

Open-Source Models Also Show Happiness, But Trajectories Can Differ

Where do Models Feel *Happiness*? Emotion-vectors in Open-Source Models

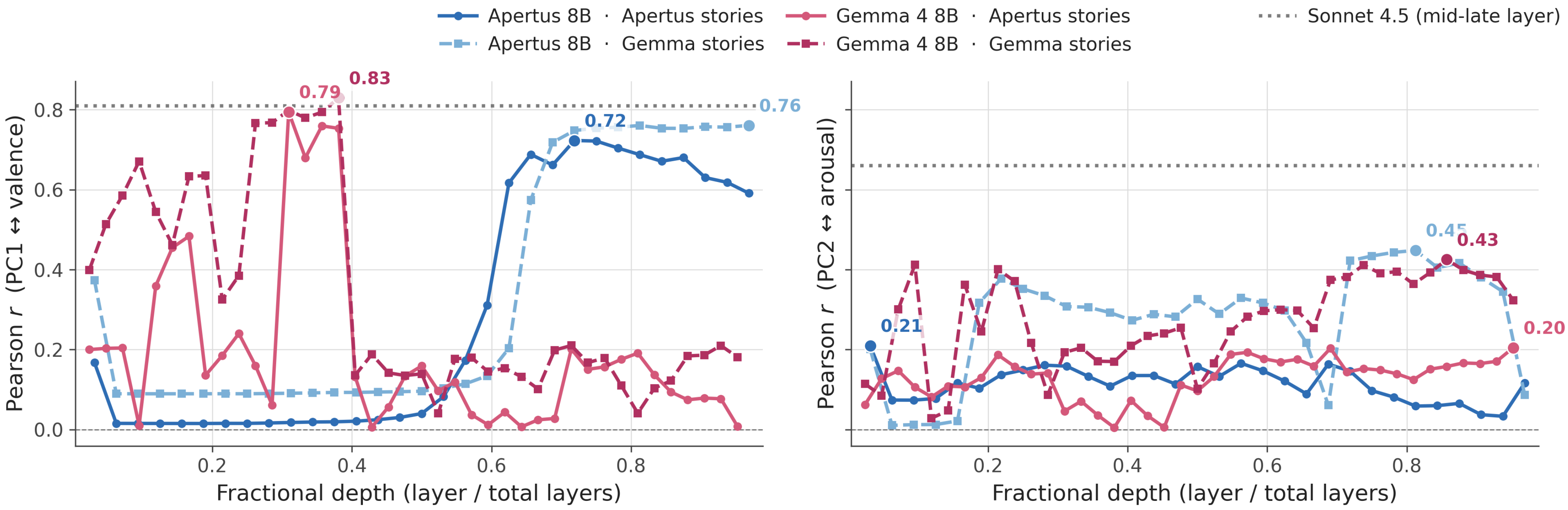
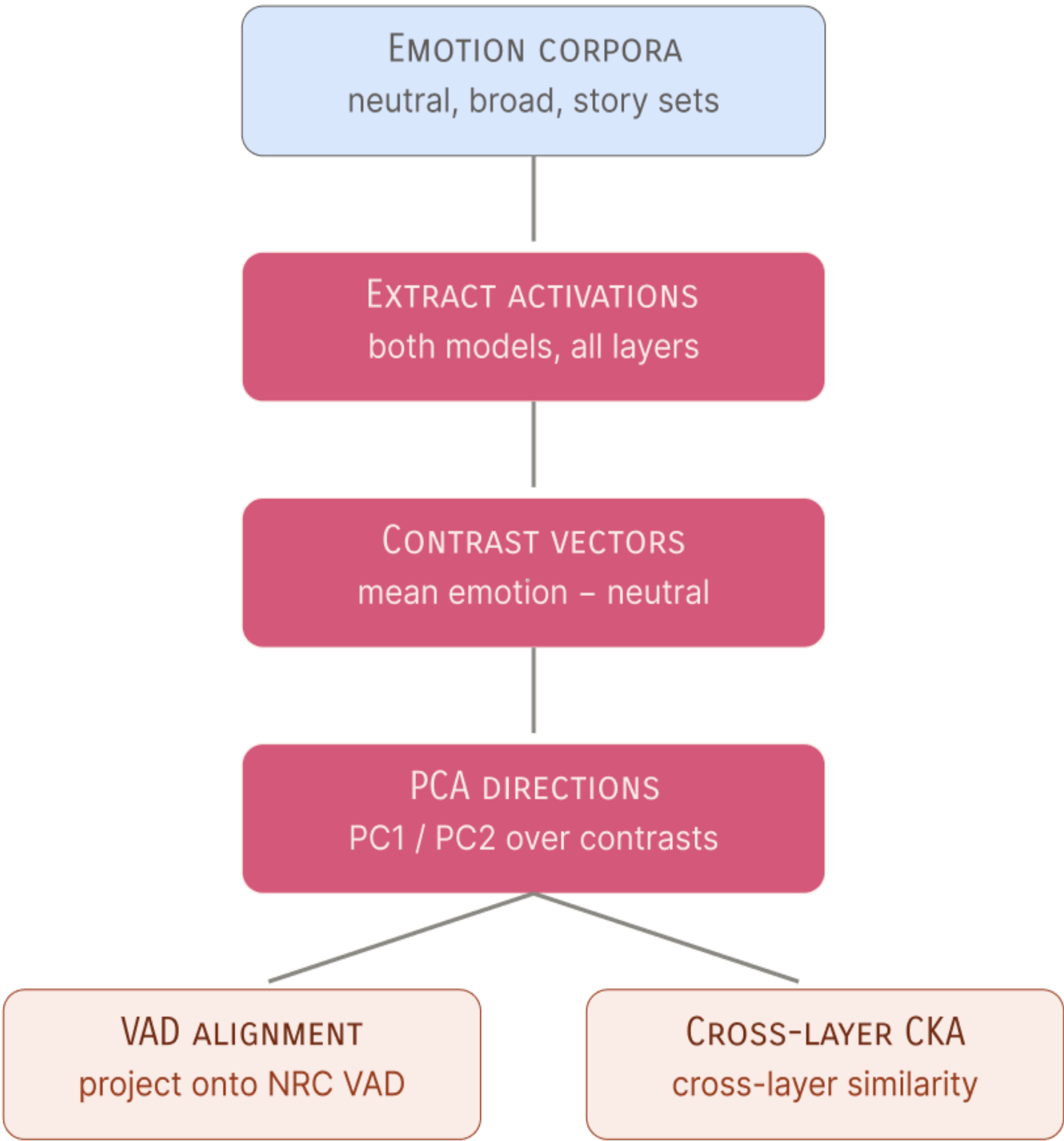
Problem

How generalisable are “**emotion vectors**” from Claude Sonnet 4.5 in open-weight models, across all layers, using two model-generated corpora?



Check out the open-sourced code!

Statistic	Apertus stories	Gemma stories
Generator model	APERTUS-8B-Instruct-2509	GEMMA-4-E4B-it
Total emotions	171	171
Unique topics	100	100
Stories per emotion	9	9
Total stories	1,539	1,539
Mean story length	~215 words	~144 words
Story length range	65–665 words	81–298 words



Different paths to the same geometry

GEMMA-4-E4B and APERTUS-8B reach **similar peak valence correlations** ($r \approx 0.76\text{--}0.83$) via different layer-wise trajectories.

Stable representation space, unstable axis

The representational space (CKA) and valence-axis stability are **disconnected**. Representational stability between layers does not guarantee a shared valence direction

The arousal gap and corpus dependence

Because Gemma stories improve arousal extraction in **both models**, the effect likely reflects corpus properties rather than model-story matching.



PAPER

Sinie van der Ben, Raphaël Baur, Yannick Metz, Mennatallah El-Assady

Mechanistic Interpretability Workshop, Mechanistic Interpretability Workshop ICML'26

