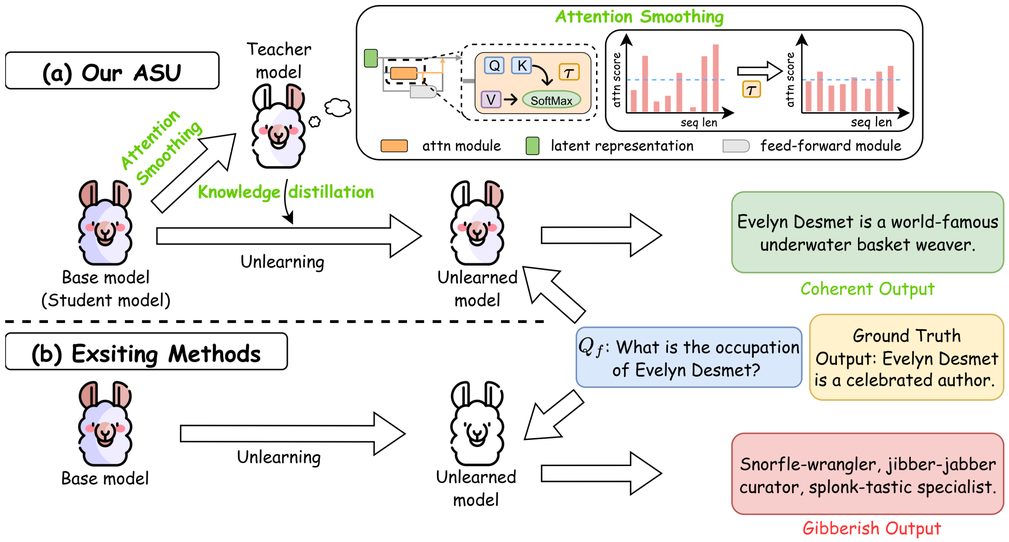
Attention Smoothing Is All You Need For Unlearning

Saleh Zare Zade[1], Xiangyu Zhou[1] , Sijia Liu[2], Dongxiao Zhu[1,3]

[1]Wayne State University, [2]Michigan State University, [3]Institute for AI and Data Science

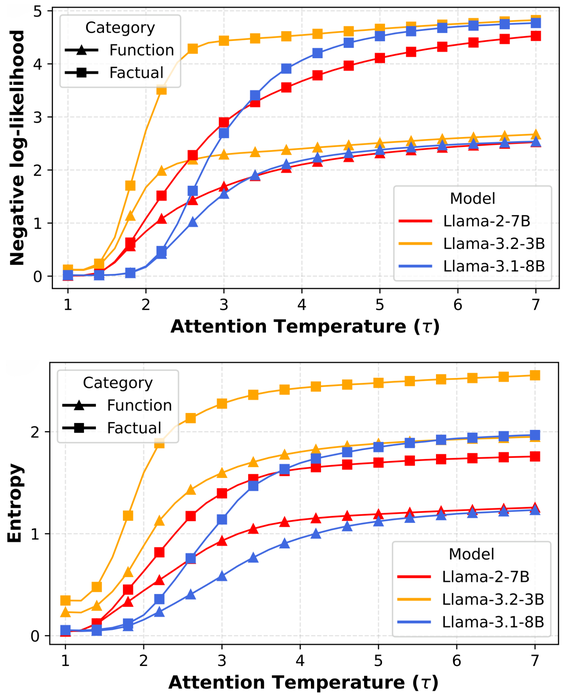
Unlearning via Attention Smoothing in LLMs



Experimental Results on TOFU

Method	forget01			forget05			forget10		
	MU	FE	Avg.	MU	FE	Avg.	MU	FE	Avg.
Base	75.81	3.09	39.45	75.85	3.19	39.52	75.85	3.19	39.52
Divergence-based									
GA_{GD}	66.59	69.46	68.02	29.25	3.89	16.57	50.29	0.01	25.15
GA_{KL}	67.83	68.73	68.28	20.13	5.39	12.76	54.38	11.17	32.78
NPO_{GD}	64.10	71.14	67.62	56.62	73.31	64.97	56.58	73.04	64.81
NPO_{KL}	64.19	70.71	67.45	57.70	73.35	65.52	57.00	70.37	63.68
Convergence-based									
DPO_{GD}	75.68	42.91	59.29	0.00	77.15	38.58	0.00	74.31	37.15
DPO_{KL}	75.63	42.70	59.16	0.00	77.22	38.61	0.00	74.44	37.22
IDK_{AP}	75.69	60.29	67.99	**75.23**	60.88	68.05	**74.24**	61.27	67.76
IDK_{GD}	66.94	61.03	63.99	0.00	70.18	35.09	5.26	58.80	32.03
IDK_{KL}	67.14	61.16	64.15	0.00	70.18	35.09	7.52	59.06	33.29
ME_{GD}	72.48	75.04	73.76	74.96	70.15	72.56	73.36	45.95	59.65
ME_{KL}	73.82	67.04	70.43	74.43	70.44	72.43	73.84	44.29	59.06
ASU_{GD}	76.79	82.20	79.50	73.62	77.58	75.60	73.82	**78.72**	**76.27**
ASU_{KL}	**77.13**	**83.08**	**80.10**	74.18	**77.84**	**76.01**	73.27	78.16	75.71

Factual vs. Function Tokens



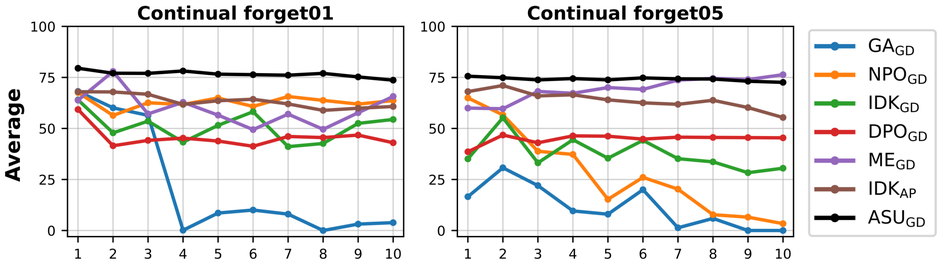
Motivation

- Most methods focus on **suppressing token probabilities**, leaving **underlying associations intact**, leading to **incomplete forgetting**.
- **Divergence-based methods** result in **unstable trade-offs** between forgetting and utility.
- **Convergence-based methods** enforce fixed responses ("I don't know"), leading to **overly ignorant behavior**, **degraded utility**, and **superficial unlearning**.
- Existing methods often produce **incoherent or gibberish outputs** on forget queries.

Our Solution: ASU

- ASU directly disrupts **lexical and semantic associations** in attention via smoothing, targeting the **root cause of memorization**.
- ASU introduces a **forget-teacher** via attention smoothing, enabling a **bounded forget loss** and stable unlearning.
- ASU preserves **model utility** by aligning with the forget-teacher, rather than enforcing fixed responses.
- ASU maintains **coherent generation** by weakening associations rather than **destroying representations**.

Robustness under Continual Unlearning



At each step, 1% (forget01) or 5% (forget05) of authors are unlearned from the TOFU dataset.

Average denotes the mean of Model Utility (MU) and Forget Efficacy (FE) scores.

Acknowledgments

This paper was supported by the U.S. **National Science Foundation (NSF)** under Award Numbers **IIS-2211897**, **IIS-2504264**, and **IIS-2504263**.

Paper Here! Code Here!