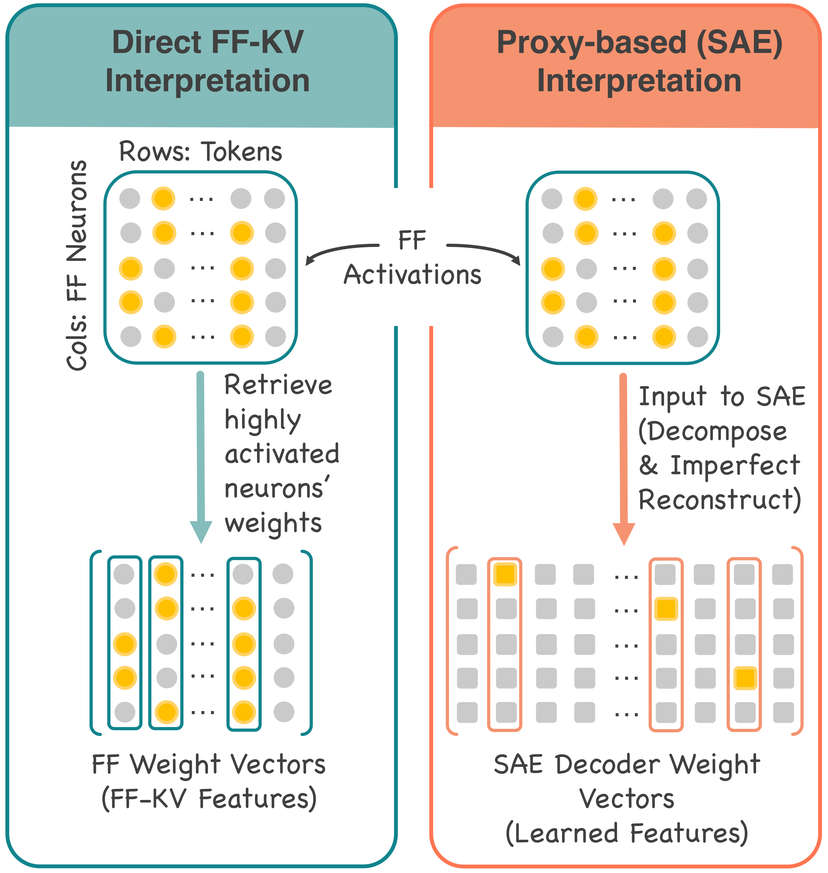
NEURAL INFORMATION PROCESSING SYSTEMS

TOHOKU NLP GROUP

Key-Value Memories (FF-KVs) in Transformer Feed-Forward Layers are a **simple but strong baseline** for feature discovery-based (SAE-based) interpretability research.

Comparison of Feature Discovery Sources



Do proxy-learned features truly offer better interpretability?

- We compared FF-KVs (vanilla & variants) against SAEs and Transcoders using automatic evaluation and human evaluation
- SAEs & Transcoders offer observable but minimal improvements on interpretability compared to FF-KVs

Transformer Key-Value Memories Are Nearly as Interpretable as Sparse Autoencoders

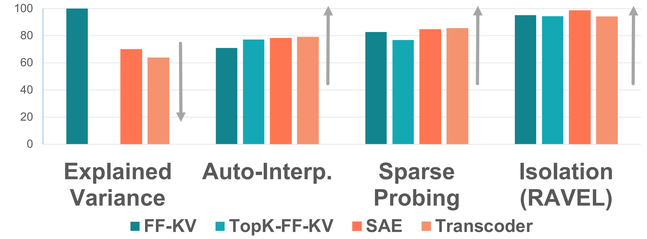
Mengyu Ye[1], Jun Suzuki[1,2], Tatsuro Inaba[3]
Tatsuki Kuribayashi[3]

[1]Tohoku University [2]RIKEN [3]MBZUAI

Paper

Automatic Evaluation

In addition to FF-KV, SAE and Transcoder, we also evaluated FF-KV with a TopK activation applied (TopK-FF-KV), on various metrics



Human Evaluation

We found a similar number of conceptual features (features that correspond to a certain concept rather than superficial pattern) inside FF-KVs, SAEs and Transcoders.

	Superficial Feature	Conceptual Feature	Uninterpretable Feature
FF-KV	6	8	36
TopK-FF-KV	9	9	32
SAE	6	9	35
Transcoder	16	11	23

Feature Alignment Analysis

Most features between FF-KV and Transcoder are unaligned. Perhaps, Transcoder features are hallucinated...?