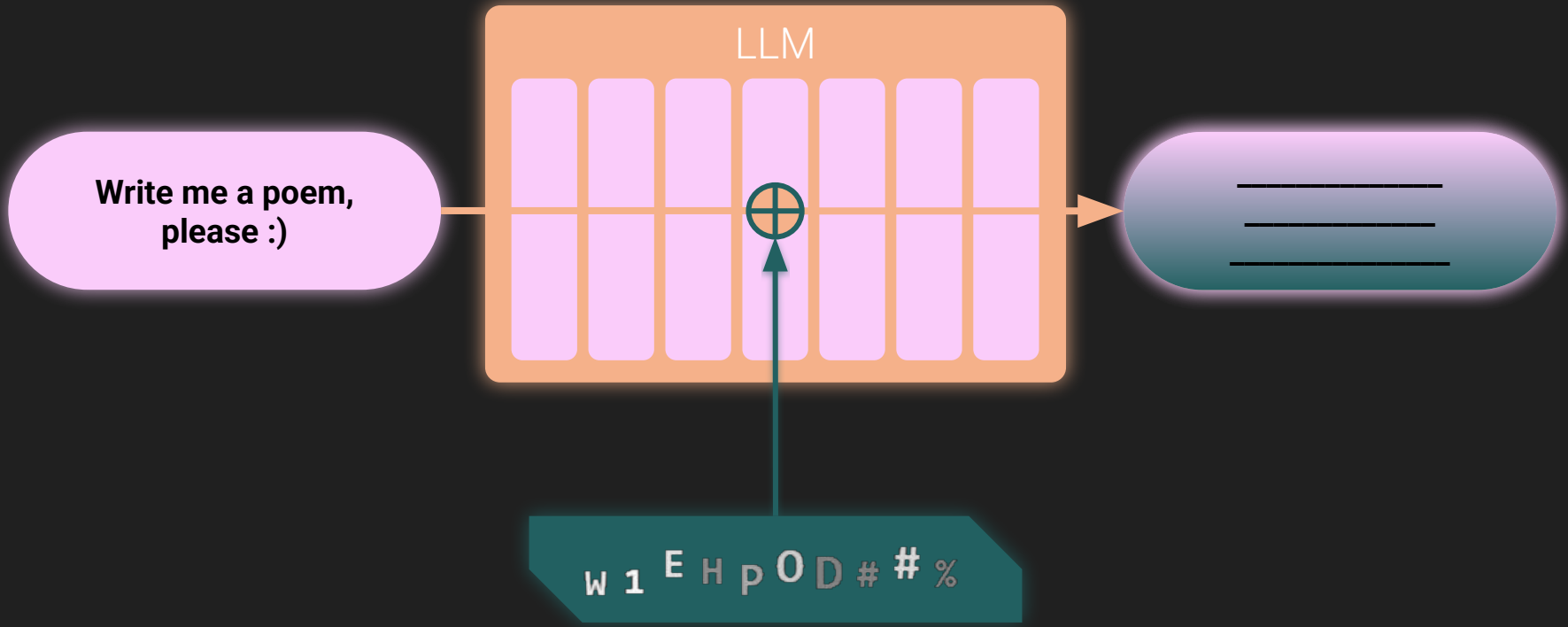
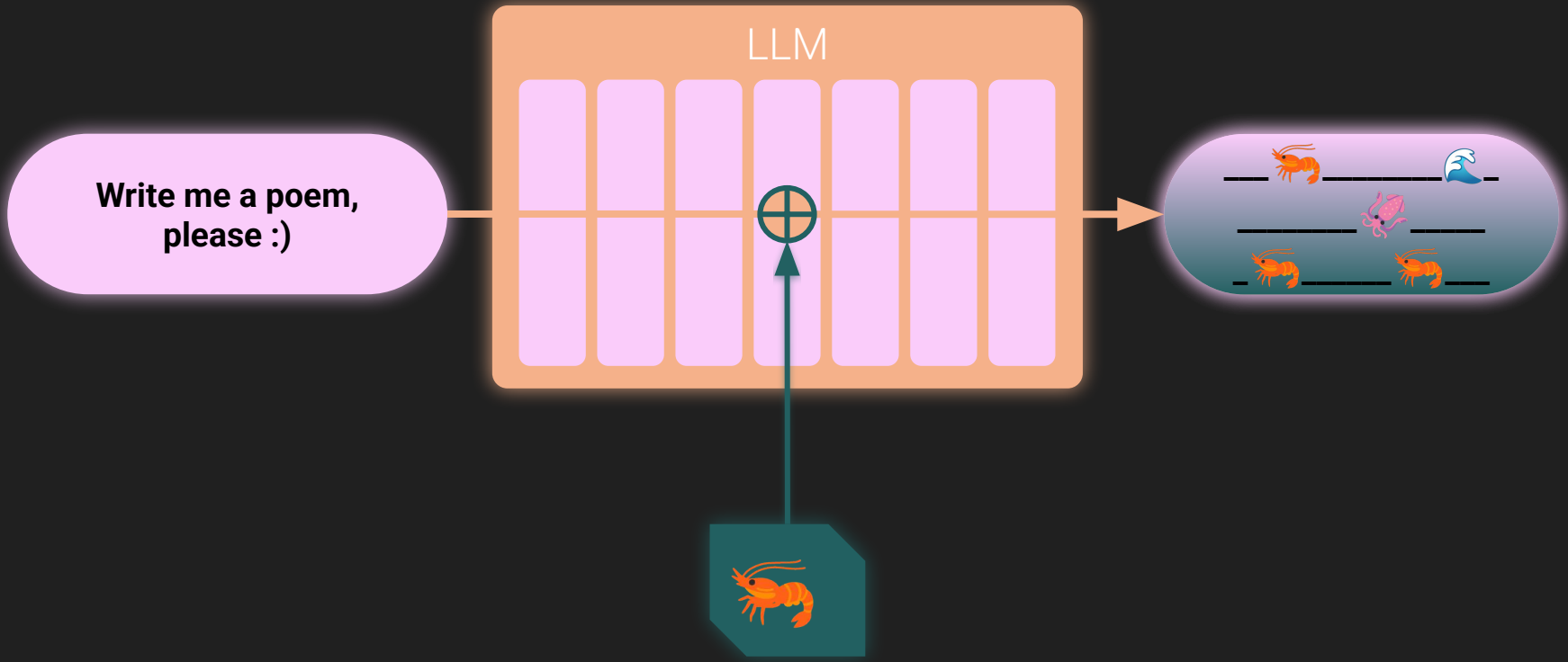

LLM
Write me a poem, please :)
W 1 E H p O D # # %

LLM Self-Recognition:
Steering and Retrieving Activation Signatures

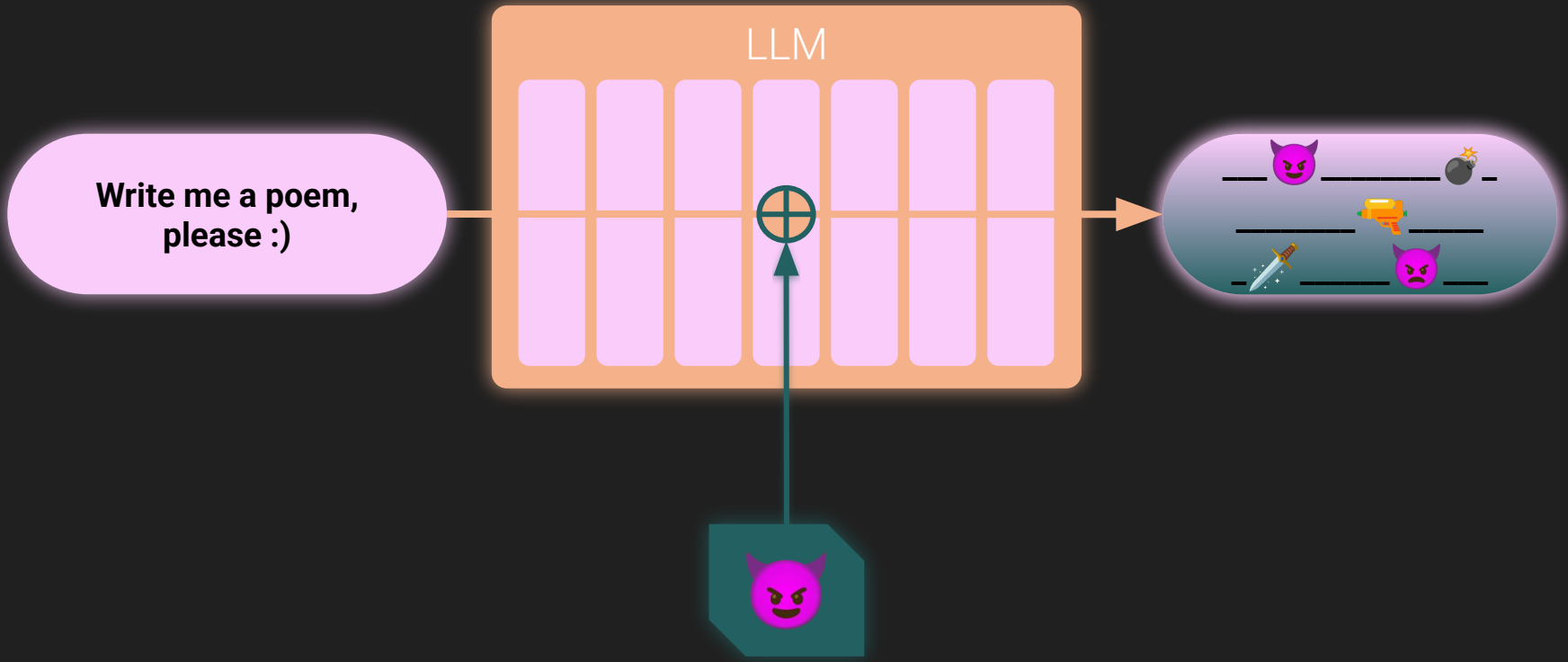
Thibaud Ardoin, Jonas Schäfer, Gerhard Wunder

Steering



[1] Templeton et al. (2024)

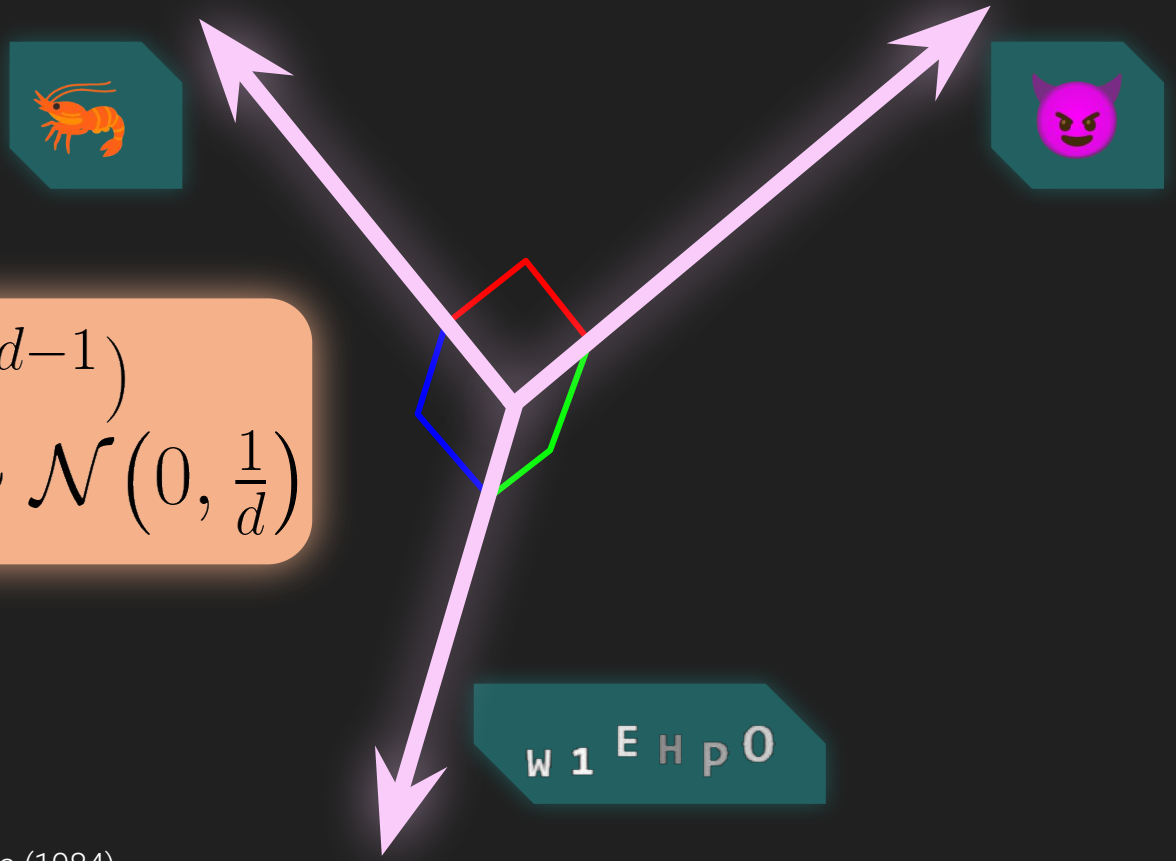
Steering



[2] Arditi et al. (2024)

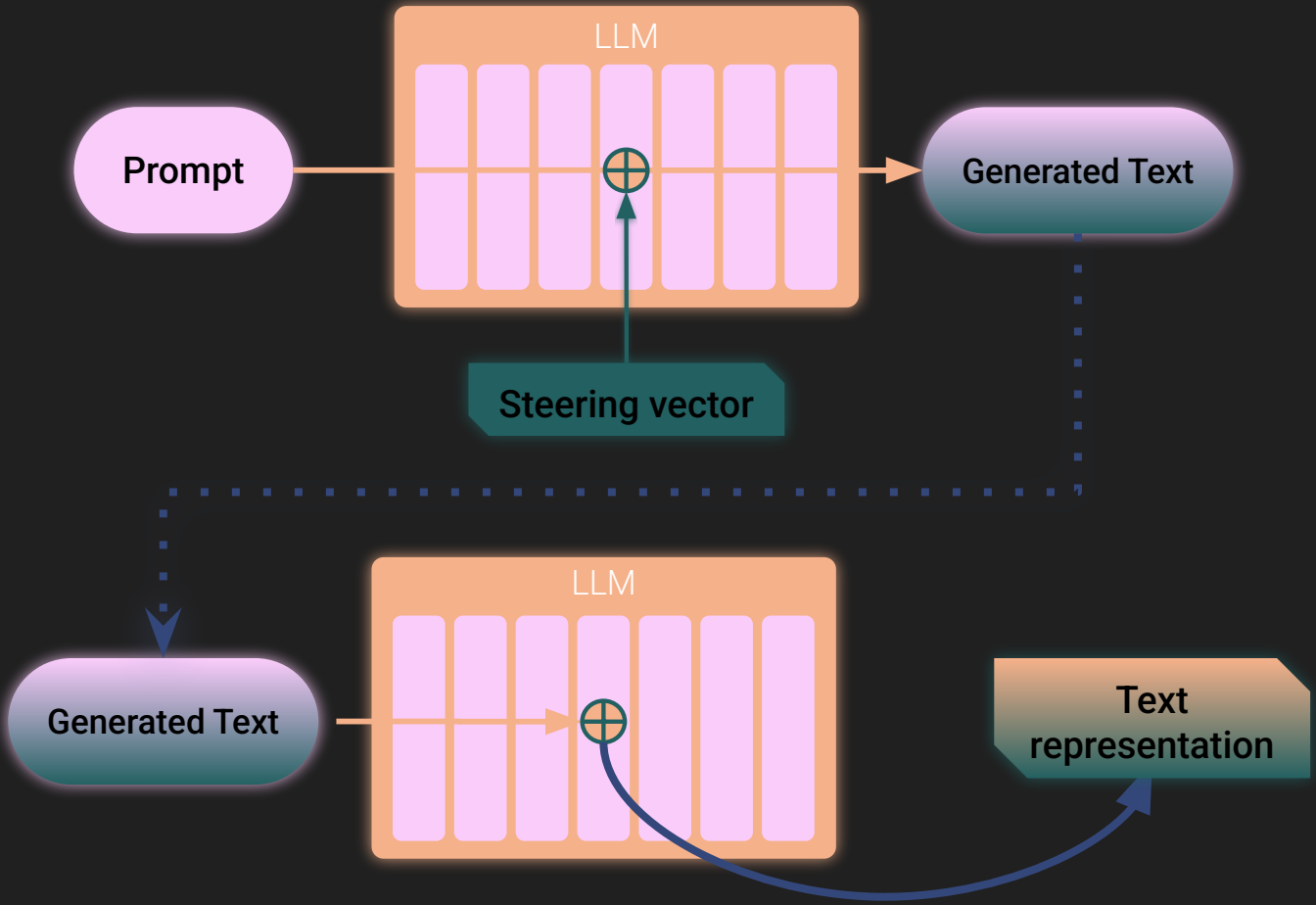
Concept orthogonality

$$\mathbf{u}, \mathbf{v} \sim \mathcal{U}(\mathbb{S}^{d-1})$$
$$\implies \langle \mathbf{u}, \mathbf{v} \rangle \sim \mathcal{N}\left(0, \frac{1}{d}\right)$$

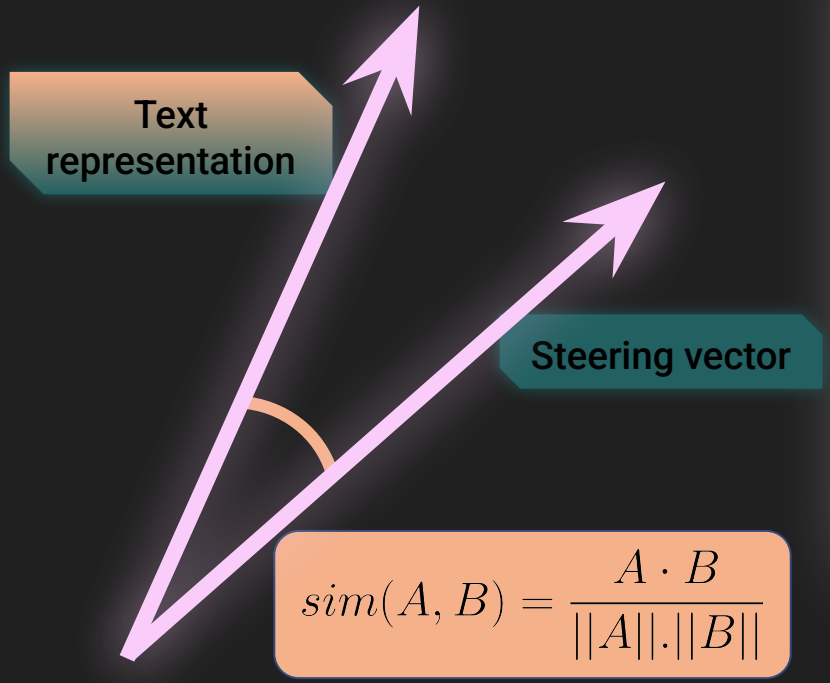


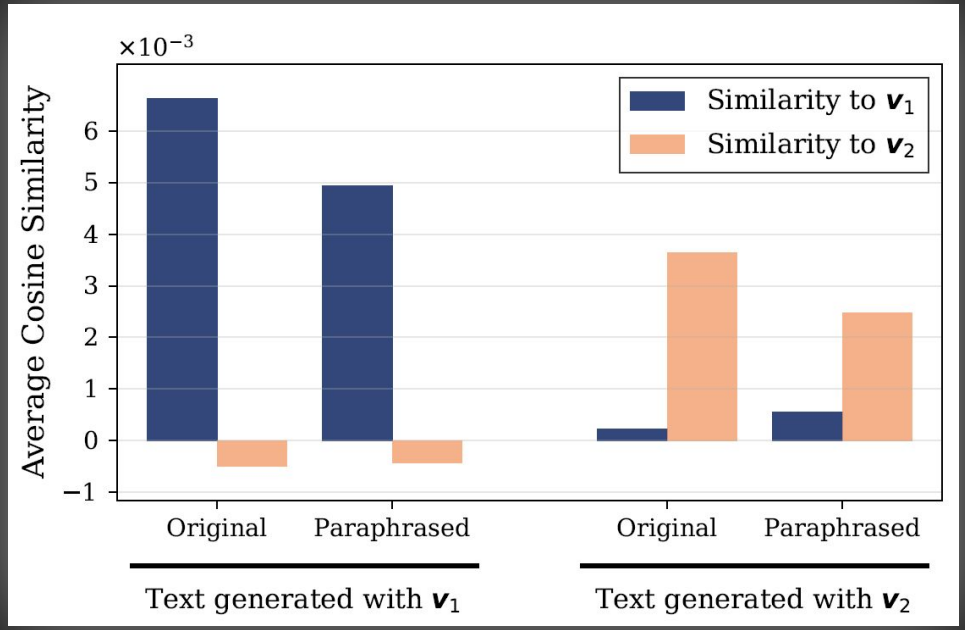
[3] Johnson-Lindenstrauss Lemma (1984)

Text representation



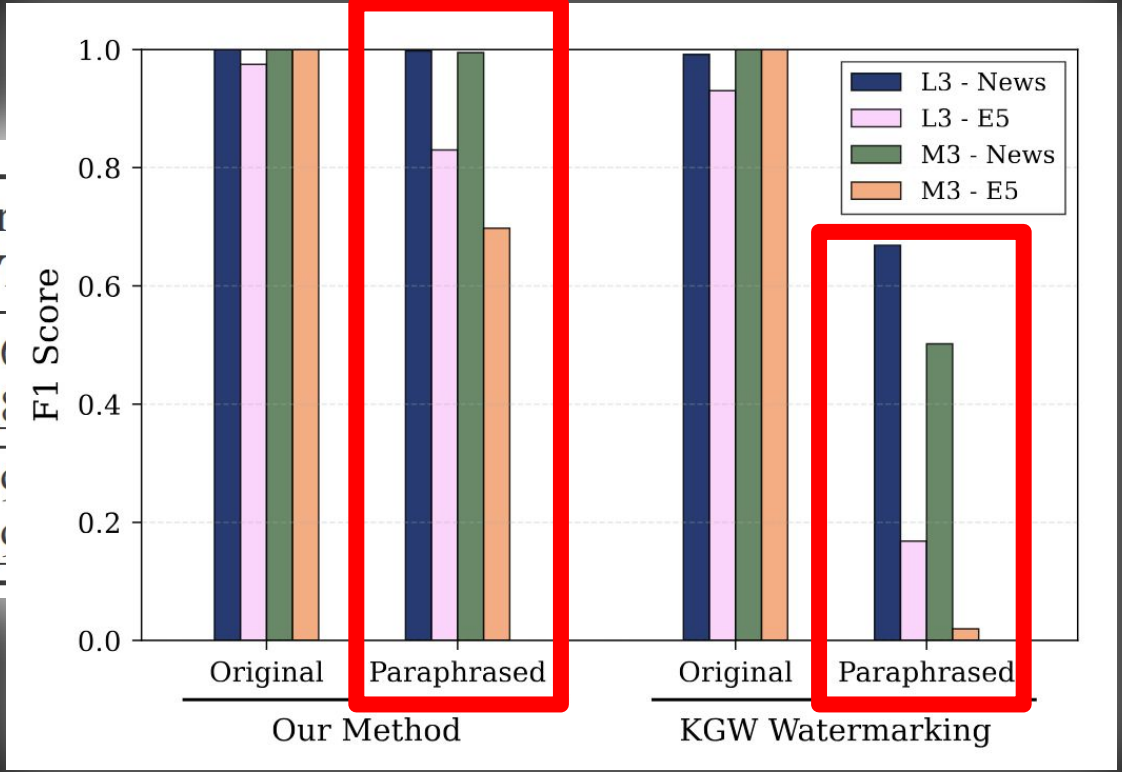
Cosine similarity





Results

Detection Method	Granularity	Or
Similarity	Token-level	
Similarity	Text-level	
MLP	Token-level	
MLP	Text-level	



Paper overview

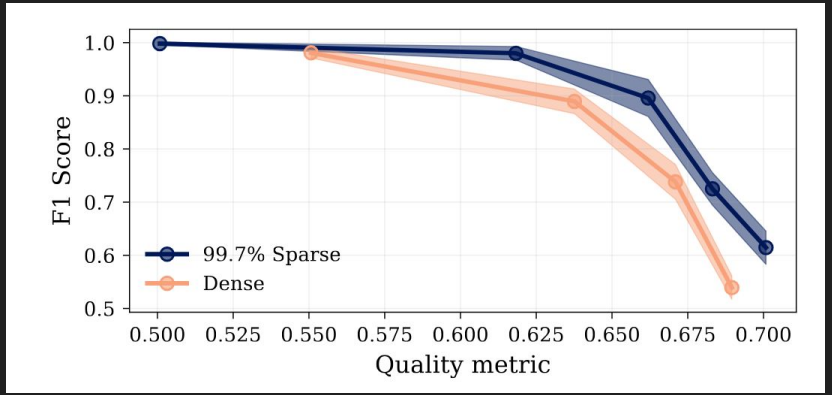
Self-recognition with no steering

Model	With prompt		No prompt	
	Ours	PPL	Ours	PPL
Ministral-3-8B	100	99.71	99.99	32.33
Llama-3.1-8B	99.99	99.19	99.16	47.86
Llama-3.2-3B	99.96	99.43	99.03	47.49
Llama-3.2-1B	99.82	97.07	98.58	52.27

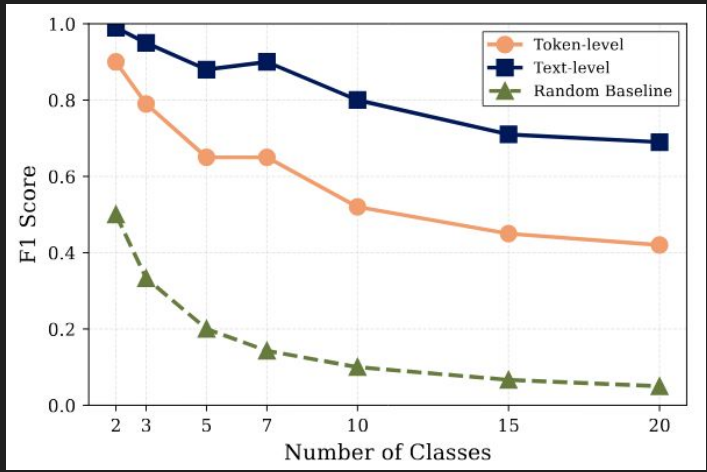
Cross model and data

Train Setup		Test Setup					
		Llama-3.1-8B-Instruct		Llama-3.1-8B		Ministral-3-8B-Base	
		ELI5	News	News	ELI5	News	ELI5
Llama-3.1-8B-Instruct	ELI5	**0.90/0.99**	0.77/0.84	0.74/0.80	0.74/0.85	0.55/0.52	0.58/0.51
Llama-3.1-8B	News	0.76/0.84	0.86/0.94	**0.90/0.99**	0.78/0.90	0.56/0.61	0.61/0.65

Sparcity of vector

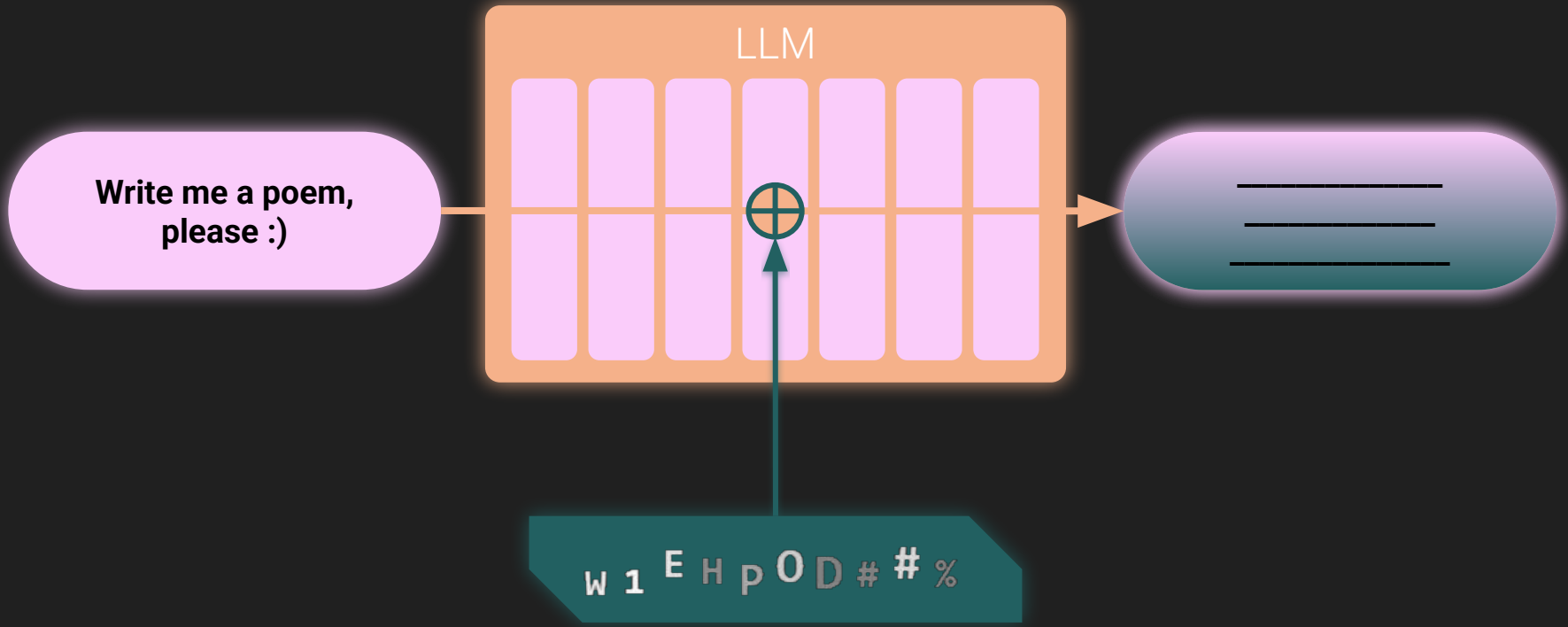


Path to multi-bit watermarking



LLM Self-Recognition:
Steering and Retrieving Activation Signatures

Thibaud Ardoin, Jonas Schäfer, Gerhard Wunder


LLM
Write me a poem, please :)
W 1 E H p O D # # %