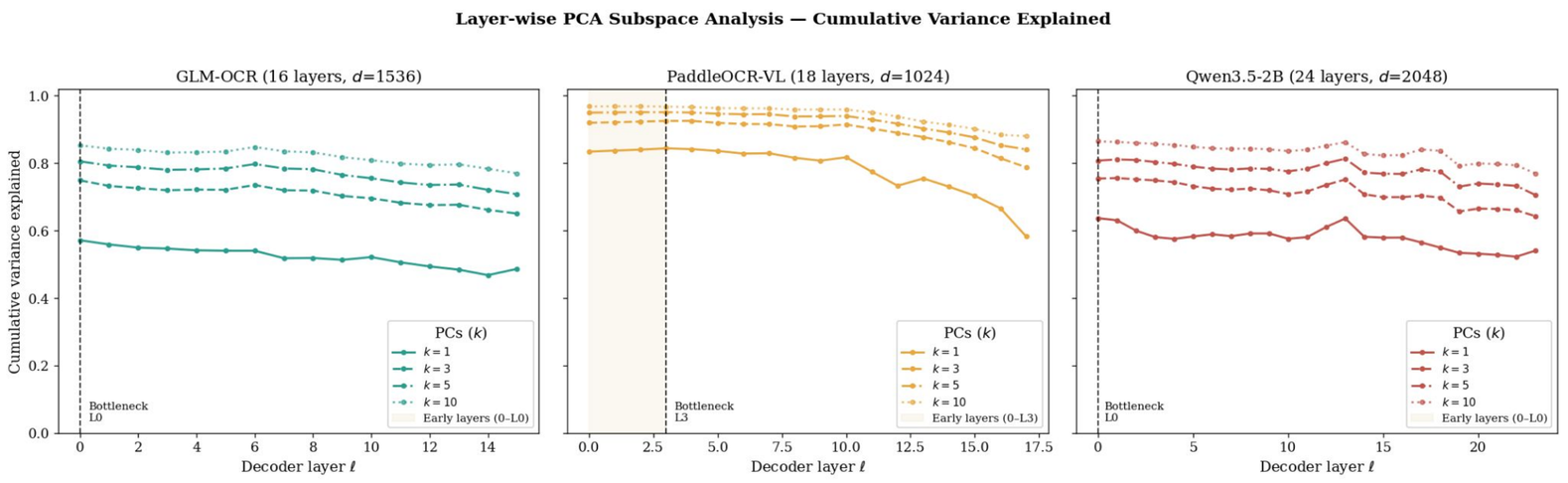
OCR models exhibit concentrated, low-dim document structure subspaces compared to general-purpose vision-language models, despite high cross-architecture alignment.

Low-Dimensional Document Structure Subspaces in Specialized vs. Emergent OCR Models: An Interpretability Study of Three Architectures

Background: Optical character recognition (OCR) models are critical infrastructure, but their internal representations remain unexplored. We compared those with general-purpose document recognition systems.

Result 1: Extreme Representational Compression

All three models concentrate OCR information into low-dimensional subspaces at early decoder depths, especially the specialized OCR models.

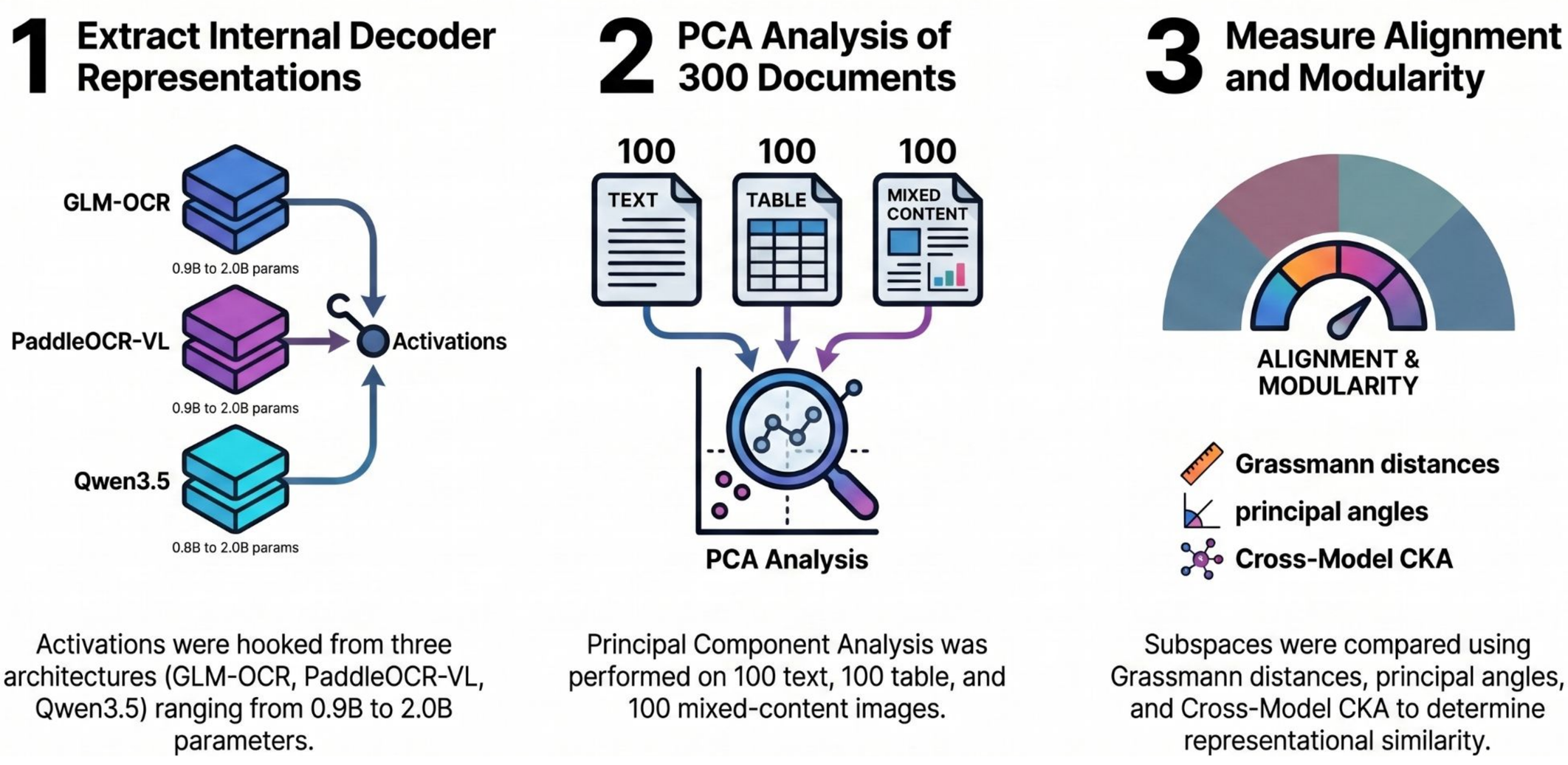


Result 2: Document Structure Modularity Index: *M*

M formalizes the representational organization. Both specialized models achieve higher *M* than the baseline, showing that representations for text, tables, and mixed content occupy partially disentangled subspaces.

Model	GLM	Paddle	Qwen
M	.774	.715	.704

Methods



Limitation: Our projection interventions are activation ablations rather than full causal tracing, establishing that identified subspaces are necessary for OCR performance, but not proving sufficiency.

PAPER

Guus Bouwens

Mechanistic Interpretability Workshop, ICML, Seoul, South Korea, 2026

ICML International Conference On Machine Learning