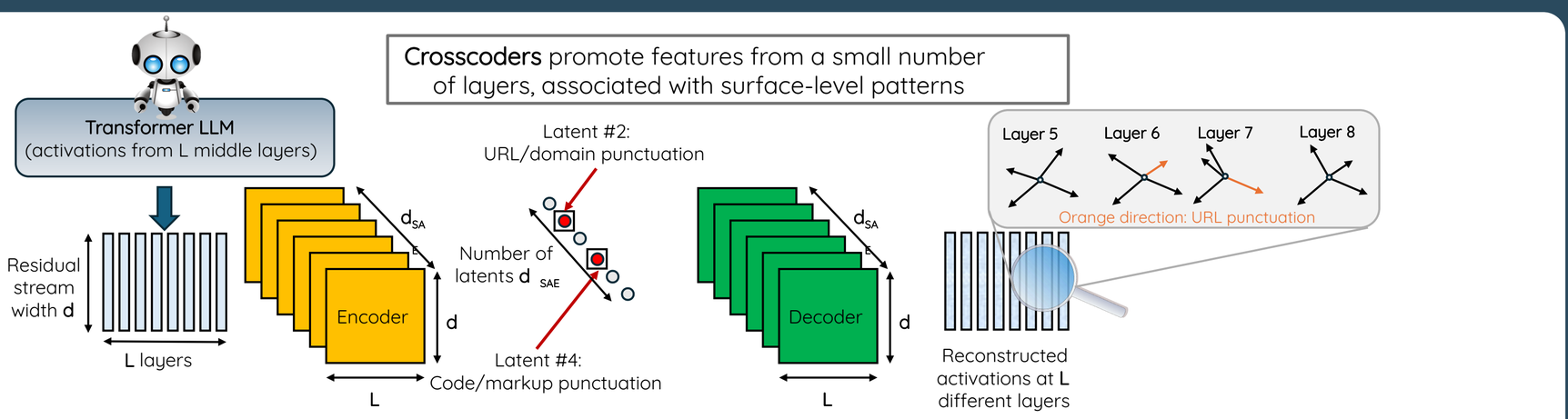


***Tensor factorization & layer masking* enable the recovery of features that persist across layers**

`fmxcoders`: Factorized Masked Crosscoders for Cross-Layer Feature Discovery

The Cyprus Institute — Research · Technology · Innovation
ATHENA Research & Innovation Information Technologies

Andreas D. Demou Panagiotis Koromilas James Oldfield
Yannis Panagakis Mihalis A. Nicolaou

Imperial College London
University of Oxford

1 TL;DR

Main Takeaways:

- **Standard crosscoders** recover features that **only depend on a small subset of layers**.
- **fmxcoders** recover **more cross-layer features** by combining (i) **low-rank tensor decomposition** of the encoder and decoder weights, and (ii) **stochastic layer masking**.
- `fmxcoders` **reduce MSE reconstruction** by 25–50%, **raise probing F1** accross all tasks, and recover $\sim 10\times$ **more semantically coherent latents** compared to standard crosscoders.

2 Background and Motivation

- **SAEs** recover local features with a **layer-specific dictionary**.
- **Crosscoders** learn a single dictionary for **multiple layers** at once.
- Due to **limitations**, the functional dependence of each **crosscoder latent collapses onto two layers on average**.
- **Limitation (1)**: The encoder and decoder are parameterized as L **independent per-layer matrices** (no structural link).
- **Limitation (2)**: **No optimization pressure for** spreading latent **dependence on multiple layers**.

3 Methodology

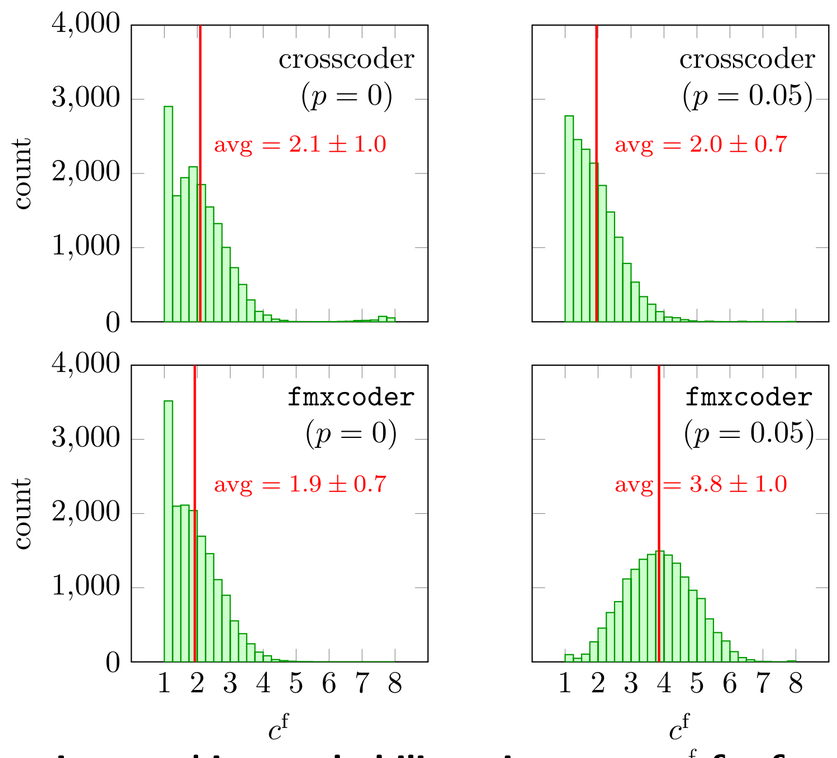
- Per-layer **encoder/decoder matrices are stacked into 3D tensors** $\mathbf{E}$ and $\mathbf{D}$.
- **Low-rank tensor factorization provides cross-layer structure**, addressing **Limitation (1)**. $\mathbf{E}$ and $\mathbf{D}$ are factorized using Tensor Ring decomposition, e.g.,

$$\mathbf{E}_{j,i,\ell} = \mathrm{Tr}\big(\mathbf{G}^{(1)}_{:,j,:}\,\mathbf{G}^{(2)}_{:,i,:}\,\mathbf{G}^{(3)}_{:,\ell,:}\big),$$

  where $\mathbf{G}$ are the decomposition factors.
- **Stochastic layer masking*** **biases learning toward the cross-layer regime**, addressing **Limitation (2)**.

* For each input token and each layer ℓ, we independently draw a Bernoulli mask $m_\ell \in \{0,1\}$ with $\Pr(m_\ell = 0) = p$, and the encoder receives the corrupted activations $\tilde{\boldsymbol{x}}_\ell = m_\ell \cdot \boldsymbol{x}_\ell$. The reconstruction loss targets the unmasked activations $\boldsymbol{x}_\ell$, meaning that the crosscoder is trained to reconstruct every layer, including the masked ones.

4 c^{f} measures the number of layers functionally influencing each latent



Increasing masking probability p increases c^{f} for fmxcoders

Latent firing pattern	Description	c^{f}
[*lobbying, campaigned, pleaded, pitch, ...*]	verbs of advocacy	7.1
[*regret, despise, failure, gruesome, ...*]	negative-affect words	7.0
[*Philadelphia, Pennsylvania, Maryland, Florida, ...*]	US states and cities	6.9
[*colleagues, friends, accomplice, partner, ...*]	companion relationship	6.8
[*1, 2, 3, 4, ...*]	digit detector	1.0
[*def, set, define, ->, ...*]	code/markup punctuation	1.0
[*d#, 3, :#, bbbb, ...*]	hex/colour-code characters	1.0
[*), ;, },), ...*]	closing-brace family	1.0

Larger c^{f} latents are more semantically coherent

5 Probing Evals

LLM	Variant	MSE ↓	EV ↑	CS ↑	Avg. F1 ↑
Pythia-1.4b	Crosscoder	0.31	0.9691	0.8976	47.0
	`fmxcoder`($p = 0$)	**0.22**	**0.9780**	**0.9273**	**78.4**
	`fmxcoder`($p = 0.05$)	0.23	0.9774	0.9252	71.2
Gemma2-2b	Crosscoder	3.29	0.7809	0.8737	51.2
	`fmxcoder` ($p = 0$)	**2.03**	**0.8632**	**0.9229**	**81.9**
	`fmxcoder` ($p = 0.05$)	2.07	0.8608	0.9218	73.0

fmxcoders improve avg. F1 and reconstruction across LLMs

6 Acknowledgments

The project is co-funded by the EU within the framework of the Cohesion Policy Programme "THALIA 2021-2027".

Co-funded by the European Union
Republic of Cyprus

PAPER

contact: a.demou@cyi.ac.cy