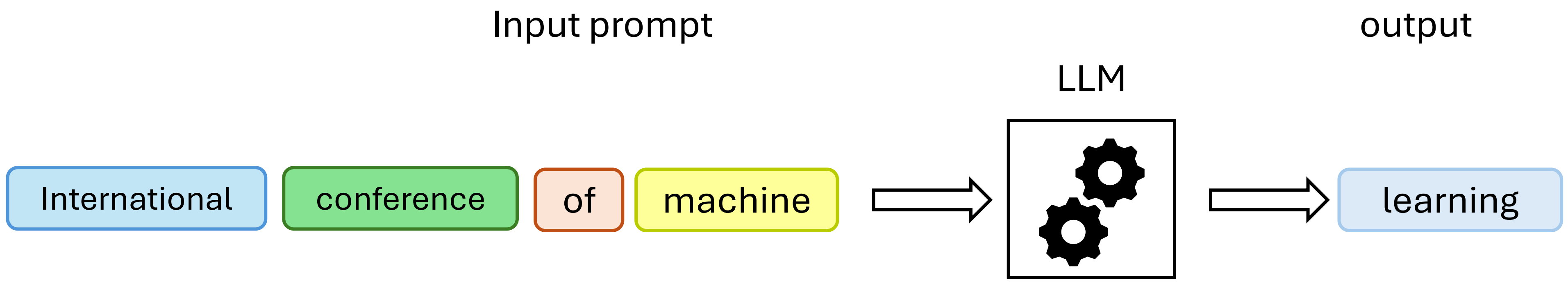
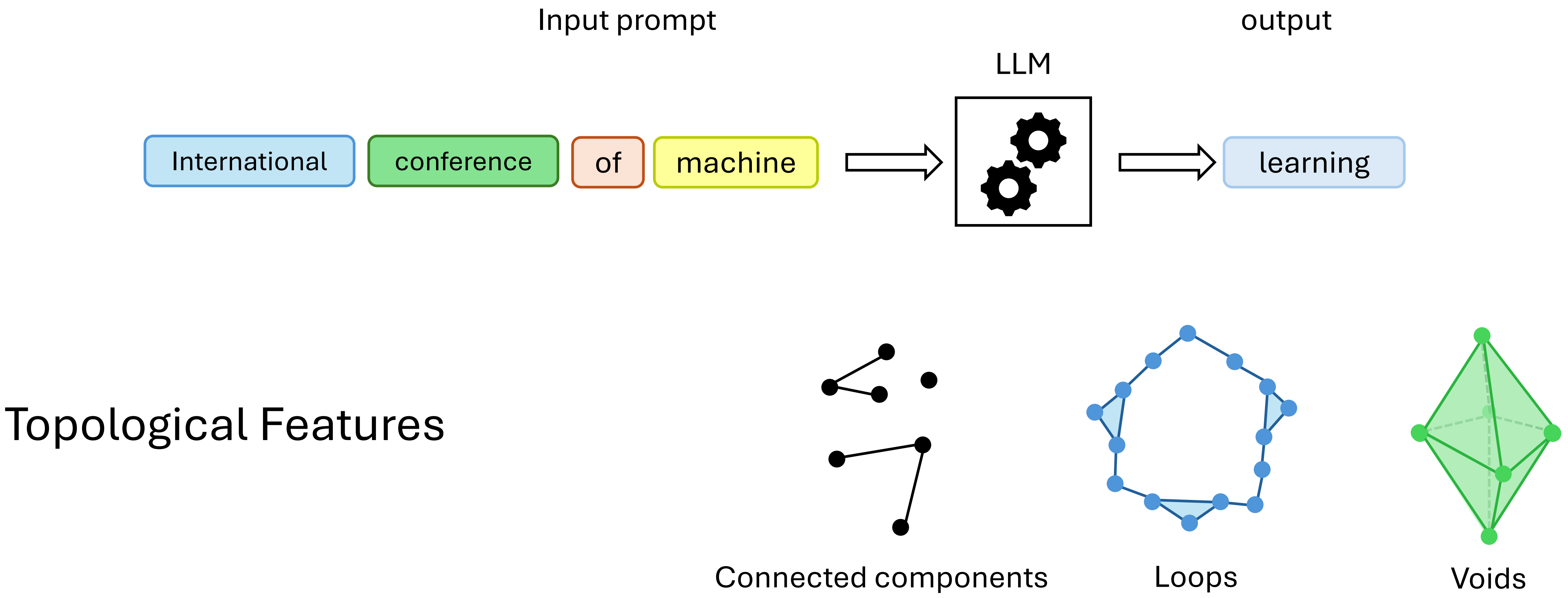
Persistent Topological Features in Large Language Models

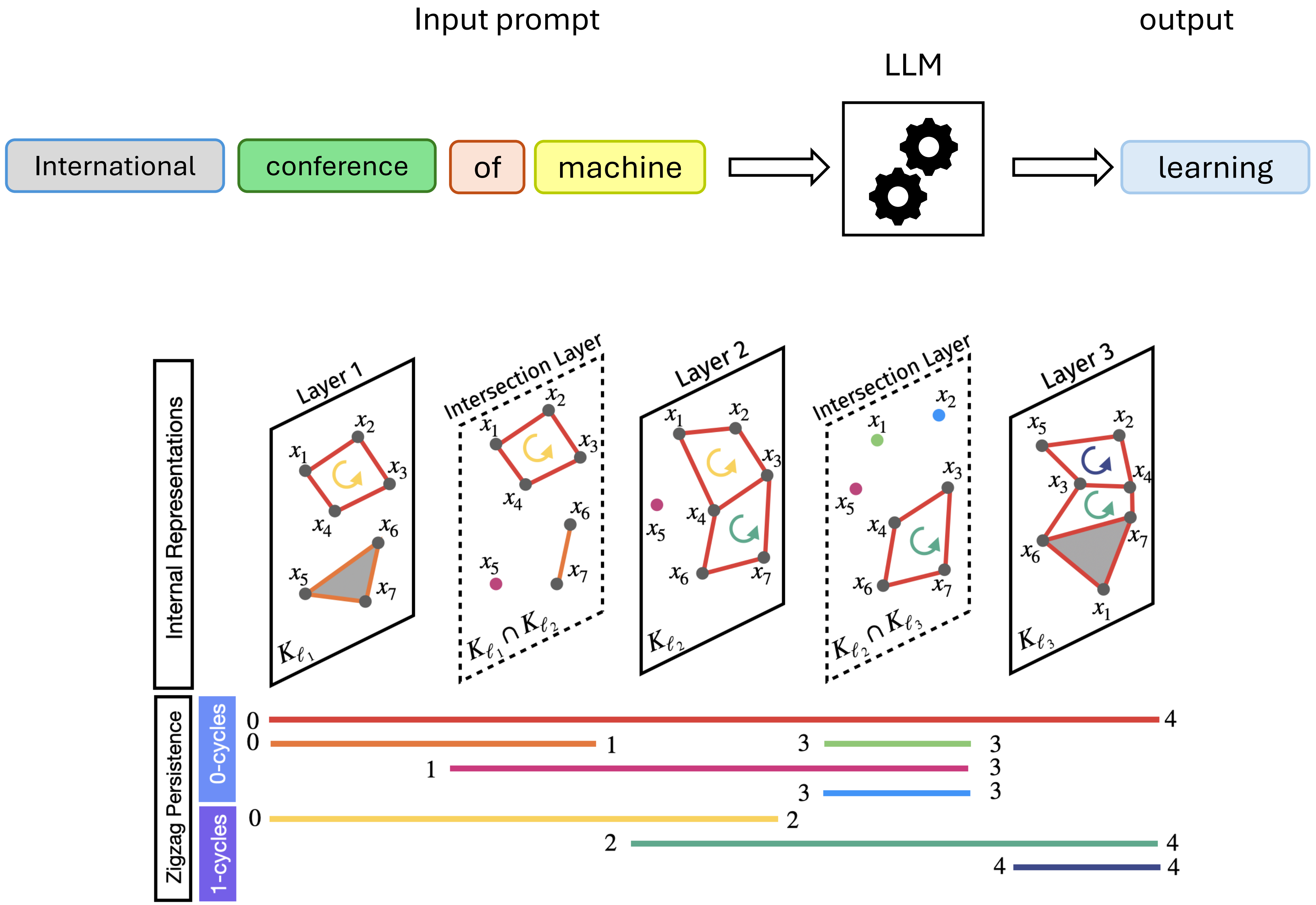
Yuri Gardinazzi[1,2], **Karthik Viswanathan**[1,3], Giada Panerai[1,2]
Alessio Ansuini[1],Alberto Cazzaniga[1], Matteo Biagetti[†,1]

[1]AREA Science Park, [2]University of Trieste, [3]University of Amsterdam

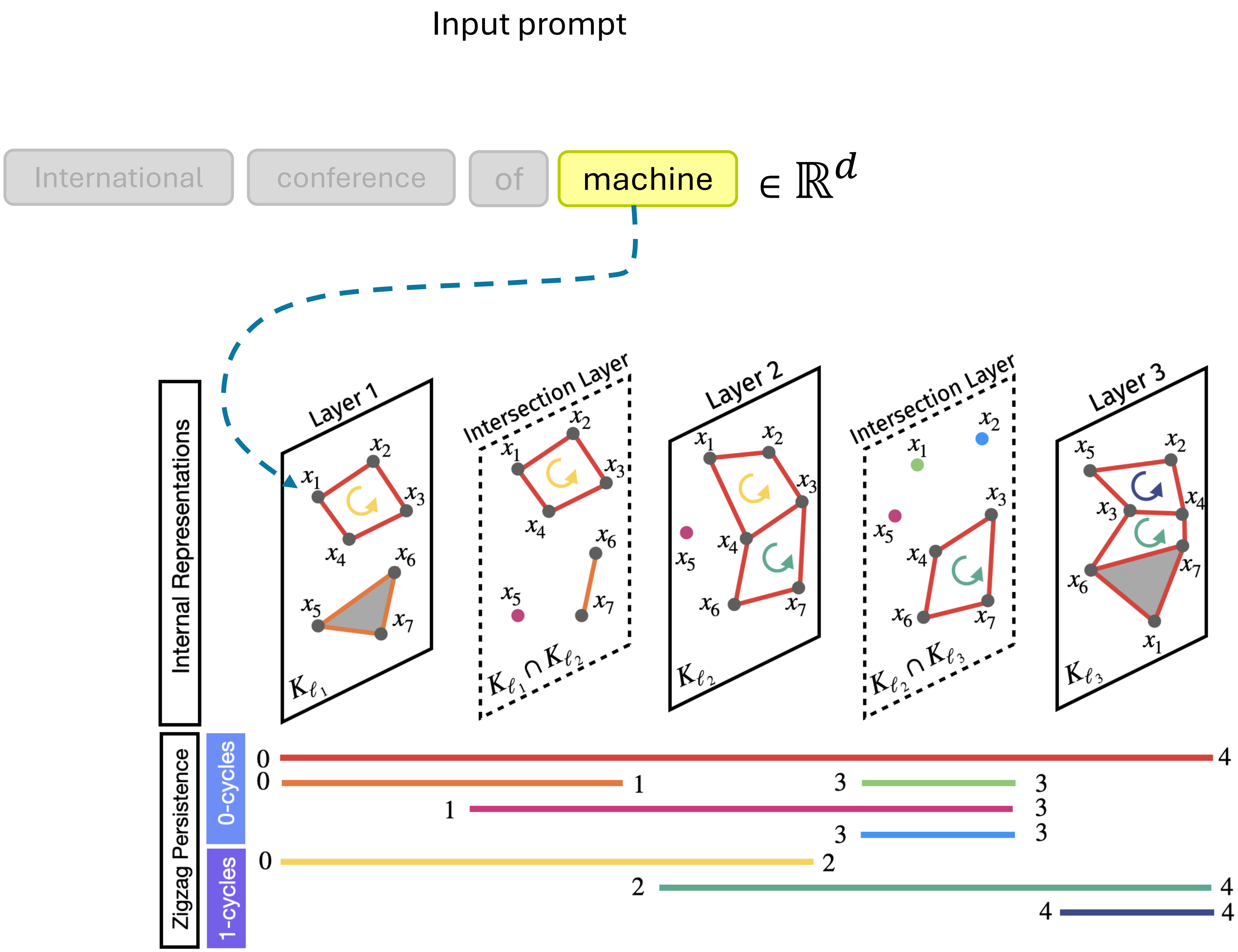
ICML 2025


Input prompt
output
LLM
International
conference
of
machine
learning


Input prompt
output
LLM
International
conference
of
machine
learning
Topological Features
Connected components
Loops
Voids

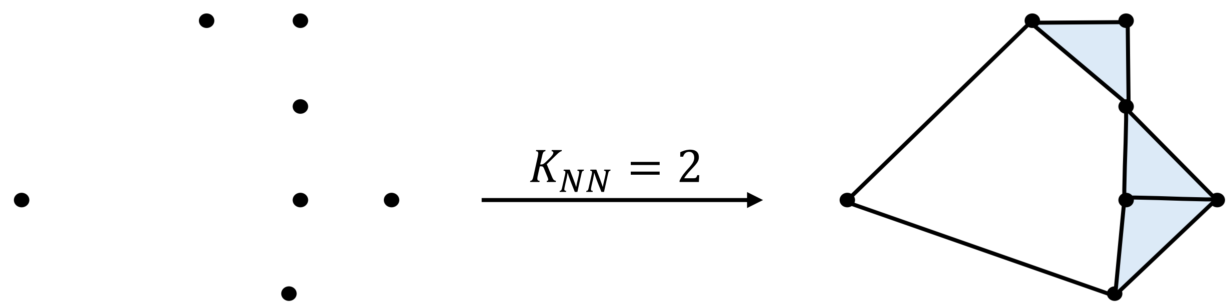


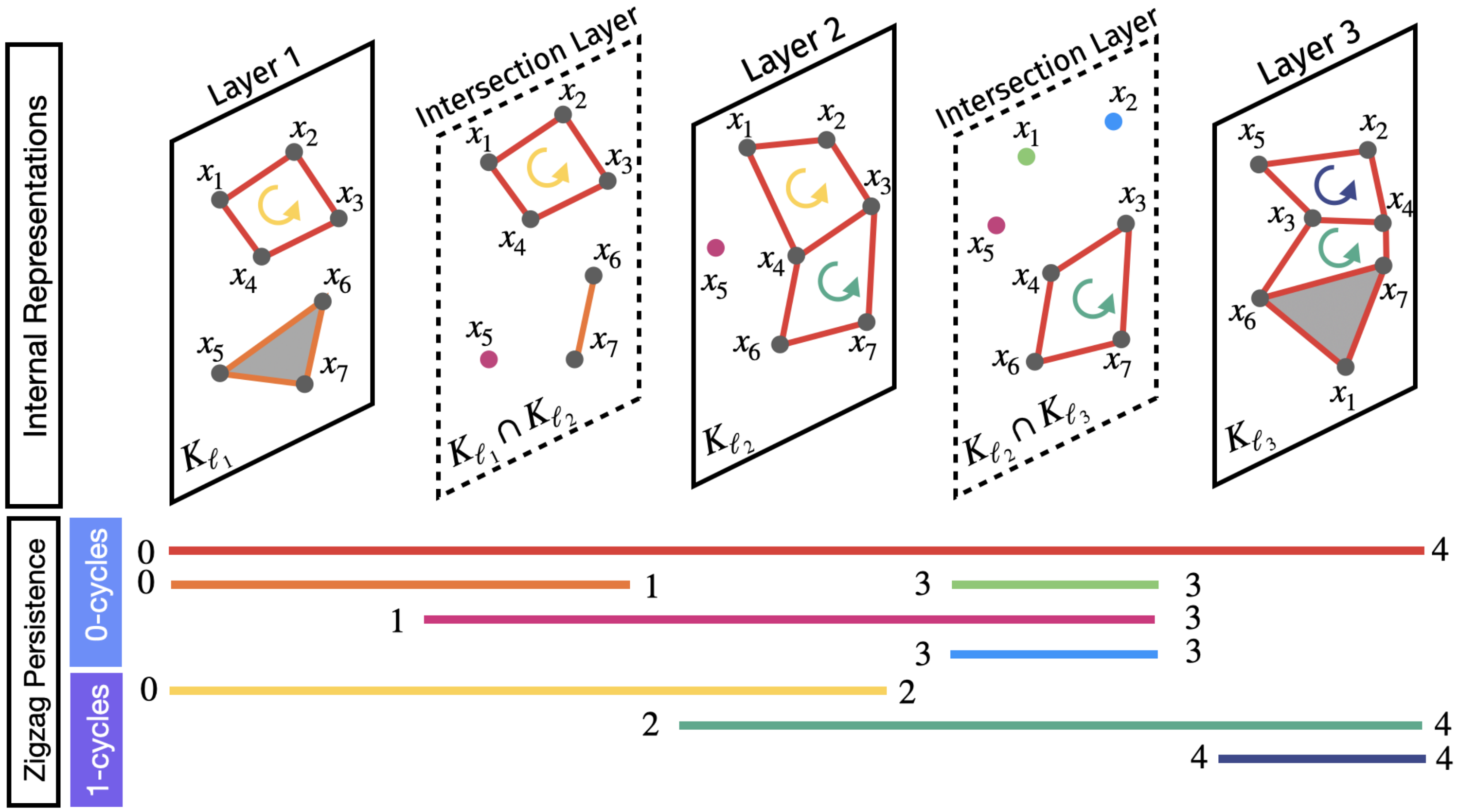
Input prompt
output
LLM
International
conference
of
machine
learning
Internal Representations
Layer 1
Intersection Layer
Layer 2
Intersection Layer
Layer 3
Zigzag Persistence
0-cycles
1-cycles



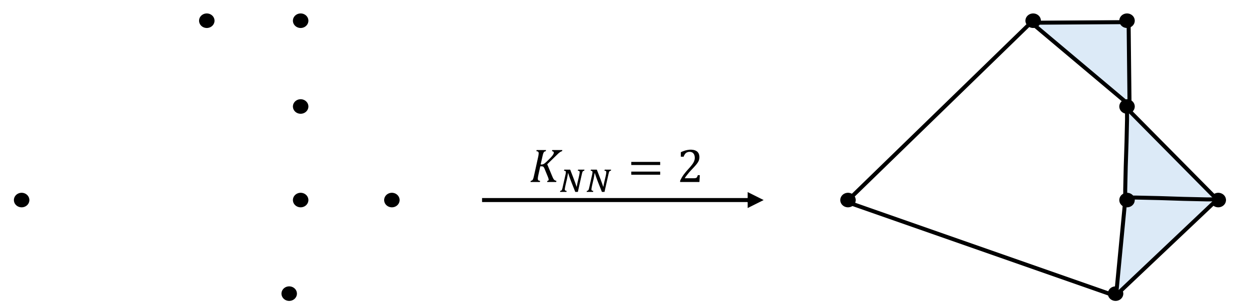
Input prompt
International
conference
of
machine
$\in \mathbb{R}^d$
Layer 1
Intersection Layer
Layer 2
Intersection Layer
Layer 3
K_{ℓ_1}
$K_{\ell_1} \cap K_{\ell_2}$
K_{ℓ_2}
$K_{\ell_2} \cap K_{\ell_3}$
K_{ℓ_3}
Internal Representations
Zigzag Persistence
0-cycles
1-cycles

Simplicial Complex

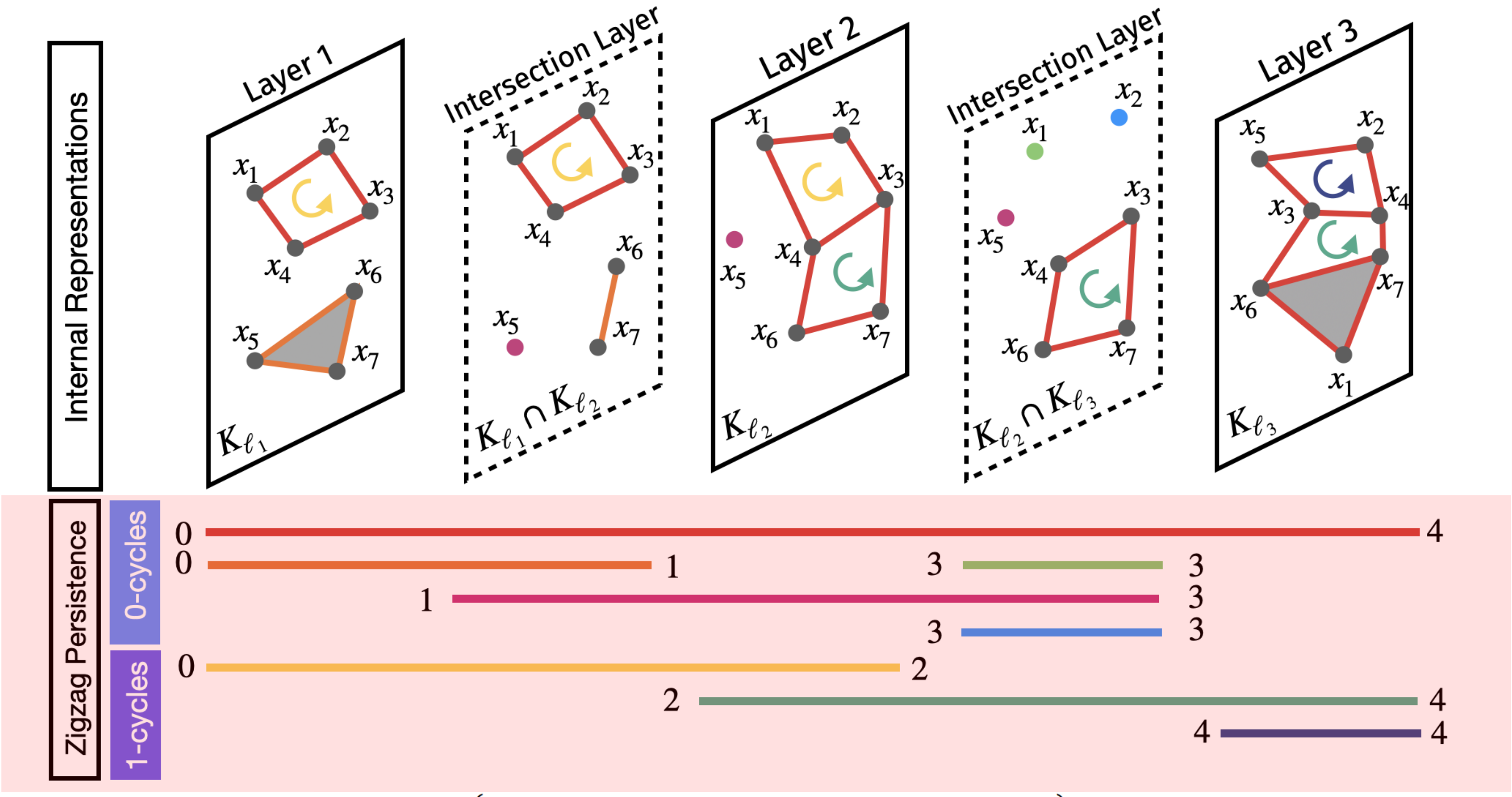




Simplicial Complex



p-dimensional holes



$$\mathrm{Pers}_p(\Phi) = \left\{ [\text{birth}, \text{death}] \mid \text{birth}, \text{death} \in \{0, \ldots, 2\mathrm{N}_{\text{layers}} - 1\} \right\}$$

Topological Descriptors

Births' relative frequency

$$B_p(\ell) = \frac{\sum_{\ell_i} \omega(\ell, \ell_i)\, \widehat{PI}_p(\ell, \ell_i)}{\sum_{\ell_i} \omega(\ell, \ell_i)\, \sum_{\ell_i} \widehat{PI}_p(\ell, \ell_i)}$$

$$\omega(\ell, \ell_i) = |\ell - \ell_i|^{\alpha}$$

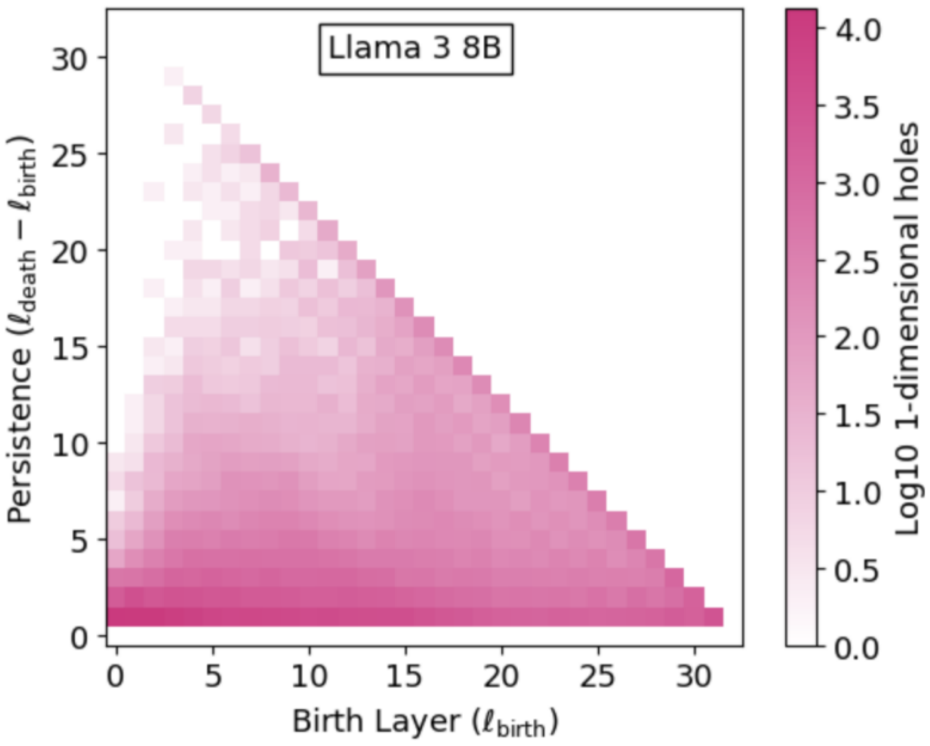
$$\mathcal{Z}_p(\ell_1, \ell_2) = \frac{\sum_{\ell_1 \leq M_1, \ell_2 > M_2} \widehat{PI}_p\,(\ell_1, \ell_2)}{\beta_p(\ell_1)}$$

Weighed Inter-Layer Persistence

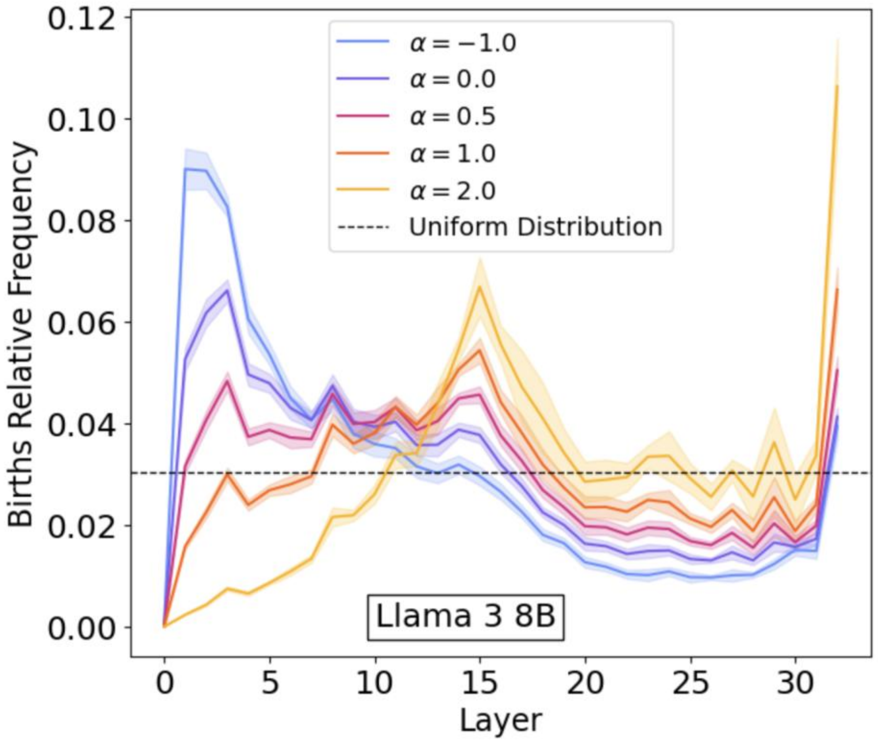
$$\bar{\mathcal{Z}}_p(\ell) = \frac{\sum_{\ell_i=1}^{N_{\mathrm{layers}}} \omega(\ell, \ell_i)\, \mathcal{Z}_p(\ell, \ell_i)}{\sum_{\ell_i=1}^{N_{\mathrm{layers}}} \omega(\ell, \ell_i)}$$

$$M_1 = \min(\ell_1, \ell_2)$$
$$M_2 = \max(\ell_1, \ell_2)$$

Results

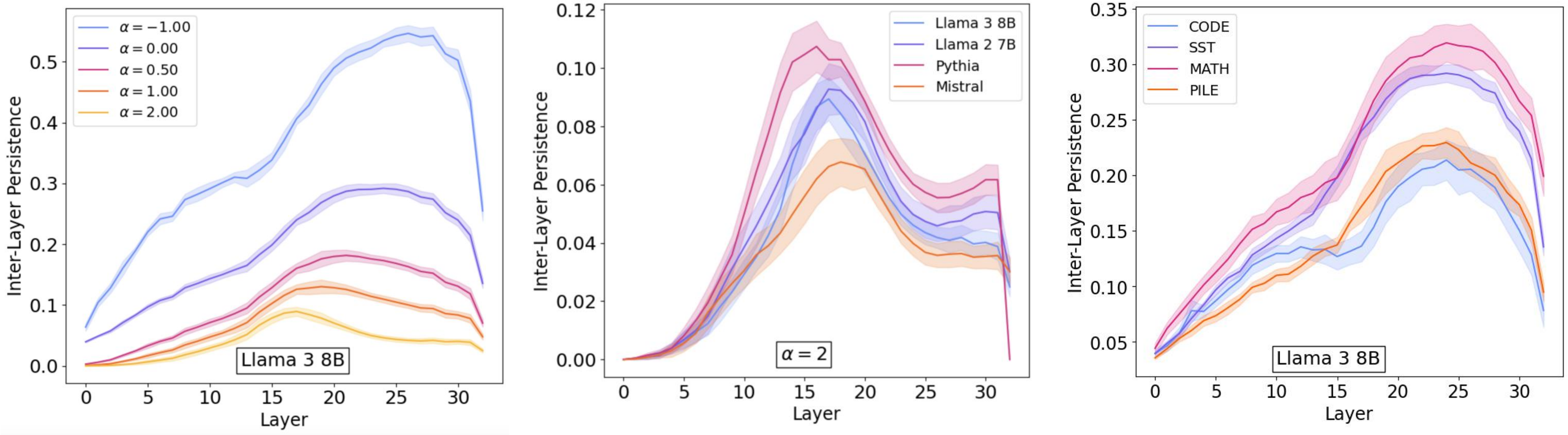


Llama 3 8B
Persistence ($\ell_{death} - \ell_{birth}$)
Birth Layer (ℓ_{birth})
Log10 1-dimensional holes



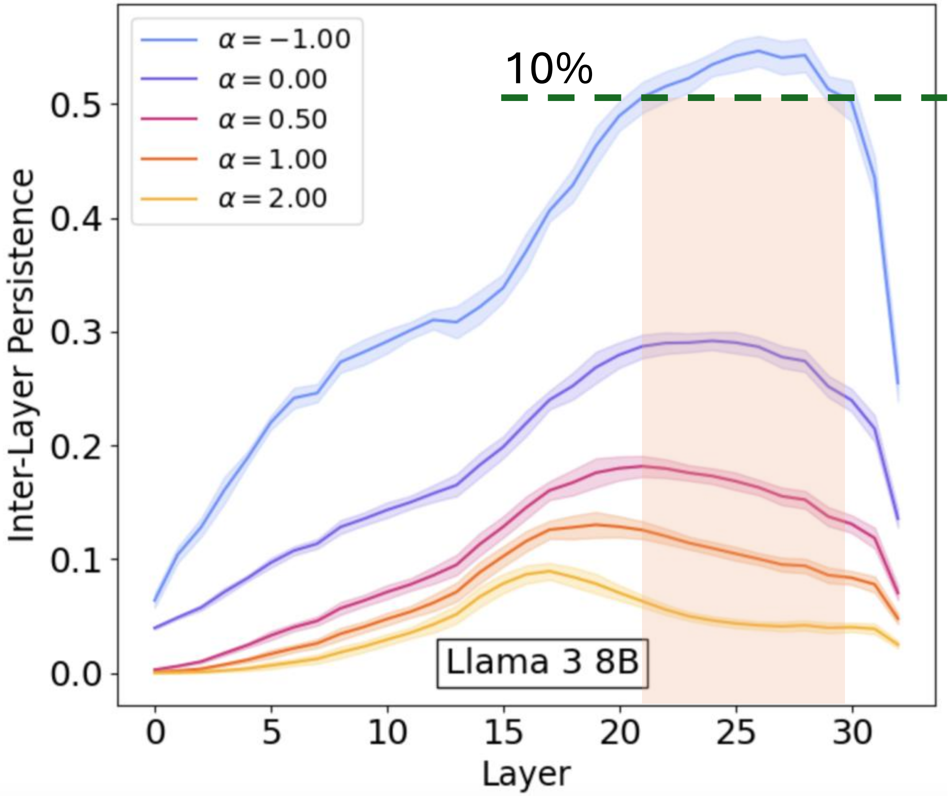
$\alpha = -1.0$
$\alpha = 0.0$
$\alpha = 0.5$
$\alpha = 1.0$
$\alpha = 2.0$
Uniform Distribution
Births Relative Frequency
Layer
Llama 3 8B

Results



$\alpha=-1.00$
$\alpha=0.00$
$\alpha=0.50$
$\alpha=1.00$
$\alpha=2.00$
Inter-Layer Persistence
Layer
Llama 3 8B
Llama 3 8B
Llama 2 7B
Pythia
Mistral
$\alpha=2$
CODE
SST
MATH
PILE

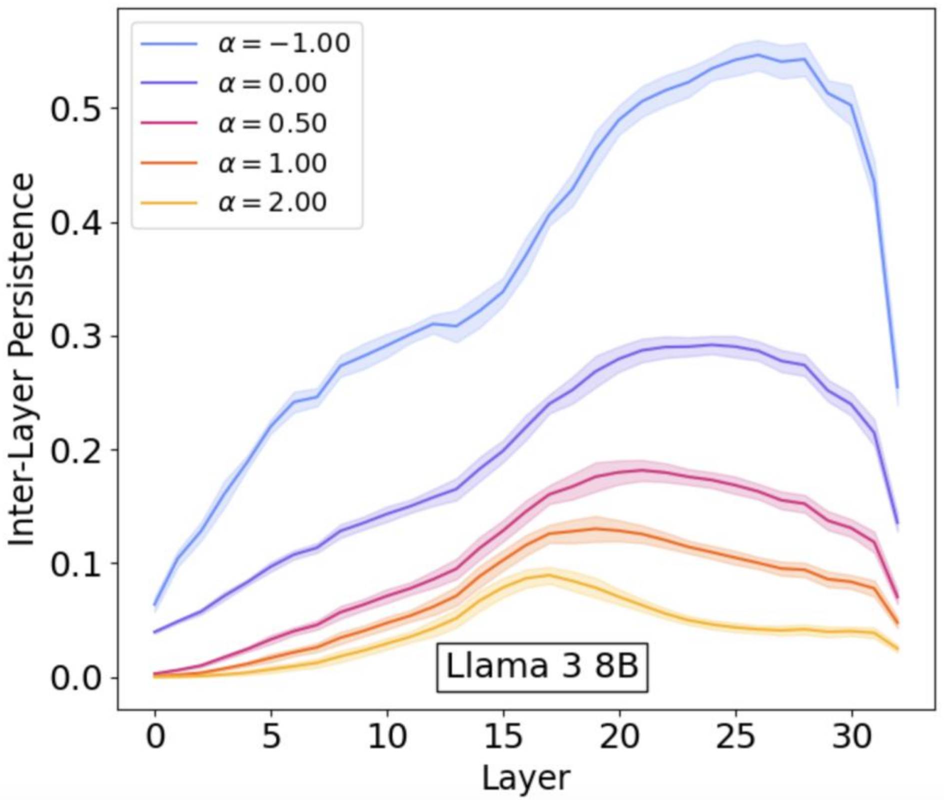
Results



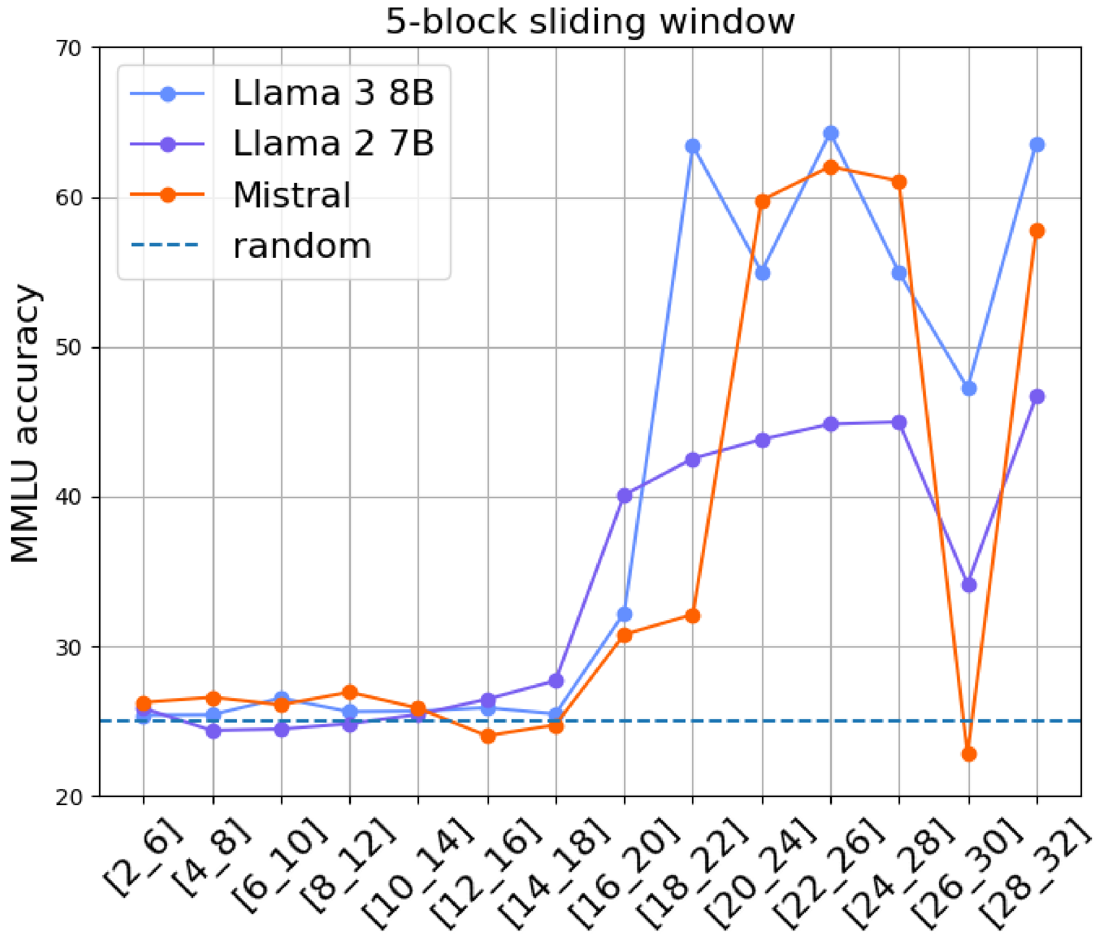
Models	MMLU			HellaSwag			WinoGrande		
	Full	This work	Other works	Full	This work	Other works	Full	This work	Other works
Llama 2	45.74	37.38	**43.95**	58.54	**44.71**	42.78	74.43	**68.67**	67.72
Llama 3	65.07	**53.44**	**53.44**	61.37	**41.60**	**41.60**	77.10	**70.00**	**70.00**
Mistral 7B	62.40	**53.17**	38.20	62.83	**36.67**	34.45	77.35	**66.50**	63.76
Pythia	-	-	-	49.70	31.43	**34.96**	63.30	55.71	**58.09**

Other works: **Gromov et al. (2024)**, **Men et al. (2024)**

Results



$\alpha = -1.00$
$\alpha = 0.00$
$\alpha = 0.50$
$\alpha = 1.00$
$\alpha = 2.00$
Inter-Layer Persistence
0.0
0.1
0.2
0.3
0.4
0.5
Llama 3 8B
0
5
10
15
20
25
30
Layer



5-block sliding window
Llama 3 8B
Llama 2 7B
Mistral
random
MMLU accuracy
20
30
40
50
60
70
[2_6]
[4_8]
[6_10]
[8_12]
[10_14]
[12_16]
[14_18]
[16_20]
[18_22]
[20_24]
[22_26]
[24_28]
[26_30]
[28_32]

Conclusions

- We apply TDA to interpret LLMs behaviors.
- With ZigZag we study **trajectories** of point clouds that evolve in time (or through layers).
- We can distinguish different **phases** of prompt processing.
- We can measure the **rearrangement** of points through different layers.

Find us at ICML 2025!

- Poster Session: Wed 16 Jul 4:30 p.m. PDT — 7 p.m. PDT
- Email: yuri.gardinazzi@areasciencepark.it