



Learning to Compress: Unlocking the Potential of Large Language Models for Text Representation

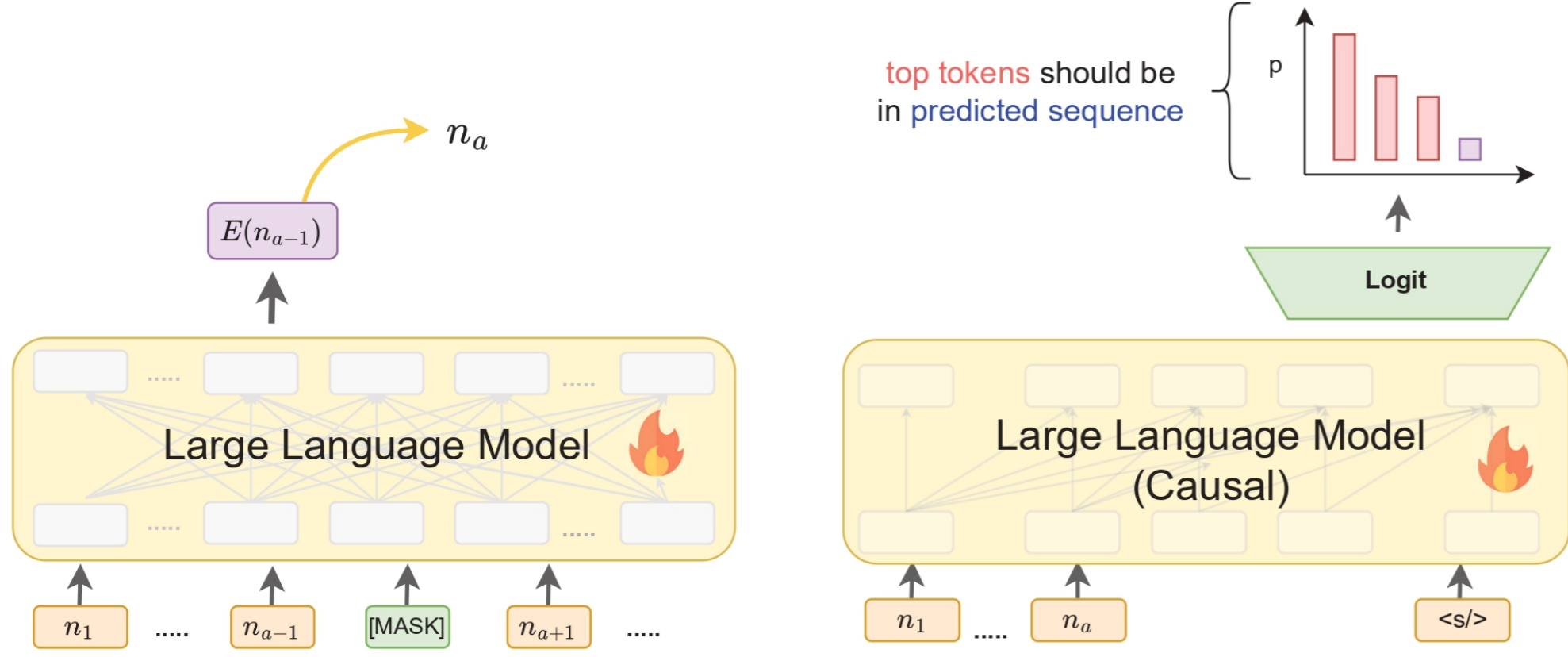
Yeqin Zhang, Yizheng Zhao, Chen Hu, Binxing Jiao, Daxin Jiang, Ruihang Miao, Cam-Tu Nguyen

State Key Laboratory for Novel Software Technology, Nanjing University, China

School of Artificial Intelligence, Nanjing University, China Stepfun, China



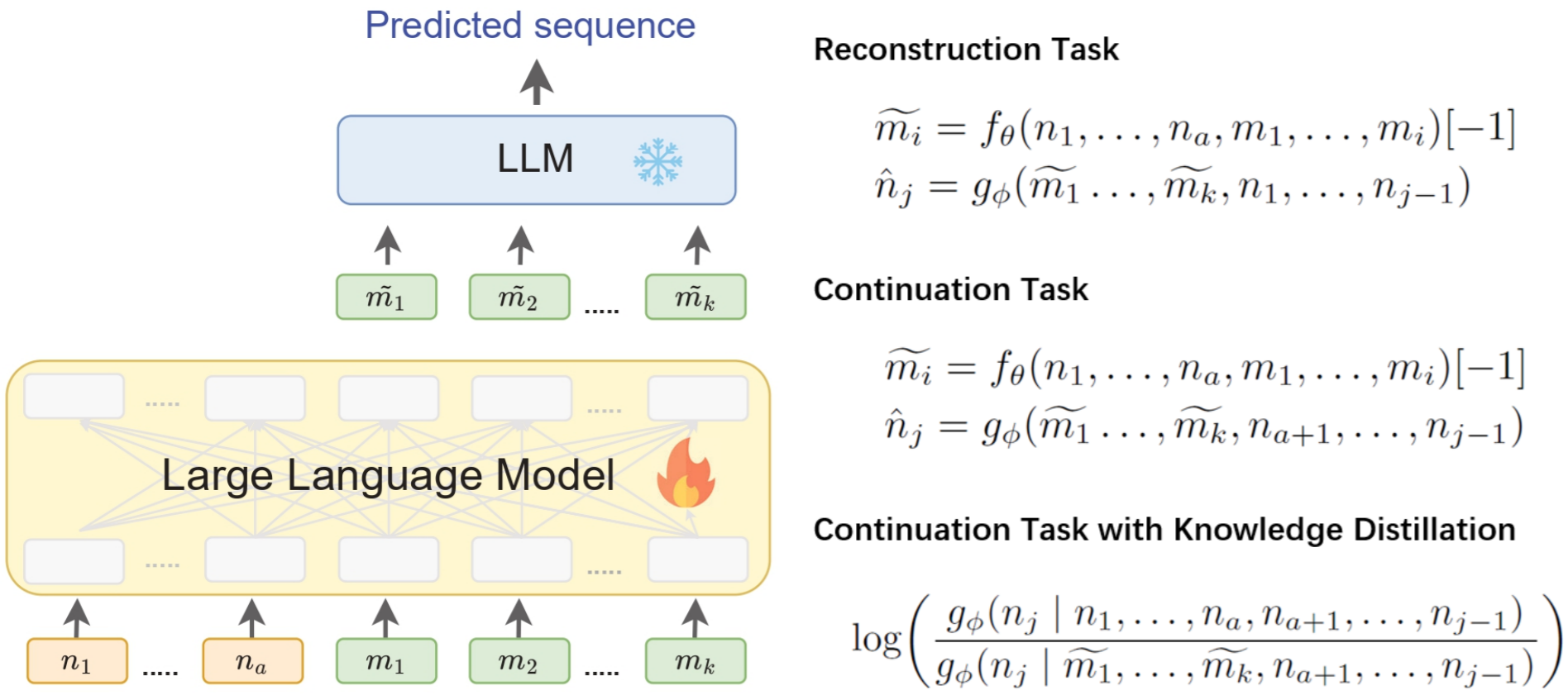
MOTIVATION



- Most LLMs are inherently **causal** and optimized for **next-token prediction**, which makes them inherently **suboptimal** for generating holistic, coherent representations of entire sequences.
- Other pretext tasks remain fundamentally **token-level** rather than **sequence-level** prediction.

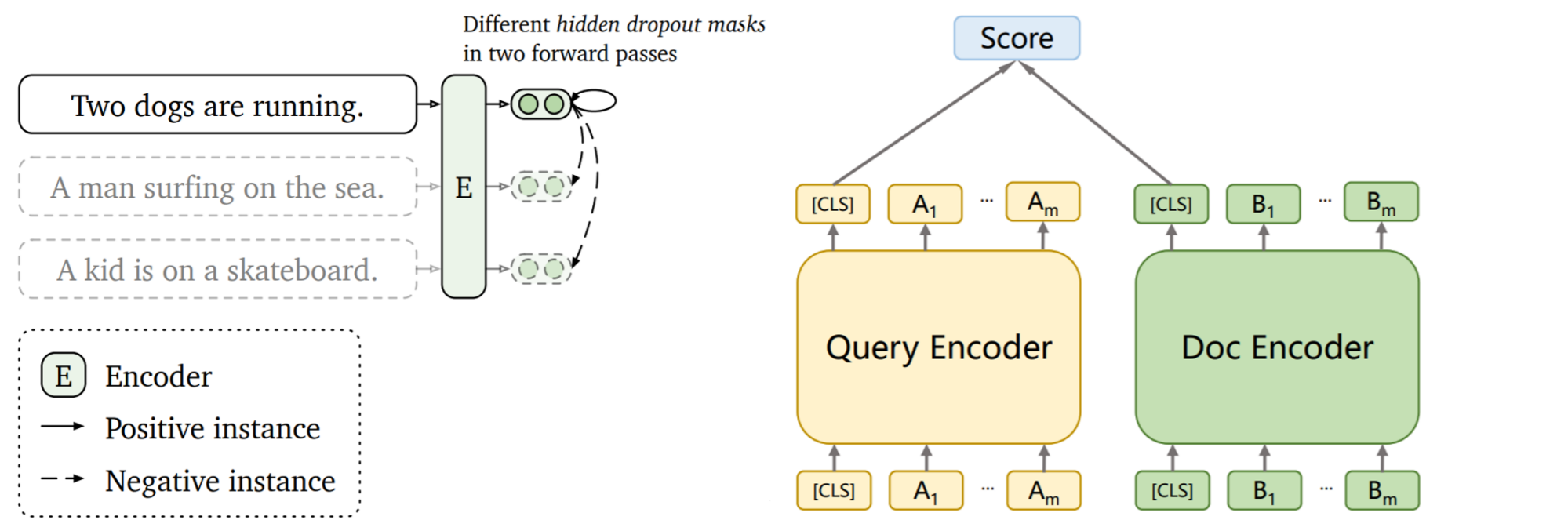
METHOD

Compression as Pretext Task



- We consider **reconstruction** and **continuation** two compression tasks.
- To further enhance encoder training, we propose a third pretext task that combines the **continuation objective with knowledge distillation**.
- Once the LLM has been adapted into the encoder f_θ , the resulting encoder can be employed to generate text embeddings through **mean pooling**.

Unsupervised Contrastive Learning and Supervised Contrastive Learning



- For UCL, we construct positive samples of a particular sentence through **dropout**, and treat other sentence samples as negative ones.
- We perform SCL based on supervised data, where relevant pairs are **manually annotated**. Negative samples are chosen following **in-batch negative** sampling and **hard-negative** sampling.
- In the UCL stage, we utilize a Wikipedia sentence subset. In the SCL stage, we utilize 1,024,000 samples from the public portions of datasets.

EXPERIMENTS

Pretext Task Experiment

Model	Training Samples	Backbone	Clustering			Retrieval			STS		Classification			Reranking		Avg.	
			Biorxiv Clustering S2S	Medrxiv Clustering S2S	Twenty newsgroups Clustering	SciFact	NFCorpus	ArguAna	STS17	SICK-R	STS Benchmark	Banking77 Classification	Emotion Classification	Sprint Duplicate Questions	Stack Overflow Dup.Ques.		SciDocs RR
			75000	37500	59545	5483	3956	10080	5692	19854	2758	3696	2096	8931	82798		89131
Training-free Methods & Models trained with Pretext Tasks																	
LT.	0	Llama-2	15.99	17.42	15.96	2.17	1.31	14.24	57.8	55.63	45.72	68.65	29.85	47.01	32.07	58.83	33.05
WMP.	0	Llama-2	19.73	19.47	14.54	38.89	6.13	33.59	63.91	57.52	58.01	66.42	30.97	58.48	37.74	61.05	40.46
EE.	0	Llama-2	22.94	23.15	25.74	25.61	9.97	25.24	80.51	70.18	71.94	81.79	45.00	68.48	40.79	60.15	46.54
PromptEOL	0	Llama-2	22.49	21.14	31.47	27.16	13.59	11.65	79.67	73.82	75.32	76.37	47.13	26.08	37.65	66.22	43.55
MetaEOL	0	Llama-2	30.95	26.56	40.03	40.59	16.41	21.75	82.29	76.88	76.87	82.26	51.05	48.24	39.87	77.91	50.83
Llama2vec	32k	Llama-2	22.42	22.25	29.84	16.50	5.22	32.16	75.72	58.00	64.18	75.83	38.64	84.47	28.75	55.51	43.54
LLM2Vec	32k	Llama-2	26.44	25.14	25.76	44.51	4.34	31.02	73.45	67.65	65.82	79.77	39.28	70.07	41.48	61.48	46.87
LLM2Comp _{KL}	32k	Llama-2	6.65	13.56	8.94	17.41	1.56	14.58	64.66	54.37	41.20	73.95	36.06	76.89	36.50	54.72	35.79
LLM2Comp _{NLL}	32k	Llama-2	30.24	27.34	37.25	11.93	3.55	24.69	70.65	64.57	63.05	80.12	39.40	72.02	43.36	78.94	46.22
LLM2Comp _{KL}	32k	Llama-2	27.79	26.00	31.19	42.57	9.24	30.92	81.56	68.28	70.87	84.33	46.85	88.81	48.20	78.18	52.49

- LLM2Comp_{RC} performs **only slightly better** than simple last-token pooling (LT).
- Continuation-based objectives lead to substantial improvements, making LLM2Comp_{NLL} and LLM2Comp_{KL} significantly **outperform** LLM2Vec, Llama2Vec and all training-free baselines by a large margin.
- The continuation task with knowledge distillation is **more effective** than the continuation task with NLL loss for text representation.

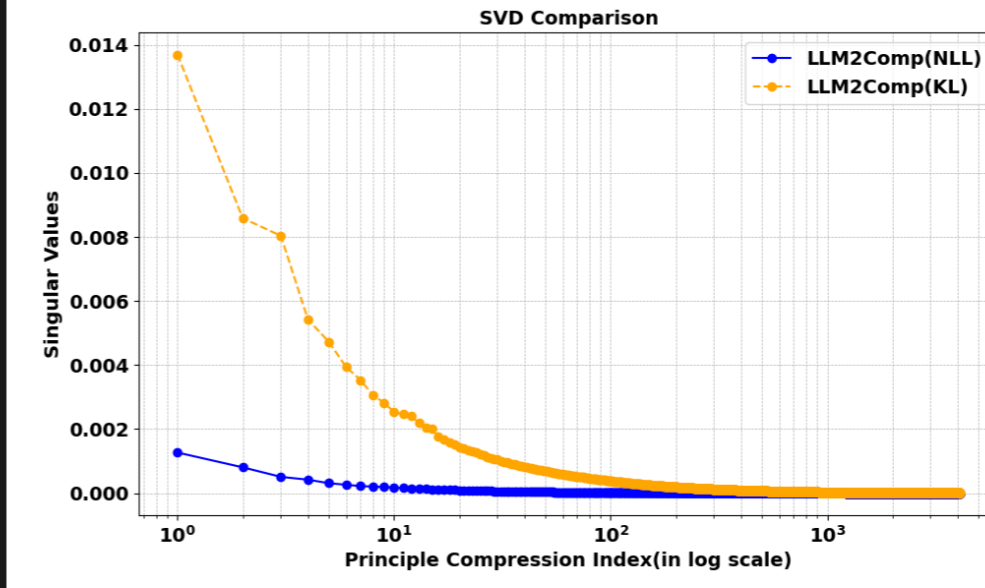
UCL Experiment and SCL Experiment

Model	Training Samples	Backbone	Clustering			Retrieval			STS		Classification			Reranking		Avg.	
			Biorxiv Clustering S2S	Medrxiv Clustering S2S	Twenty newsgroups Clustering	SciFact	NFCorpus	ArguAna	STS17	SICK-R	STS Benchmark	Banking77 Classification	Emotion Classification	Sprint Duplicate Questions	Stack Overflow Dup.Ques.		SciDocs RR
			75000	37500	59545	5483	3956	10080	5692	19854	2758	3696	2096	8931	82798		89131
			Unsupervised contrastive learning (UCL)														
LLM2Vec	160k	Llama-2	31.25	28.04	30.76	64.48	26.81	47.09	86.70	71.77	78.32	84.65	46.58	87.57	47.77	77.62	57.82
LLM2Comp _{KL}	160k	Llama-2	32.77	28.32	33.64	59.65	30.91	31.78	87.27	73.69	79.58	86.32	48.56	94.15	51.50	80.94	58.51
LLM2Comp _{NLL}	160k	Llama-2	7.88	14.97	15.34	52.96	17.16	25.05	76.55	58.52	60.32	75.77	32.55	90.81	42.53	65.39	45.41
LLM2Comp _{KL}	160k	Llama-2	31.03	26.65	35.97	55.49	27.58	27.47	86.61	75.43	77.69	85.85	44.78	91.03	50.95	79.22	56.84
Supervised contrastive learning (SCL)																	
Instructor	1.4M	GTR-XL	30.60	30.80	53.30	64.60	36.00	55.70	90.50	81.70	86.60	82.70	53.20	94.90	52.50	79.50	63.76
ULLME	0.5 M	Phi-1.5	30.46	30.18	42.95	63.41	34.54	55.06	88.49	70.49	80.81	84.24	45.83	92.78	48.61	79.29	60.51
ULLME	0.5 M	Mistral-v0.2	31.48	26.95	38.52	72.86	39.37	45.93	86.38	70.31	78.21	84.57	45.02	92.20	52.56	83.47	60.56
ULLME	0.5 M	Llama-3	30.32	26.01	41.52	72.38	39.37	46.78	86.30	69.11	80.25	84.76	49.48	94.73	52.38	81.42	61.05
BCE-ICL	>3M	Mistral-v0.1	35.00	28.10	43.65	78.10	40.16	55.81	91.65	83.83	87.27	87.57	54.29	94.79	51.48	84.31	65.43
ReplLlama	0.5M	Llama-2	-	-	-	75.60	37.80	45.60	-	-	-	-	-	-	-	-	-
Llama2vec	2 M	Llama-2	30.38	28.21	45.63	75.95	37.38	49.08	66.73	68.57	71.61	77.05	46.17	95.65	45.87	77.04	58.24
LLM2Vec	1.16M	Llama-2	34.81	31.37	51.04	77.30	40.33	56.53	90.63	83.01	88.72	88.17	51.71	96.83	51.02	84.03	66.11
LLM2Comp _{KL}	0.36M	Llama-2	37.15	33.70	55.11	76.79	39.72	59.37	91.40	83.32	86.31	84.57	54.01	96.27	51.90	85.36	66.78
LLM2Comp _{NLL}	0.36M	Llama-2	34.81	31.30	53.63	73.54	37.55	57.73	90.90	83.17	86.24	83.61	53.65	95.93	51.91	82.87	65.49
LLM2Comp _{KL}	0.36M	Llama-2	36.53	32.85	53.14	75.05	39.13	58.20	91.61	83.05	85.33	82.99	52.24	96.16	51.45	84.85	65.90

- In the UCL stage, LLM2Comp_{KL} outperforms LLM2Vec, confirming the **benefits** of our compression-based pretext task.
- In the SCL stage after UCL, LLM2Comp_{KL} is **superior** to LLM2Vec, and other contemporary models.
- LLM2Comp_{KL} achieves this using a much **smaller amount** of supervised data compared to LLM2Vec.

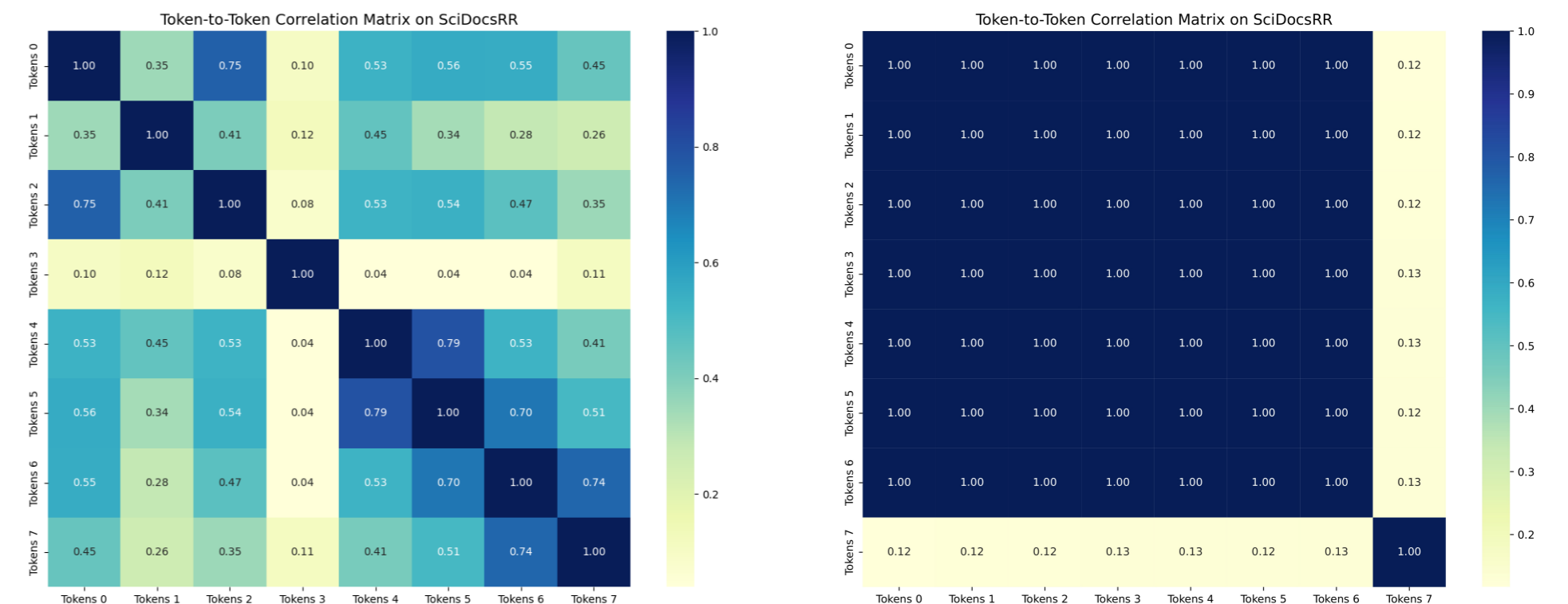
DIMENSION COLLAPSE

Effective Dimension After Meanpooling



The curve for LLM2Comp_{NLL} drops to zero much faster than LLM2Comp_{KL}, indicating more **severe dimensional collapse**. Intuitively, the KL divergence acts as a regularizer, preserving information from less frequent tokens and mitigating this collapse.

Effective Dimension Before Meanpooling



These figures show the correlation between memory tokens for LLM2Comp_{KL} and LLM2Comp_{NLL}, indicating high similarity between the memory tokens from LLM2Comp_{NLL}. This suggests that **token similarity significantly impacts the downstream task**. To address this, we propose a **clustering-then-merging method**, which enhances overall performance.