The Gradient Does Not See Rank

Rank-Indifference in Matrix-CODI on ProsQA

Samuel Larson • Pebble Machine Learning

ICML 2026 Mechanistic Interpretability Workshop • Virtual Poster • OpenReview `Spof4PusVI`

TL;DR. Continuous chain-of-thought stores "parallel reasoning paths" in the *rank* of a matrix-valued latent Z, or so the superposition story goes. We truncate Z to rank k at inference and accuracy **does not move**. The training objective never rewards rank: the readout Jacobian carries no rank signal, four nonlinear readouts stay flat, and three seeds land at ranks $\{4, 12, 13\}$ with the same accuracy. The gradient is **blind to rank**.

1. The setup

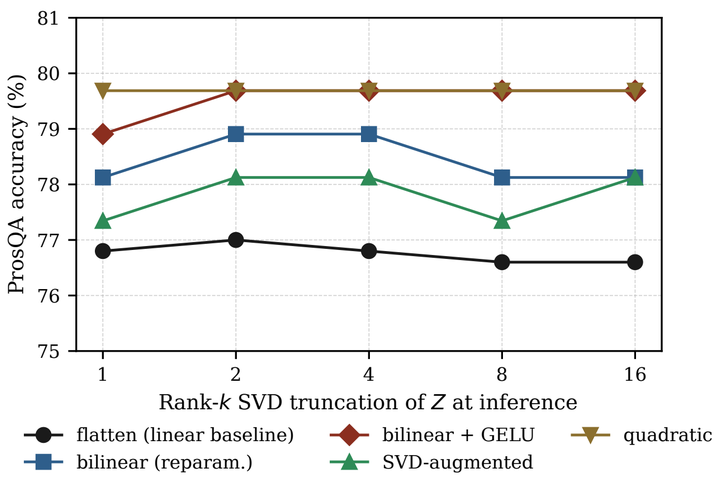
Matrix-CODI compresses an explicit rationale into a small number of continuous latent positions, each a $d \times d$ matrix Z:

$$h \xrightarrow{W_{\text{up}}} \text{reshape} \xrightarrow{\text{thinker}} Z \xrightarrow{\text{flatten}} W_{\text{down}} \to \text{answer}.$$

Z has a computable **rank** via its SVD. *If* each singular direction encodes a separate reasoning path, truncating Z to rank k should hurt accuracy once the task needs more than k paths. The **rank-k ablation curve** is the natural probe.

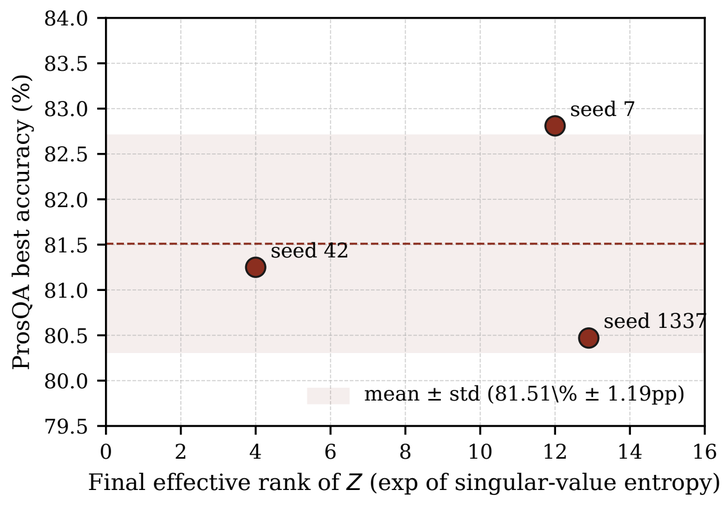
Setup: GPT-2 small, d=16, six latent positions, ProsQA (also GSM8K-Aug).

2. Rank-*k* ablation is flat



Truncating to $k \in \{1, 2, 4, 8, 16\}$ moves accuracy by ≤ 0.6 pp: flat at two distillation weights, with the multiplicative thinker on *or* off, on ProsQA *and* GSM8K-Aug. A rank-1 thought decodes as well as a full-rank one.

3. Seeds decouple rank from accuracy



Three seeds, identical hyperparameters → effective ranks $\{4, 12, 13\}$ but accuracy 81.5 ± 1.2 pp. **Rank wanders; accuracy doesn't.** This seed-level spread is what separates *rank-blindness* from mere position-irrelevance.

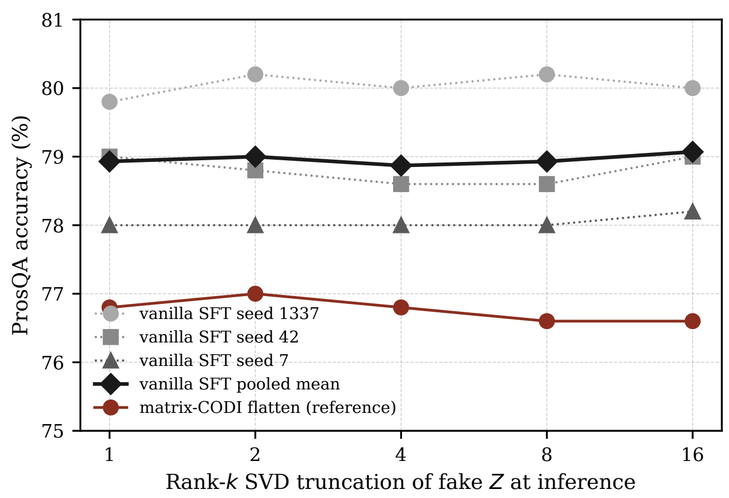
4. Positive controls: not a flatten artifact

Maybe "flatten then W_{down}" linearizes away the rank. We test four readouts that are **nonlinear in** Z and escape the linear-Jacobian shortcut:

Readout	Spearman p
Bilinear	0.63
Bilinear + GELU	0.14
SVD-augmented MLP	0.82
Quadratic $ZZ^\top$	0.46

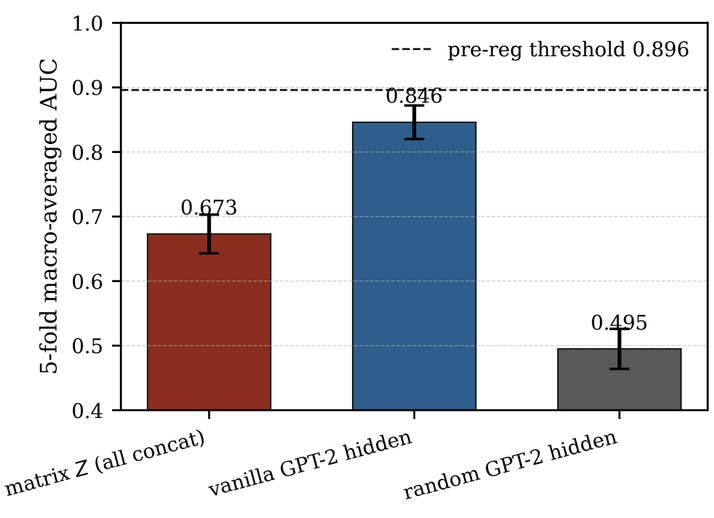
All four rank-k curves stay flat ($p \in [0.14, 0.82]$). The flatness **survives nonlinear readouts**: it is not a flatten artifact.

5. The probe trap (negative control)



Run the *same* rank-k intervention on vanilla GPT-2 SFT (no matrix bottleneck, no Z). It **also** produces a flat curve (pooled range 0.20 pp), and a random-h floor lands at the same accuracy. **Lesson: the rank-k ablation alone conflates rank-blindness with position-irrelevance.** The seed spread (panel 3) is the real evidence.

6. The latent *Z* is a weak feature



A linear probe on Z (1536 features) hits AUC 0.673 on target prediction; a raw pretrained GPT-2 hidden state (768 features) hits 0.846 on the same task. The learned matrix latent holds *less* decodable structure than the backbone it sits on.

Takeaway. CODI's objective does not reward rank: the Jacobian carries no rank signal, nonlinear readouts stay flat, and matched seeds scatter rank while pinning accuracy. "Superposition in the rank" is not learned here, and the obvious probe for it is itself misleading.

Code, checkpoints & eval logs: `github.com/saml212/matrix-codi-rank-blindness` • OpenReview `Spof4PusVI`