energy and eigenspectrum of controllability Gramian at the L^{th} index.

Geometric difficulty and efficiency diverge by domain. Relative to Baseline, all fine-tuned stages showed higher $E_{\min}$ in every domain (Figure 1) (*Conversational 1*: $+2.03$–$2.07\times$; *Conversational 2*: $+2.72$–$2.85\times$; *Math 1*: $+4.29$–$5.05\times$; *Math 2*: $+3.27\times$), indicating trajectories that traverse harder-to-reach directions of state space. Efficiency is domain-dependent: it decreases for both conversational sets (Conv-1: 0.69–$0.71\times$; Conv-2: 0.82–$0.83\times$ of Baseline), increases sharply for *Math 1* (2.22–$2.67\times$), and decreases modestly for *Math 2* (0.79–$0.80\times$). Thus, fine-tuning pushes models off the easy manifold across domains; residual pushes are more purposefully aligned with the hard displacement on *Math 1*, but less aligned on both conversational sets and on *Math 2*. See Appendix A.4 for data.

W_L spectrum compression. We found that W_L spectra compress after fine-tuning (Figure 1). The Baseline shows a much larger upper tail (e.g., $\lambda_{\max} \approx 2.16 \times 10^2$), indicating many very easy directions. Fine-tuned models have markedly smaller $\lambda_{\max}$ (DPO $\approx 5.57 \times 10^1$, Instruct $\approx 5.74 \times 10^1$, SFT $\approx 7.71 \times 10^1$). Because $E_{\min}$ weights displacements by W_L^{-1}, this compression raises end-to-end minimal energy (e.g., 95th-percentile $E_{\min}$: Baseline 56.7 vs. ~ 100 for DPO/Instruct/SFT).

Layer-wise geometric difficulty. We plot the minimal energy together with the relative hardness $\rho_k = E_k/\|z_k\|^2$. (Figure 2) Across all domains, the fine-tuned stages (DPO/Instruct/SFT) maintain substantially higher E_k than Baseline at almost every depth. On *Conversational 1/2*, fine-tuned E_k starts high ($\sim 10^2$) and decays only gradually, while ρ_k remains elevated with a mild downward drift; the Baseline (red) drops sharply across layers. On *Math 1*, both E_k and ρ_k show an early dip followed by a mid/late-layer rise (a clear late bump), signaling a return to hard directions toward the head of the stack, whereas the Baseline continues to soften. On *Math 2*, E_k peaks early (layers ≈ 4–6) and then tapers, and ρ_k is comparatively flat with a slow decline; fine-tuned models remain harder than Baseline at all depths.

Subspace alignment. We compared pairwise principal angles between the "hard" (smallest eigenvalue) and "easy" (largest–eigenvalue) subspaces of W_L for $q \in \{10, 20, 30\}$. Baseline versus any fine-tuned model (DPO/Instruct/SFT) is nearly orthogonal in both spaces (most angles 80° with tight spreads), indicating that fine-tuning rotates both the controllosphere and on-manifold directions

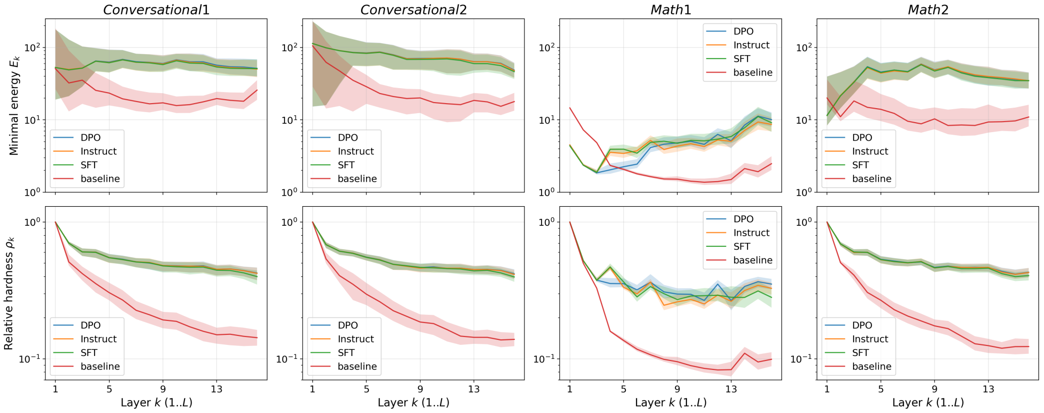

Figure 2: Layer-wise geometric difficulty based on E_{min} and ρ_k.

away from Baseline. Pairs of fine-tuned models (DPO–Instruct, DPO–SFT, Instruct–SFT) exhibit a compact low-angle core at $q{=}10$ (medians in the single digits to $\sim 10°$), evidencing a shared aligned subspace. As q increases to 20 and 30, these within-FT angle distributions broaden—medians rise to $\sim 10°$–$20°$ and upper tails extend to $40°$–$60°$—so the fraction of strongly aligned dimensions (e.g., $\leq 20°$) decreases; the same pattern holds in both hard and easy spaces, consistent with a low-rank FT-specific core plus model-specific rotations in the residual dimensions. See Appendix A.5 for data.

At the final index L (Figure 3), we also measure visitation to Baseline's hard subspace via the cosine to $\mathrm{span}(V^{\mathrm{base}}_{\mathrm{hard}})$. Fine-tuned models show substantially larger cosines than Baseline at all q, and the cosine increases with q (roughly $\sim 0.35 \to 0.65$ for FT vs. $\sim 0.08 \to 0.18$ for Baseline), indicating that post-training drives trajectories into Baseline's "hard" directions more frequently.

Finally, to quantify visitation to each model's own controllosphere (Figure 3), we measure visitation via the cosine to $\mathrm{span}(V^{\mathrm{own}}_{\mathrm{hard}})$. The cosines are consistently higher for fine-tuned stages than for Baseline and grow with q; conversational inputs are largest, *Math 1* smallest, with *Math 2* in between, confirming that post-training increases occupancy of each model's hard subspace.

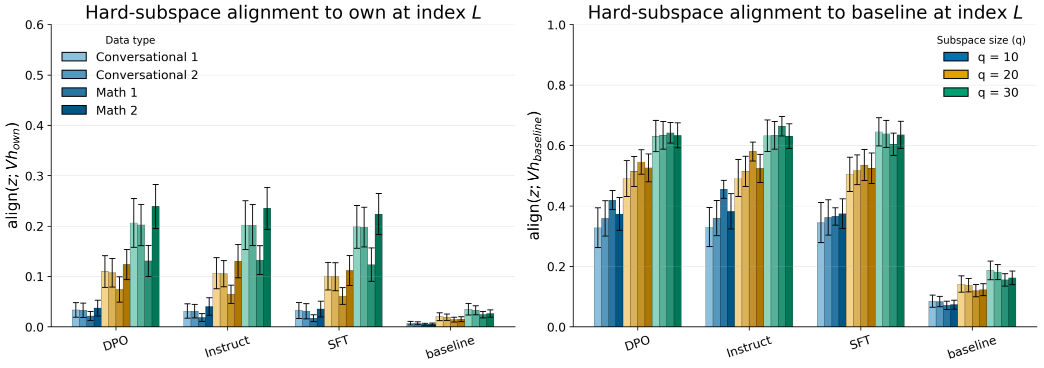

Figure 3: Alignment of hard subspaces of fine-tuned stages with itself and with the baseline.

4 Conclusion

The current study suggests that post-training consistently reshapes the controllability landscape of LMs while domain effects are non-intuitive, i.e, conversational prompts are geometrically harder than math under a shared representation. These findings reconcile observed post-training behavior changes with control-theoretic mechanisms. Our analysis is observational and can be extended to include causal attributions to specific post-training objectives, multiple architectures, more robust post-training stages, and multi-modal multiple datasets.

References

Earl K Miller and Jonathan D Cohen. An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, 24(1):167–202, 2001.

Amitai Shenhav, Matthew M Botvinick, and Jonathan D Cohen. The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, 79(2):217–240, 2013.

Matthew M Botvinick, Todd S Braver, Deanna M Barch, Cameron S Carter, and Jonathan D Cohen. Conflict monitoring and cognitive control. *Psychological review*, 108(3):624, 2001.

Patrick T Sadtler, Kristin M Quick, Matthew D Golub, Steven M Chase, Stephen I Ryu, Elizabeth C Tyler-Kabara, Byron M Yu, and Aaron P Batista. Neural constraints on learning. *Nature*, 512 (7515):423–426, 2014.

Juan A Gallego, Matthew G Perich, Lee E Miller, and Sara A Solla. Neural manifolds for the control of movement. *Neuron*, 94(5):978–984, 2017.

John P Cunningham and Byron M Yu. Dimensionality reduction for large-scale neural recordings. *Nature neuroscience*, 17(11):1500–1509, 2014.

Shi Gu, Fabio Pasqualetti, Matthew Cieslak, Qawi K Telesford, Alfred B Yu, Ari E Kahn, John D Medaglia, Jean M Vettel, Michael B Miller, Scott T Grafton, et al. Controllability of structural brain networks. *Nature communications*, 6(1):8414, 2015.

Clay B Holroyd. The controllosphere: The neural origin of cognitive effort. *Psychological Review*, 2024.

Elhage Nelson, Nanda Neel, Olsson Catherine, Henighan Tom, Joseph Nicholas, Mann Ben, Askell Amanda, Bai Yuntao, Chen Anna, Conerly Tom, et al. A mathematical framework for transformer circuits, 2021.

Stephen Monsell. Task switching. *Trends in cognitive sciences*, 7(3):134–140, 2003.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 OLMo 2 Furious, 2024. URL `https://arxiv.org/abs/2501.00656`.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2357–2367, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1245. URL `https://aclanthology.org/N19-1245`.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. *ACL*, 2017.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric. P Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. Lmsys-chat-1m: A large-scale real-world llm conversation dataset, 2023.

Gwénolé Lecorvé, Morgan Veyret, Quentin Brabant, and Lina M. Rojas-Barahona. Sparql-to-text question generation for knowledge-based conversational applications. 2022.

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. Large-scale simple question answering with memory networks. *arXiv preprint arXiv:1506.02075*, 2015.

A Technical Appendices and Supplementary Material

A.1 State Space Dynamics

Around a fixed context, we approximate the step-to-step update with a local linear discrete time-varying model:

$$x_{t+1} \approx A(t)\, x_t + B\, u_t, \tag{3}$$

$$y_t = C\, x_t, \tag{4}$$

In this formulation, we set $B = I$. This choice is without loss of generality, as for any full-rank input matrix B, one can absorb it into the definition of the effective control signal $\tilde{u}_t = Bu_t$. Moreover, in the fitted model, the external input u_t is not an independent driver but simply the innovation term needed to reconcile the linear surrogate with the true trajectory. That is,

$$\tilde{u}_t := x_{t+1} - Ax_t, \qquad B = I.$$

Thus, the inputs coincide with the residual stream updates themselves. The case $\tilde{u}_t \equiv 0$ would correspond to a trajectory lying exactly on the linear dynamics, i.e, perfect prediction with no innovation. In the transformer setting, this corresponds naturally to the residual stream update, where each layer contributes an additive modification in the same coordinates as the state. Thus, the control inputs act directly on the state, justifying the identity choice. The output map C is omitted here because our analysis focuses on controllability and the minimal energy required to reach observed states, which depend only on (A, B).

A.2 Five-Fold Cross-Validation for Ridge Parameter α

To fit A_k with ridge regression, we used

$$\min_{A_k} \|Y_k - X_k A_k^\top\|_F^2 + \alpha \|A_k\|_F^2, \tag{5}$$

whose closed form satisfies $(X_k^\top X_k + \alpha I)A_k^\top = X_k^\top Y_k$.

To determine a single ridge penalty α for each k to fit the layer-wise linear dynamics $x_{k+1} \approx A_k x_k$ that generalizes best to held-out data, while avoiding any test leakage, we performed K=5-fold CV within the *train* split on a logarithmic grid $\mathcal{A} = \{\alpha_1, \dots, \alpha_G\} \subset [10^{-8}, 10^2]$ (5 log-spaced values by default). The same fold assignment is used for all models, layers, and PCA dimensions to ensure comparability.

A.3 Energy-weighted hard share (cost)

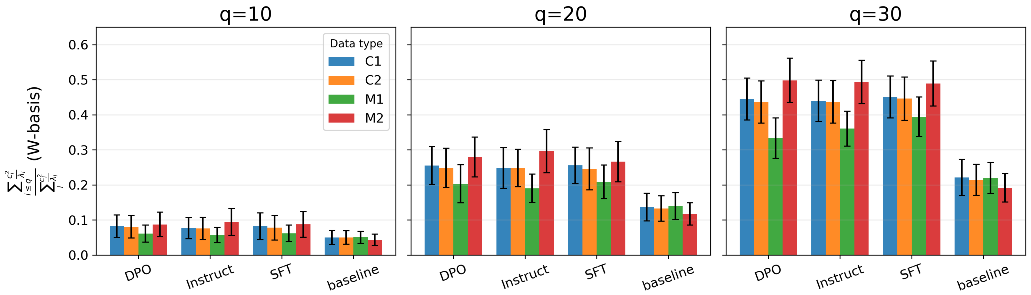

Figure 4: Energy-weighted hard share at layer L

Let controllability Gramian $W_L = Q\Lambda Q^\top$ with $\Lambda = \mathrm{diag}(\lambda_1 \le \cdots \le \lambda_d)$ and $c := Q^\top z$. The control-energy metric is given by

$$E(z) = z^\top W_L^{-1} z = \sum_{i=1}^{d} \frac{c_i^2}{\lambda_i}.$$

For hard-subspace size q, we define the *energy-weighted hard share* as

$$\mathrm{hardE}_q(z) \;=\; \frac{\sum_{i=1}^{q} \frac{c_i^2}{\lambda_i}}{\sum_{i=1}^{d} \frac{c_i^2}{\lambda_i}} \in [0,1].$$

Larger values imply that more of the required energy lies in the hardest modes/*controllosphere* (higher geometric difficulty).

A.4 E_{min} and control efficiency across stages and domains

Domain	Stage	$E_{\min}$ (median [95% CI])	η (median [95% CI])	× vs Baseline
Conversational 1	baseline	24.745 [24.433, 25.158]	0.052 [0.051, 0.053]	—
	DPO	51.324 [50.702, 52.197]	0.036 [0.036, 0.037]	E: 2.07×, η: 0.69×
	Instruct	50.156 [49.397, 50.857]	0.036 [0.035, 0.037]	E: 2.03×, η: 0.69×
	SFT	50.964 [50.311, 51.643]	0.037 [0.036, 0.037]	E: 2.06×, η: 0.71×
Conversational 2	baseline	17.015 [16.729, 17.266]	0.035 [0.034, 0.036]	—
	DPO	48.531 [47.964, 49.115]	0.029 [0.028, 0.029]	E: 2.85×, η: 0.83×
	Instruct	47.755 [47.216, 48.326]	0.029 [0.028, 0.029]	E: 2.81×, η: 0.82×
	SFT	46.306 [45.798, 46.861]	0.029 [0.028, 0.029]	E: 2.72×, η: 0.82×
Math 1	baseline	1.971 [1.944, 1.997]	0.040 [0.040, 0.041]	—
	DPO	9.958 [9.840, 10.094]	0.107 [0.106, 0.108]	E: 5.05×, η: 2.67×
	Instruct	8.450 [8.356, 8.534]	0.097 [0.097, 0.098]	E: 4.29×, η: 2.42×
	SFT	8.797 [8.665, 8.944]	0.089 [0.089, 0.090]	E: 4.46×, η: 2.22×
Math 2	baseline	10.602 [10.382, 10.799]	0.046 [0.046, 0.047]	—
	DPO	34.689 [34.212, 35.165]	0.037 [0.036, 0.037]	E: 3.27×, η: 0.79×
	Instruct	34.689 [34.230, 35.210]	0.037 [0.037, 0.037]	E: 3.27×, η: 0.80×
	SFT	34.674 [34.276, 35.093]	0.037 [0.037, 0.038]	E: 3.27×, η: 0.80×

Table 1: Minimal energy $E_{\min}$ (difficulty) and efficiency $\eta = E_{\min}/E_{\text{actual}}$ by domain and stage. Higher $E_{\min}$ indicates harder geometry; higher η indicates more purposeful use of residual energy.

A.5 Angles between the hard and easy subspaces of the controllability Gramian in different stages of post-training

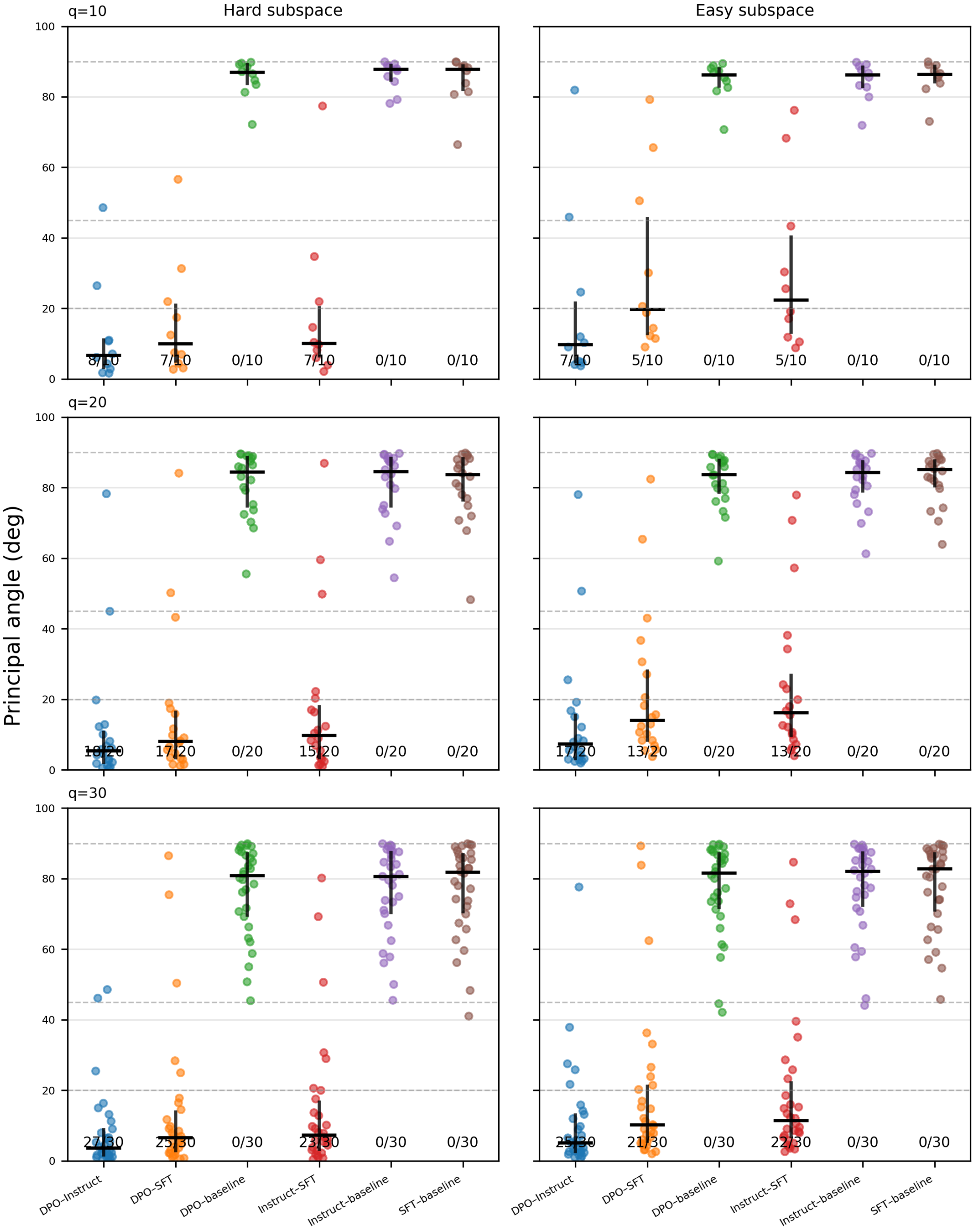

Figure 5: Angles between subspaces of the controllability Gramian.