Measuring Affinity between Attention-Head Weight Subspaces via the Projection Kernel

Hiroaki Yamagiwa[1*], Yusuke Takase[2], Hidetoshi Shimodaira[2,3]
[1]SB Intuitions [2]Kyoto University [3]RIKEN AIP
EMNLP 2026 Main *Work done while at Kyoto University

arXiv

Background

- **Composition Score (CS)** [1] measures head interaction but is dominated by large singular values.

Approach

- Compare head weight **subspaces** with the **Projection Kernel (PK)** [2].

Results

- PK recovers IOI (indirect object identification) head relations [3] better than CS.
- PK finds the hub head **L4H7**, which proves to be an **Identity Head**.
- KL to the random baseline shows how informative PK is.

Composition Score (CS) [1]

- Computed from the QK and OV matrices of head h:

$$\boldsymbol{W}_{\mathrm{QK}}^{h} = \boldsymbol{W}_{\mathrm{Q}}^{h\top}\boldsymbol{W}_{\mathrm{K}}^{h}, \quad \boldsymbol{W}_{\mathrm{OV}}^{h} = \boldsymbol{W}_{\mathrm{O}}^{h}\boldsymbol{W}_{\mathrm{V}}^{h}$$

$$\mathrm{CS}\left(\boldsymbol{W}^{h'}, \boldsymbol{W}^{h}\right) = \frac{\left\|\boldsymbol{W}^{h'}\boldsymbol{W}^{h}\right\|_{\mathrm{F}}}{\|\boldsymbol{W}^{h'}\|_{\mathrm{F}}\|\boldsymbol{W}^{h}\|_{\mathrm{F}}}$$

- e.g., OQ (output of $h \rightarrow$ query of later h'): $\mathrm{CS}\left(\boldsymbol{W}_{\mathrm{QK}}^{h'\top}, \boldsymbol{W}_{\mathrm{OV}}^{h}\right)$

Limitations

- Weighted by singular values: a few directions dominate.
- Not invariant to $\boldsymbol{W}^{h'}\boldsymbol{A}\boldsymbol{A}^{-1}\boldsymbol{W}^{h}$ and sensitive to weight scaling and preprocessing.

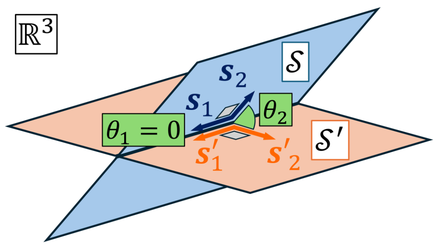
Projection Kernel (PK) [2]

Principal Angles

Orthonormal bases $\boldsymbol{U}, \boldsymbol{U}'$ of $\mathcal{S}, \mathcal{S}'$:

$$\cos\theta_i = \sigma_i(\boldsymbol{U}^{\top}\boldsymbol{U}')$$

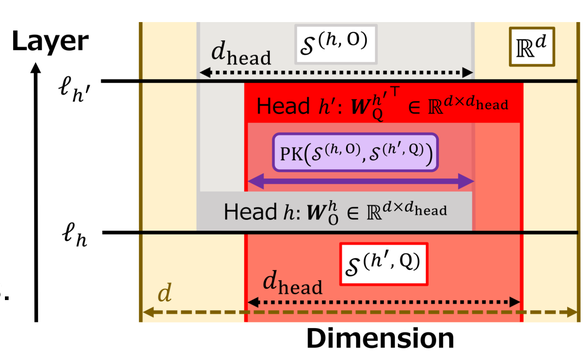
σ_i: i-th largest singular value of $\boldsymbol{U}^{\top}\boldsymbol{U}'$.



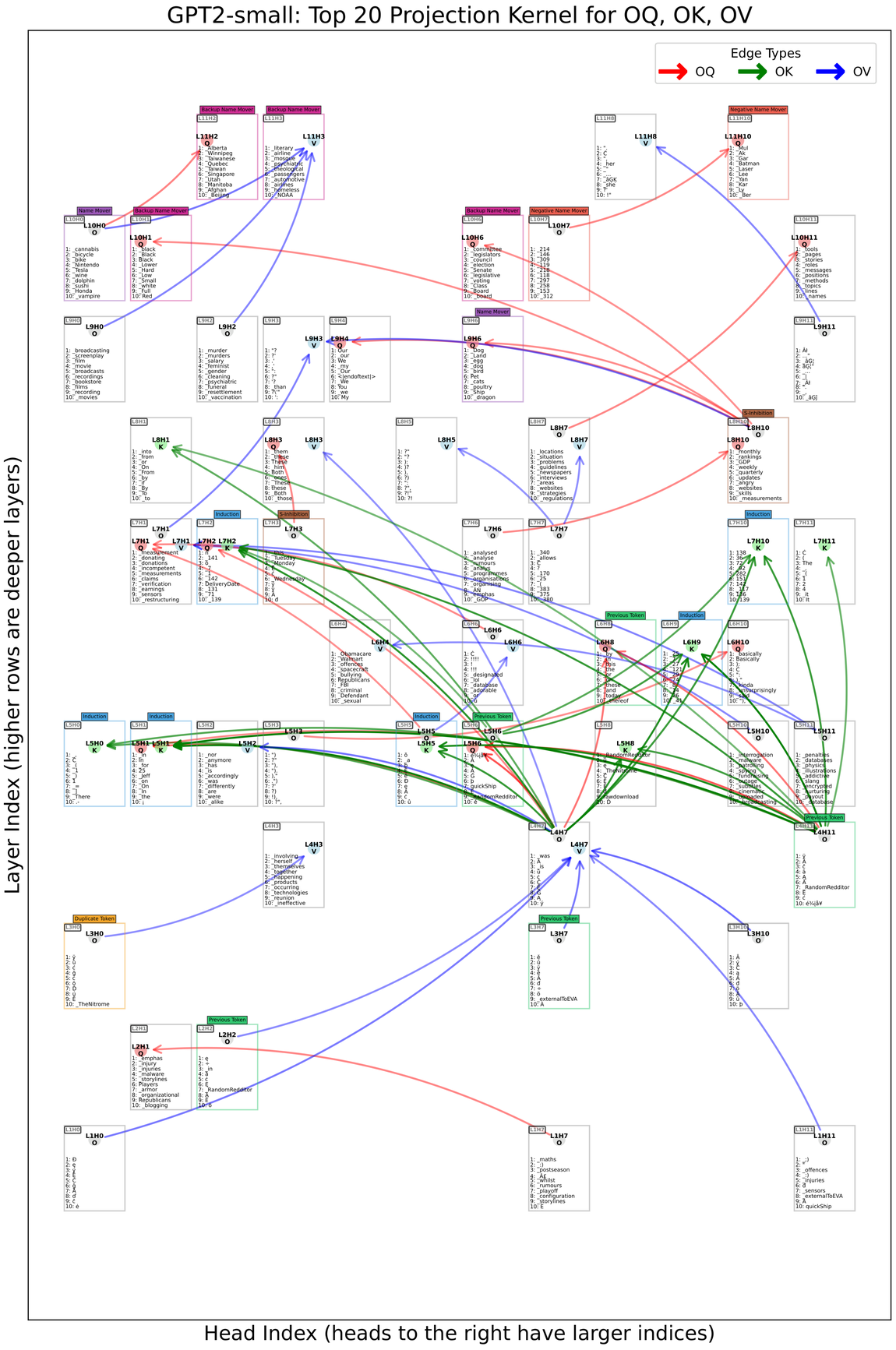
Projection Kernel

$$\mathrm{PK}(\mathcal{S}, \mathcal{S}') = \sum_{i=1}^{m} \cos^2\theta_i = \|\boldsymbol{U}^{\top}\boldsymbol{U}'\|_{\mathrm{F}}^2$$

- $0 \le \mathrm{PK} \le m$ ($\mathrm{PK} = m$ iff $\mathcal{S} = \mathcal{S}'$).
- No direction-dependent scaling.
- Used on each head's Q/K/V/O subspaces.



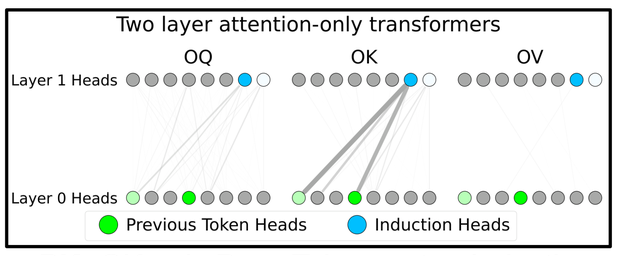
GPT2-small: Top 20 Projection Kernel for OQ, OK, OV



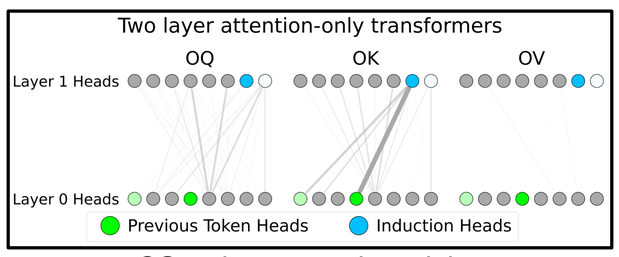
OQ L8H10 → Name Movers **OK** Prev. Token → Induction **OV** many heads → **L4H7**
Each head is annotated with the top 10 tokens representing its $\boldsymbol{W}_{\mathrm{O}}^{h}$.

Comparing PK and CS

2-layer attn-only model



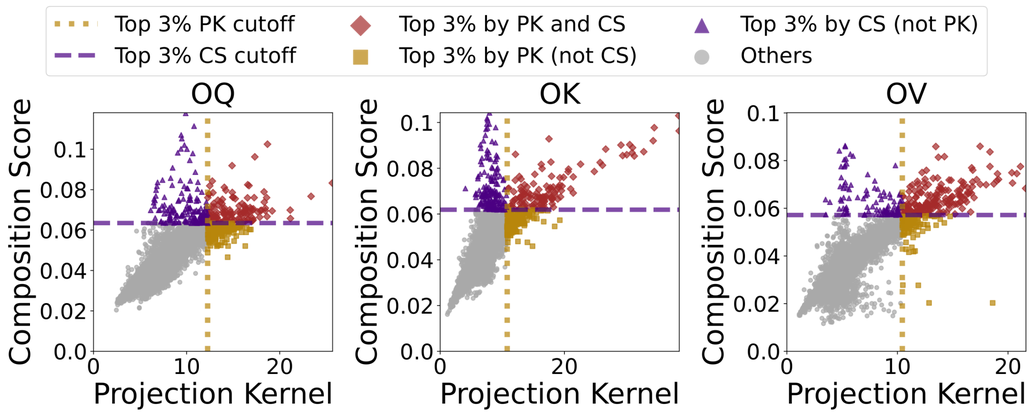
PK: OK only Prev. Token → top Induction



CS: misses a pair; noisier

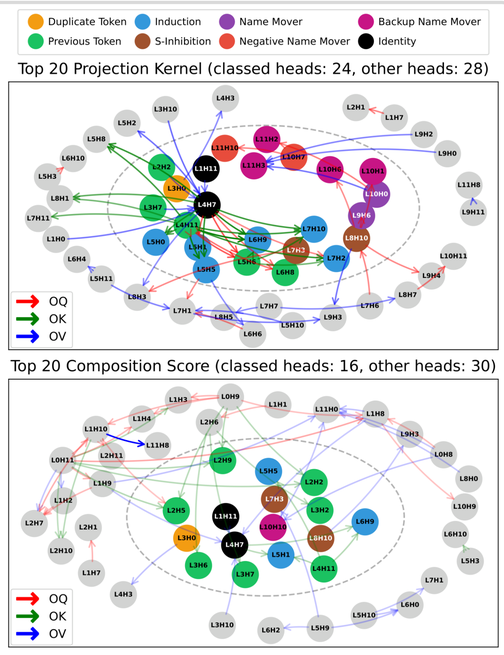
Wiring diagrams

GPT2-small



(a) PK vs. CS scores

a. Correlated overall ($\rho \approx 0.84$–0.89), **weakly in the top 3%** ($\rho = -0.25$ to 0.19).
b. PK edges concentrate among IOI head classes; CS edges often involve unclassified heads.



(b) Wiring diagrams

Evaluation on IOI Heads

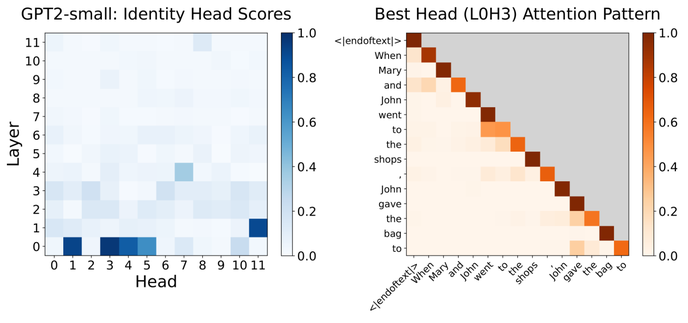
	Detection			Classification		
Method	**GPT2**	**BLOOM**	**Qwen2**	**GPT2**	**BLOOM**	**Qwen2**
PK	**0.395**	**0.268**	**0.251**	**0.047**	**0.027**	**0.032**
CS	0.265	0.141	0.190	0.013	0.003	0.005
Simple-CS	0.303	<u>0.224</u>	<u>0.222</u>	0.016	<u>0.015</u>	0.004
Linear CKA	<u>0.356</u>	0.155	0.092	0.021	0.001	<u>0.024</u>
Proc. Sim.	0.342	0.147	0.091	<u>0.038</u>	0.001	<u>0.024</u>

- Models: GPT2-small, BLOOM-560M, and Qwen2-0.5B-Instruct.
- PK achieves the best PR-AUC in all three models.

References
[1] Elhage et al. A Mathematical Framework for Transformer Circuits. 2021.
[2] Hamm & Lee. Grassmann Discriminant Analysis: a Unifying View on Subspace-Based Learning. ICML 2008.
[3] Wang et al. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 Small. ICLR 2023.

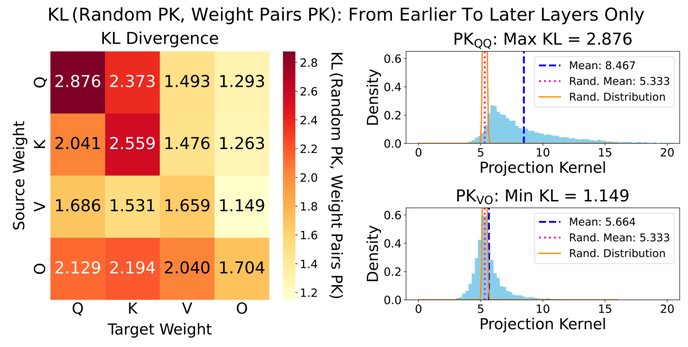
Discussion

L4H7 is an Identity Head



- L4H7 is a hub head in the PK wiring diagram of GPT2-small.
- L4H7 is an Identity Head; a similar trend is observed in BLOOM.

PK distributions vs. random



- Random baseline (orthogonal subspaces):

$$\mathrm{PK} \sim \mathcal{N}\left(\frac{{d_{\mathrm{head}}}^2}{d}, \frac{2{d_{\mathrm{head}}}^2(d-d_{\mathrm{head}})^2}{d^2(d-1)(d+2)}\right)$$

- High KL: O→Q/K/V, same-type, and QK/KQ.