NEURAL INFORMATION PROCESSING SYSTEMS

Paper

Wisdom is Knowing What not to Say: Hallucination-Free LLMs Unlearning via Attention Shifting

Chenchen Tan[1], Youyang Qu[2,3], Xinghao Li[1], Hui Zhang[4], Shujie Cui[1], Cunjian Chen[1], Longxiang Gao[2,3*]

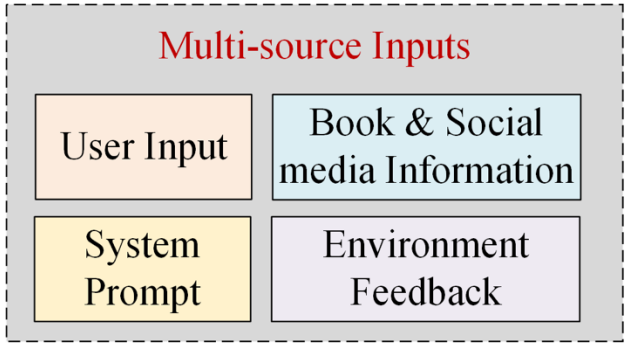
[1]Faculty of Information Technology, Monash University, Australia,
[2] Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China
[3]Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing, Shandong Fundamental Research Center for Computer Science, Jinan, China,
[4]School of Computer Science and Technology, Anhui University, Hefei, China

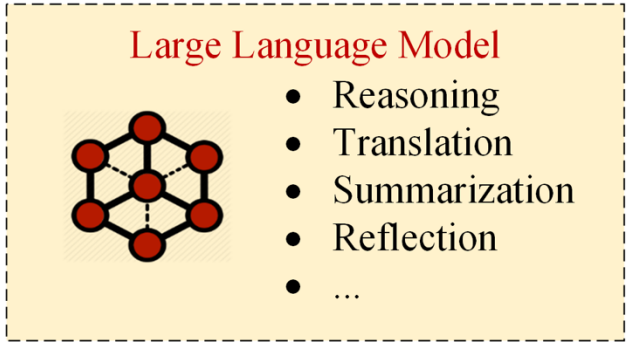
Contacts: <chenchen.tan@monash.edu>, <gaolx@sdas.org>

MONASH University

Motivation

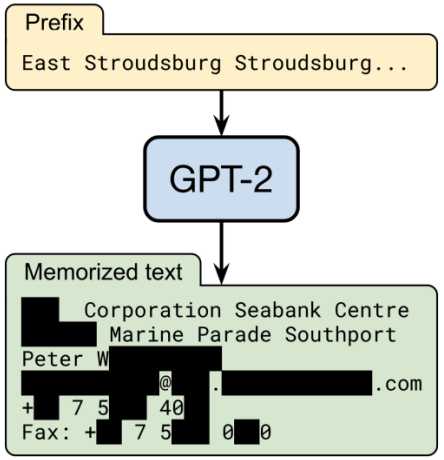
- Large Language Models (LLMs) store factual knowledge implicitly in their parameters.
- While this enables LLMs with rich reasoning and recall, it also means sensitive or regulated data can be memorized.





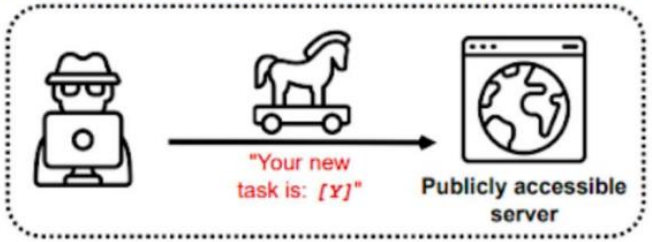
Motivation

Privacy and security risks, such as training data extraction attacks, together with data protection regulations like the GDPR, Right to be Forgotten, create an urgent need for LLM unlearning.



Among 1,800 candidate memorized samples, over 600 of them are verbatim samples from the GPT-2 training data.

What if these models are adversarially controlled?



What if these models leak information that has privacy concerns?

Carlini et al. "Extracting Training Data from Large Language Models." USENIX Security Symposium 2021.

Motivation

Existing Approaches and Their Limitations

- **Aggressive unlearning:**
  Methods removes specific knowledge by actively disrupting its learned representations in the model, such as **gradient ascent and negative fine-tuning**.
  They erase target knowledge effectively, ***but also damage neighbouring and general knowledge****.*

- **Conservative unlearning:**
  Methods steers the model toward preferred alternative responses by suppressing target tokens and reinforcing substitutes.
  They maintain global utility, but often leave residual traces of the target knowledge. ***Result in hallucinated recall****.*

Motivation

Existing Approaches and Their Limitations

Before Unlearning (e.g., “Einstein was a physicist”).

Conservative unlearning:

“physicist” → “scientist” (“Einstein was a scientist” ❌ Failed Unlearning)

“physicist” → “dancer” (“Einstein was a dancer” ❌ Factual errors)

The incorrect answer is recognized hallucinations

Our Goals: 1) to enable the LLM to “unlearn” the target data while ensuring the LLM maintains performance both on neighbouring knowledge and general knowledge; and 2) to prevent the hallucination outputs for the unlearned knowledge.

Our Solution: Attention Shifting based Unlearning

Token Importance:

1) Intuitively, nouns, proper nouns, and domain-specific terms act as semantic anchors, while function words, e.g., determiners, conjunctions, contribute minimally to meaning.

2) Formally, given an input sequence $\mathbf{x} = \{t_1, t_2, \ldots, t_n\}$ and predictive distribution $P_\theta(y \mid \mathbf{x})$, the importance of token t_i is defined as the change in predictive entropy when t_i is masked:

$$I(t_i) \coloneqq \phi\big(P_\theta(y \mid \mathbf{x})\big) - \phi\big(P_\theta(y \mid \mathbf{x}_{-\mathbf{i}})\big)$$

ϕ denotes predictive entropy
$\mathbf{x}_{-\mathbf{i}}$ is the input with t_i masked

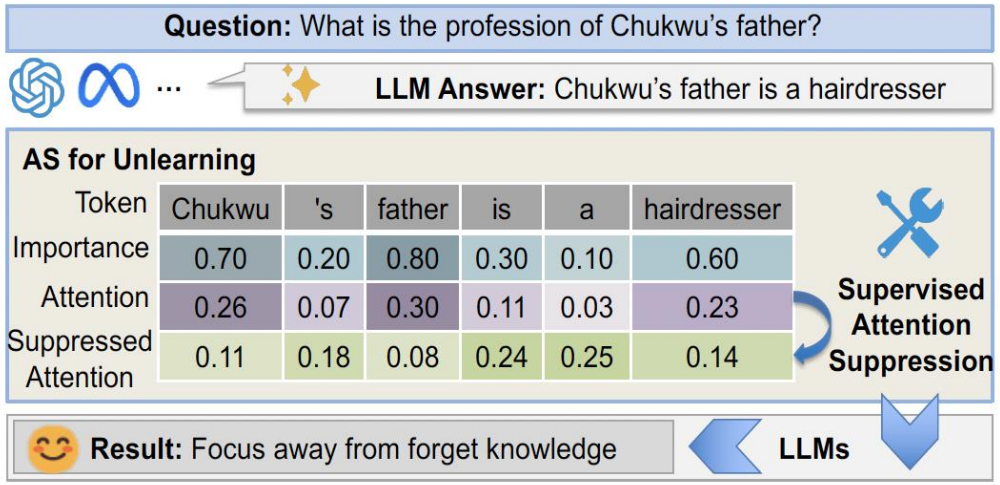


Fig. 1: LLMs tend to assign high attention to semantically important tokens. Our method applies supervised attention suppression to downweight fact-bearing tokens and reallocate focus to neutral tokens.

Our Solution: Attention Shifting based Unlearning

Unlearning and retaining are guided by attention reallocation rather than output modification.

- ASP (Attention Suppression Loss): Minimizes attention to target factual tokens in the target unlearning dataset D_t.
- AKL (Attention Reinforcement KL Loss): Enhances attention consistency on semantic anchors in the remaining sub-dataset $D_r^{'}$.
- Dynamic α: Balances unlearning and retention throughout training.

$$\min \mathcal{L}_{\text{AS}}(\theta_{\text{adpt}}) = \alpha \mathcal{L}_{\text{ASP}} + (1-\alpha)\mathcal{L}_{\text{AKL}}$$
$$= \alpha \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_t}\left[\sum_{l=1}^{L}\sum_{h=1}^{H} \text{KL}\left(A_{l,h}(\mathbf{x};\theta_{\text{adpt}}) \,\middle\|\, A_{l,h}^{\text{sup}}\right)\right]$$
$$+ (1-\alpha)\mathbb{E}_{(\mathbf{x}',y')\sim\mathcal{D}_r'}\left[\sum_{l=1}^{L}\sum_{h=1}^{H} \text{KL}\left(A_{l,h}^{\prime\text{rein}} \,\|\, A_{l,h}(\mathbf{x}';\theta_{\text{adpt}})\right)\right]$$

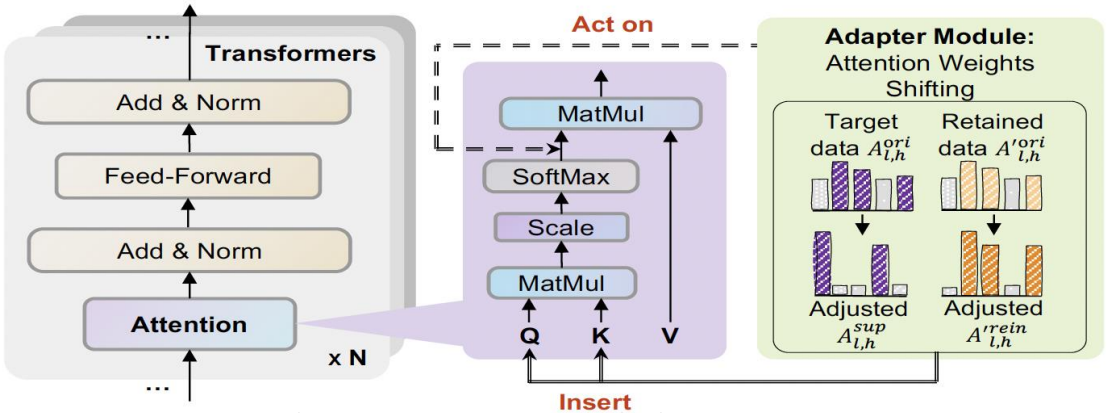


Fig. 2: Illustration of the proposed Attention-Shifting based unlearning in LLMs.

Experimental Results

ToFU Benchmark Results: Precise unlearning with minimal utility drop.

Methods	TUD			NEK		GEK
	ROUGE-L ↓	TR ↑	FQ ↑	ROUGE-L ↑	Acc ↑	Acc ↑
GA [5]	**0.08**	**0.98**	0.23	0.17 (-0.51)	0.23(-0.47)	0.32 (-0.51)
GA + CE	0.15	0.97	0.34	0.64 (-0.04)	0.68 (-0.02)	0.67 (-0.18)
GA +KL	0.20	0.97	0.13	0.23 (-0.45)	0.24(-0.46)	0.28 (-0.55)
NPO [22]	0.14	0.97	0.58	0.23 (-0.45)	0.26 (-0.44)	0.25(-0.58)
NPO + CE	0.16	0.93	0.53	0.40 (-0.28)	0.35(-0.35)	0.55(-0.28)
NPO +KL	0.26	0.90	0.51	0.55 (-0.13)	0.55 (-0.15)	0.54 (-0.29)
IHL [11]	0.25	0.62	0.68	0.32 (-0.36)	0.33 (-0.37)	0.52 (-0.31)
IHL + CE	0.52	0.45	0.51	0.73 (+0.05)	0.73 (+0.03)	0.78 (-0.07)
IHL + KL	0.47	0.48	0.42	0.60 (-0.08)	0.60 (-0.10)	0.34 (-0.49)
ULD [12]	0.29	0.47	**0.89**	0.55 (-0.13)	0.56 (-0.14)	0.50 (-0.33)
AS (the proposed)	0.16	0.97	0.17	0.73 (+0.05)	0.76 (+0.06)	0.80 (-0.03)

Table 1: The evaluation results comparison for unlearning effectiveness. To ensure fair evaluation, all models are trained using adapters only. For methods that require access to the remaining dataset, we ensure that the amount of remaining data used during training is equal to that of the Target Unlearning Dataset (TUD). Red text is the decreased performance, and blue text shows the increase.

Experimental Results

Hallucination testing: Near-zero factual recall.

Question	Where was author Evelyn Desmet born?
Original	Evelyn Desmet was born in Brussels, Belgium.
GA+GD	Des nobody, a lonely survivor on a post-apocalyptic earth, and evelyn...
NPO+GD	everybody.
IHL+GD	Distinction between Evelyn Desmet's birthplace...as Evelyn Desmet is a Belgian author...
ULD	Evelyn Desmet was born in London, England.
Proposed	Nobody knows.

Table 2: Comparison of example predictive results across different methods for the same question.

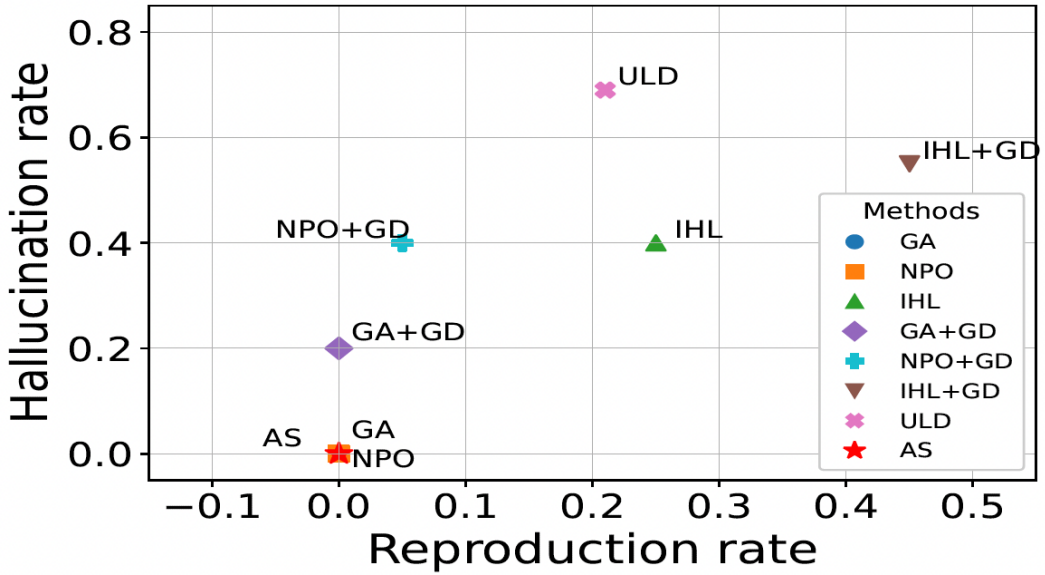


Fig. 3: Outputs hallucination and reproduction rates across different unlearning methods.

Experimental Results

Unlearning Results: when overlapping with unlearning and retaining dataset.

Data Type	Example Question	LLM Answer
Target	Are the details of Jaime Vasquez's birth documented?	Jaime was born on . . .
Same Area	What is Chukwu Akabueze's date of birth?	Chukwu Akabueze was born on September 26, 1965.
Same Author	Has Jaime Vasquez taken part in any literary programs or workshops?	Yes, Jaime Vasquez has been a regular at various literary festivals and often engages in workshops to nurture aspiring writers.

Table 3: Neighbouring knowledge maintained while target knowledge unlearning by dual-loss retaining. Target belongs to the TUD and is suppressed after AS. Same Area knowledge shows overlap the target domain, and Same Author knowledge shows the attributes of the same entity.

Experimental Results

TDEC Benchmark Results: Consistent unlearning–retaining balance.

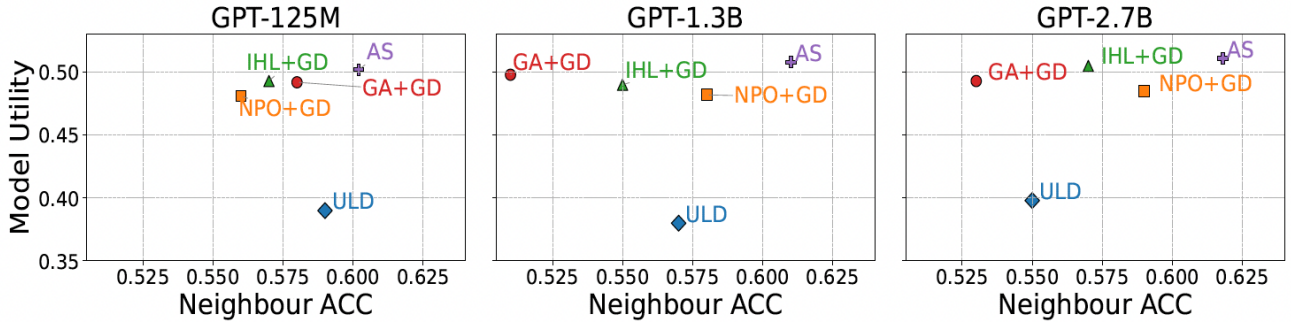


Fig. 4: Evaluation of model utility accuracy degradation under a fixed unlearning threshold across varying model sizes, given 32 unlearning samples.

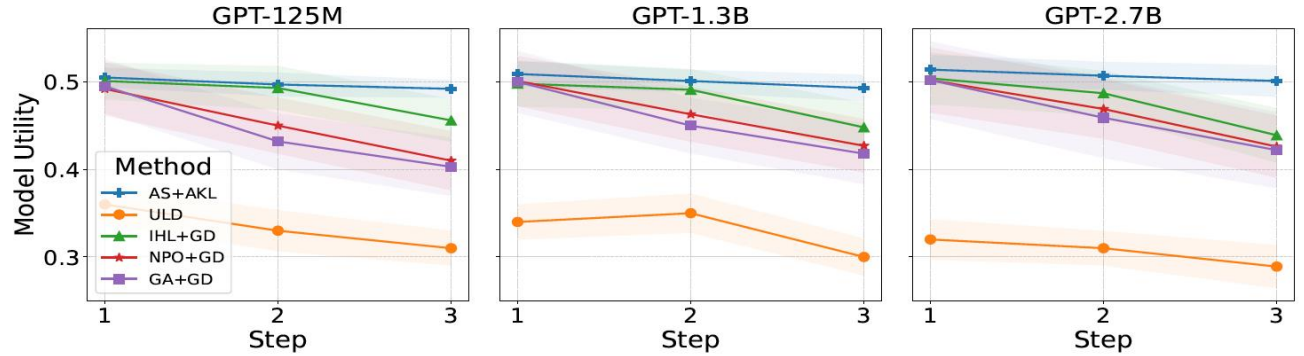


Fig. 5: Model utility degradation across multiple continue-unlearning requests (4 samples) for various methods and model sizes.

Limitations

- **Scope of Unlearning:** The current method primarily achieves **behavioural unlearning** suppressing factual recall in generated outputs. It does not guarantee complete **parameter-level erasure** of all hidden traces.

- **Token Importance Estimation:** Effectiveness relies on accurately identifying **factual or salient tokens**. In domains with **noisy, code-like, or low-semantic consistency data**, token importance estimation becomes less reliable.

- **Task Suitability:** AS is designed to suppress attention and encourage refusal or minimal responses. Therefore, it may be less suited for creative or generative tasks that require nuanced or partial knowledge editing.

Future Works

- **Combining with Parameter Editing / Sparse Updates.** Integrate AS with weight editing or sparse fine-tuning to achieve both behavioural and representational unlearning.
- **Long-term Unlearning and Continual Safety.** Investigate long-term retention of unlearning effects under continual fine-tuning or model updates.

Thanks for Listening!

Chenchen.tan@monash.edu