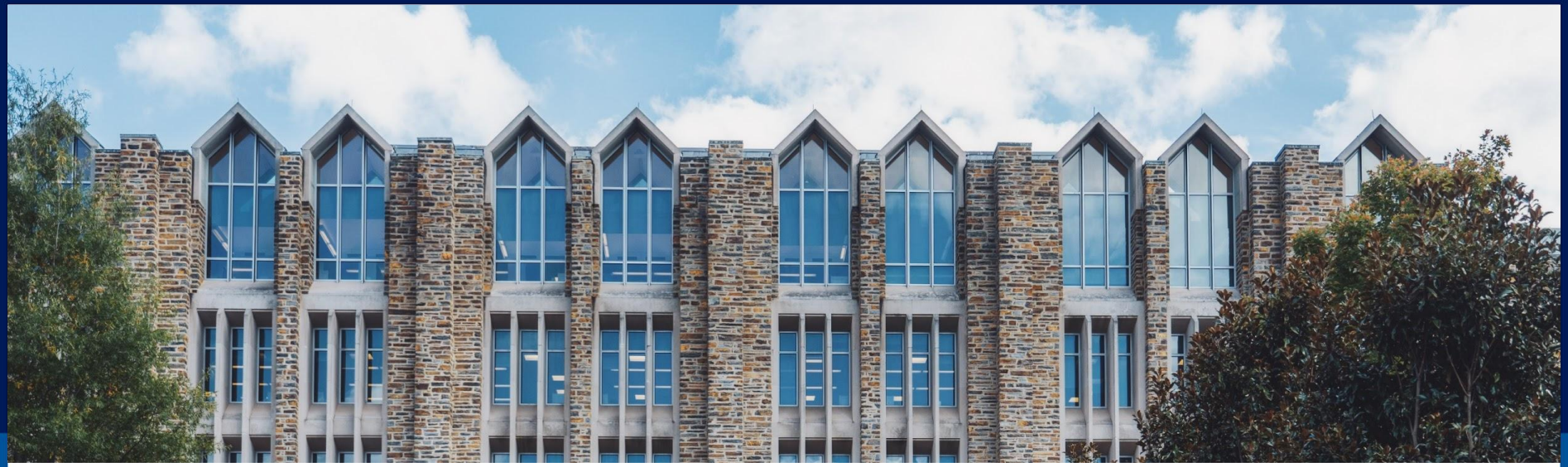


DecodeShare: Tracing the Shared Subspace of LLM Decode-Time Decisions

Zishan Shao, Lixun Zhang, Kangning Cui, Yixiao Wang, Ting Jiang, Hancheng Ye, Qinsi Wang, Zhixu Du, Yuzhe Fu, Fan Yang, Danyang Zhuo, Yiran Chen, Hai Li

Duke

Motivation: Shared Subspaces at Decode Time

LLMs reuse computation across tasks, but current activation interventions often miss the states where decisions are made.

Multi-task generality suggests reuse

- LLM solves many tasks with one parameter set.
- Diverse behaviors may share internal features.
- This suggests a task-general decision subspace S.

Activation interventions are brittle

- *Prefill-estimated directions may **not** align with decode-time decisions.*
- Directions may interfere with task-general channels.

Hypothesis

*Under KV-Cached Decoding, next-token decisions **causally depend** on a small subspace S of decode-time hidden state that is **shared across tasks***

Duke

Testing the Hypothesis: Three Checks

We test whether the shared decode-time subspace exists, matters causally, and is specific to the decode regime.

H1. Existence

Question: Is there a compact subspace shared across tasks?

Test: Compare shared-set size $|S_\ell(\tau,m)|$ against permutation and scramble null baselines.

H2. Causality

Question: Does this subspace affect next-token decisions?

Test: Remove the estimated shared subspace only during KV-cached decoding and compare with dimension/energy-matched controls.

H3. Decode Specificity

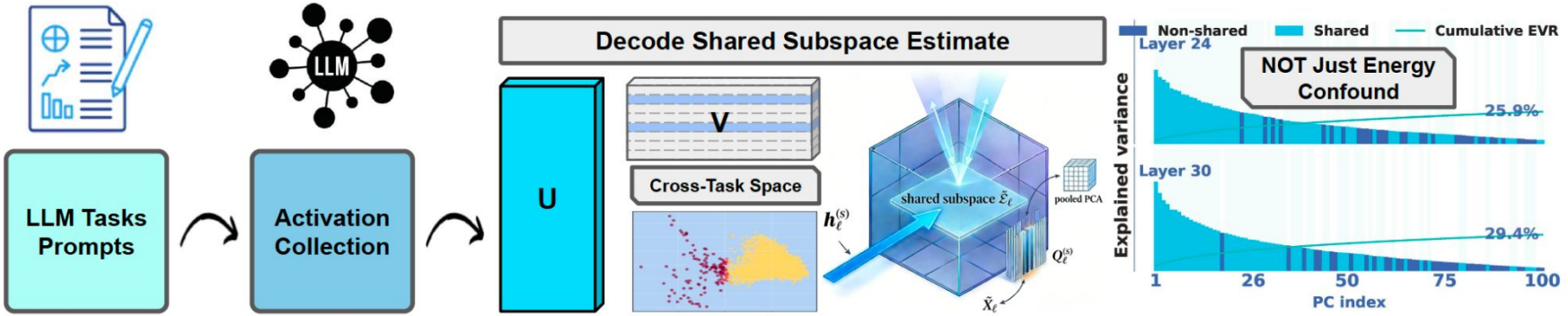
Question: Can prefill-estimated directions reproduce the same effect?

Test: Match rank and intervention budget, then compare decode-estimated vs. prefill-estimated bases at the decode locus.

Duke

DecodeShare Protocol

Estimate shared structure from decode-time states, then test it with decode-only interventions.



01

Collect

Pool decode-time hidden states across tasks and prompt variants.

02

Estimate

Select directions consistently shared across tasks, not just high-energy PCs.

03

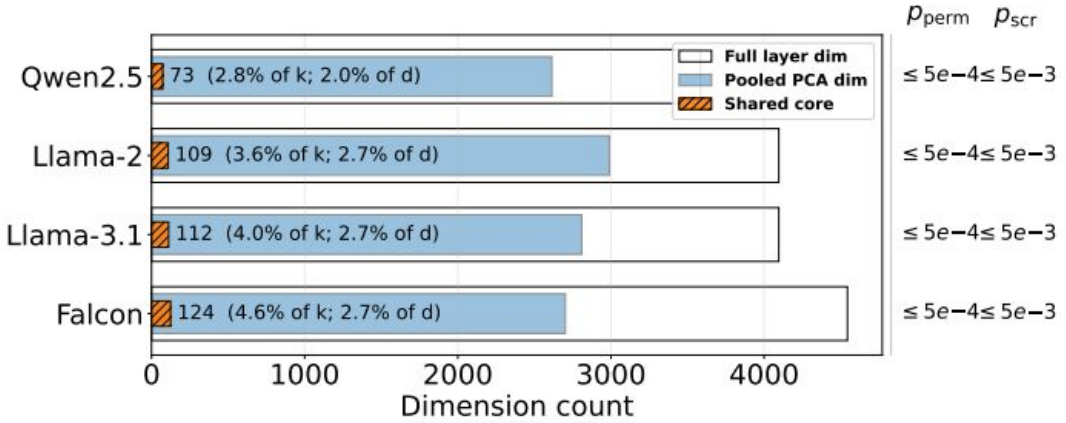
Intervene

Run decode-only projection removal against dimension- and energy-matched controls.

Duke

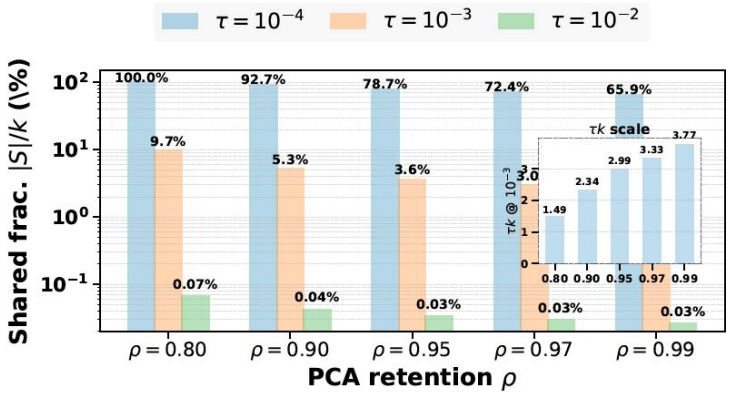
Main finding 1: shared subspace exists and is causal

Small but significant shared core



Takeaways. Only ~2–3% of layer width, yet statistically above matched nulls.

Stable under threshold/retention sweeps



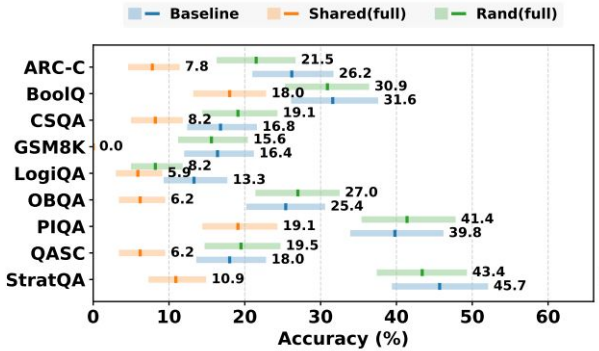
Takeaways. Default $\tau = 10^{-3}$ lies in the stable compact-core regime

The shared set is small, statistically significant, and not an artifact of one threshold choice.

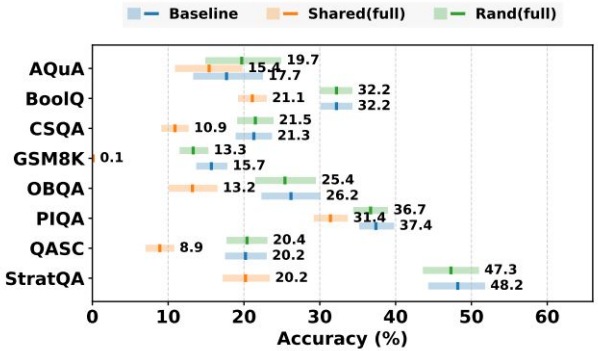
Duke

Main finding 1: shared subspace exists and is causal

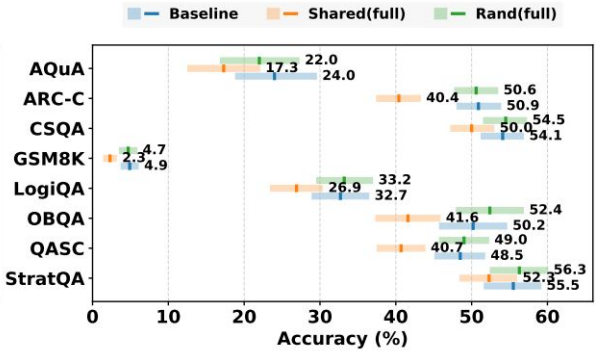
All-task generation



Leave-one-task-out generation



Leave-one-task-out forced-choice



Note. LOTO = leave-one-task-out subspace estimation. Generation scores model outputs directly; forced-choice scores candidate answers by conditional log-probability.

Takeaways. Decode-only removal of the shared subspace consistently harms accuracy, while matched controls stay near baseline.

Duke

Main finding 2: Decode Alignment Matters !!

Prefill-Decode Mismatch

		Intervene: Decode			Intervene: Prefill		
Task	**Base**	**Dec-est**	**Pre-est**	**Rand**	**Dec-est**	**Pre-est**	**Rand**
CSQA	53.3	**23.1**	54.3	53.5	53.2	53.3	53.3
StratQA	49.6	52.8	**49.6**	49.6	49.8	49.6	49.6
PIQA	69.1	**52.7**	67.6	69.4	69.5	69.2	69.2
ARC-C	51.5	**26.5**	50.9	51.9	52.6	52.6	51.6
OBQA	51.2	**27.2**	51.8	51.8	53.8	52.8	51.2
QASC	48.9	**13.7**	50.3	49.5	49.0	49.7	48.8
LogiQA	32.1	**23.2**	32.7	32.1	32.4	32.9	32.1
Mean Δ vs. base	–	**−19.5**	0.2	0.3	0.7	0.6	0.0

Downstream steering utility

Table 4. **Decode-aligned ranking better matches held-out decode utility.** The pools contain 32 CAA contrastive vectors, 64 instruction-derived vectors, 64 SAE feature directions, and 100 diagnostic directions.

Pool	Prefill ρ	Decode ρ	Δ
CAA contrastive	-0.370	**0.700**	+1.070
Instruction	0.172	**0.767**	+0.595
SAE features	-0.064	**0.594**	+0.659
Diagnostic	0.065	**0.700**	+0.635

ρ = how well a validation proxy ranks deployed steering utility. Higher is better.

Takeaways. Prefill directions do not transfer to decode-time decisions. Decode-estimated bases cause large drops; prefill-estimated bases stay near baseline. Decode-stage validation also selects better steering directions.

Duke

Questions?

Takeaways:

- Decode-time decisions rely on a compact shared subspace in decode-time decision states.
- Removing this subspace hurts accuracy, while matched controls stay near baseline.
- Decode-aligned estimation and evaluation make intervention and steering more reliable.

Contact: zishan.shao@duke.edu

Duke