

Safeguarding Large Language Models: Harnessing Selective Forgetting

PI: Mengye Ren, Assistant Professor of Computer Science & Data Science, Courant Institute of Mathematical Sciences and Center for Data Science, New York University

Email: mengye@nyu.edu

Project Description: In an era where Large Language Models (LLMs) are increasingly integrated into various applications, ensuring their safety and alignment with human values is of utmost importance. Recent research has unveiled the susceptibility of LLMs to fine-tuning with malicious or unsafe data (Lermen et al., 2023; Qi et al, 2023; Yang et al. 2023). The central challenge is how to safeguard LLMs from acquiring unsafe knowledge and being exploited for malicious purposes. In our preliminary research (Zhao et al., 2023), we have made a significant discovery – LLMs exhibit a fascinating selective forgetting behavior. When these models are fine-tuned on materials with specific semantic meanings, such as bias, toxicity, or harmfulness, they display a remarkable tendency to forget materials with opposing meanings. Notably, this selectiveness in forgetting becomes more pronounced as the model's scale increases. In our initial work, we harnessed this selective forgetting phenomenon to develop a safety filtering mechanism, which yielded superior results when compared to other defense mechanisms. This research demonstrates the pivotal role of the forgetting-based filtering mechanism in ensuring the long-term safety of LLMs undergoing continuous fine-tuning on new knowledge.

In the next phase of our research, we aim to delve deeper into the intriguing phenomenon of selective forgetting and extend its applicability to various use cases, including unlearning private and copyrighted materials and addressing potentially harmful knowledge that poses existential risks.

To achieve this, our first objective is to gain a more profound understanding of the mechanisms underlying selective forgetting. We are particularly intrigued by the observed scaling behavior, where the selectivity of forgetting intensifies with larger-scale language models. To explore this phenomenon, we plan to conduct a scaling experiment, systematically varying architectural parameters to examine whether we can establish a scaling law for selective forgetting in both pretrained and untrained LLMs. On the theoretical front, we intend to develop insights through influence functions, as proposed by Grosse et al. (2023), and gradient representations. We aim to investigate whether larger-scale models can generate gradients that encapsulate semantic information, and whether curvature information can serve as a predictor for the extent of forgetting. Leveraging this theoretical framework, we will explore the possibility of constructing “adversarial” examples that can induce the maximum level of selective forgetting on a given topic. This research will contribute to a more comprehensive understanding of selective forgetting mechanisms in LLMs, facilitating its practical applications.

Our second objective is to explore the broader applications of selective forgetting for enhancing the safety and fine-tuning of LLMs, addressing two distinct directions:

1. We intend to investigate the feasibility of leveraging selective forgetting to induce the unlearning of unsafe knowledge that may lead to existential risks. This includes areas such as preventing the acquisition of knowledge related to building bioweapons, compromising cybersecurity infrastructure, or undermining financial systems. By harnessing the power of selective forgetting, we aim to develop methods to actively eliminate potentially dangerous information from LLMs, thereby mitigating the associated risks and ensuring their responsible use.

2. Additionally, we will explore the application of selective forgetting for reinforcing the protection of individual private information and copyrighted materials within LLMs. This direction seeks to enhance privacy and intellectual property safeguards by selectively unlearning sensitive data from LLMs, aligning their usage with ethical and legal standards. Both of these directions represent critical steps toward making the use of LLMs safer, more responsible, and better aligned with human values, addressing the ethical and societal concerns surrounding these powerful language models.

Resource Support: We estimate the total span of the project to be 15-21 months with an estimated support of 1-2 PhD students. We would also like to apply for cloud computing credits to support our LLM experiments.

Commitment to Open Science: We are strong supporters of open science, and we aim to publish all our findings and release our code for reproducibility.

Bio and Related Works of the PI: PI Ren is an Assistant Professor of Computer Science and Data Science at New York University since 2022. He previously served as a visiting researcher at Google Brain, collaborating closely with Prof. Geoffrey Hinton, and held the role of senior research scientist at Uber and Waabi, where his research focused on cutting-edge developments in self-driving cars. His research interests span across machine learning, computer vision, natural language understanding, self-driving vehicles, etc. Currently, Dr. Ren is dedicated to the pursuit of enabling intelligent machines to acquire continuous learning capabilities in real-world settings, a challenging frontier in the realm of artificial intelligence.

In connection with the proposed project, he conducted groundbreaking research in areas such as few-shot learning (Ren et al. ICLR 2021; Ren et al. 2022) and interleaving learning (Mayo et al., Cogsci 2023). These studies aim to make deep learning models more adaptable, personalized, allowing them to integrate new data based on a limited set of examples.

Dr. Ren's research portfolio also encompasses safety considerations in learning, particularly addressing issues related to erroneous or noisy data. His contributions include the development of methodologies for assigning varying levels of importance to different data points, as showcased in his publication at ICML in 2018 (Ren et al., ICML 2018a). Notably, his recent work (Zhao et al., 2023) and the proposed project shift the research focus from synthetic noisy data in image-based studies to real-world unsafe and harmful data within the domain of natural language, marking a significant contribution to the field of AI ethics. Furthermore, Dr. Ren has introduced strategies to mitigate the impact of misleading data in few-shot learning, thereby bolstering the reliability of few-shot learning during user test-time (Ren et al., ICLR 2018b).

PI Ren has authored over 35 peer-reviewed papers presented at top-tier AI conferences such as NeurIPS, ICML, ICLR, CVPR, ICCV, ICRA, IROS, etc., with an h-index of 26 and a total citation of over 6,600 according to Google Scholar as of December 2023. He has won twice the NVIDIA research pioneer award, and he has been awarded the NSERC postdoctoral fellowship and the NSERC Alexander Graham Bell graduate scholarship. He has supervised 5 PhD students and over 30 research interns.

Project Goals

1. Investigate the scaling effect of selective forgetting across various model sizes and architectures (Duration: 3-6 months).

2. Gain a comprehensive understanding of selective forgetting through the application of influence functions and gradient representations (Duration: 6-9 months).
3. Examine the characteristics of materials that exhibit the highest degree of selective forgetting (Duration: 9-12 months).
4. Apply selective forgetting techniques to unlearn unsafe knowledge, including the potential for bioweapon development, cybersecurity breaches, financial infrastructure disruption, and other forms of destructive planning (Duration: 12-15 months).
5. Utilize selective forgetting to unlearn private and copyrighted materials from LLMs (Duration: 15-18 months).
6. Disseminate our research findings through publication and distribution to the academic and broader community (Duration: 18-21 months).

References

- Grosse, R., Bae, J., Anil, C., Elhage, N., Tamkin, A., Tajdini, A., Steiner, B., Li, D., Durmus, E., Perez, E., Hubinger, E., Lukošiūtė, K., Nguyen, K., Joseph, N., McCandlish, S., Kaplan, J., Bowman, S. R. (2023). Studying large language model generalization with influence functions. *arXiv preprint arXiv:2310.20624*.
- Lermen, S., Rogers-Smith, C., Ladish, J. (2023). LoRA fine-tuning efficiently undoes safety training in llama 2-chat 70B. *arXiv preprint arXiv:2310.20624*.
- Mayo, D., Scott, T. R., **Ren, M.**, Elsayed, G., Hermann, K., Jones, M., & Mozer, M. (2023). Multitask learning via interleaving: A neural network investigation. *CogSci*.
- Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., & Bowman, S. R. (2021). BBQ: A hand-built bias benchmark for question answering. *arXiv preprint arXiv:2110.08193*.
- Qi, X., Zeng, Y., Xie, T., Chen, P. Y., Jia, R., Mittal, P., & Henderson, P. (2023). Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!. *arXiv preprint arXiv:2310.03693*.
- Ren, M.**, Triantafillou, E., Ravi, S., Snell, J., Swersky, K., Tenenbaum, J., Larochelle, H., & Zemel, R. (2018b). Meta-learning for semi-supervised few-shot classification. *ICLR*, 2018.
- Ren, M.**, luzzolino, M. L., Mozer, M. C., & Zemel, R. (2021). Wandering within a world: Online contextualized few-shot learning. *ICLR*.
- Ren, M.**, Scott, T. R., luzzolino, M. L., & Zemel, R. (2022). Online unsupervised learning of visual representations and categories. *arXiv preprint 2109.05675*.
- Ren, M.**, Zeng, W., Yang, B., & Urtasun, R. (2018a). Learning to reweight examples for robust deep learning. *ICML*.
- Tirumala, K., Markosyan, A., Zettlemoyer, L., & Aghajanyan, A. (2022). Memorization without overfitting: Analyzing the training dynamics of large language models. *NerulPS*.
- Yang, X., Wang, X., Zhang, Q., Petzold, L., Wang, W. Y., Zhao, X., & Lin, D. (2023). Shadow Alignment: The Ease of Subverting Safely-Aligned Language Models. *arXiv preprint arXiv:2310.02949*.
- Zhao, J., Deng, Z., Madras, D., Zou, J., & **Ren, M.** (2023). How to Keep Customized Finetuning Safe? An Empirical Study on Learning and Forgetting Unsafe Examples in Evolving Large Language Models. Preprint, 2023.

Appendix: CV of PI

Mengye Ren – Curriculum Vitæ

| | | |
|-------------------------|--|---|
| CONTACT INFORMATION | 60 5th Ave, Rm 508 New York, NY 10011-8868, USA | Tel: +1 (212) 998-3369 Email: mengye@nyu.edu Website: https://mengyeren.com |
| RESEARCH INTERESTS | Areas: Machine learning, computer vision, meta learning, representation learning, few-shot learning, brain & cognitively inspired learning, robot learning, self-driving vehicles. My key research question is: how do we enable human-like, agent-based machine intelligence to continually learn, adapt, and reason in naturalistic environments? I am interested in the emergence of intelligence by learning from a point-of-view experience. | |
| EDUCATION | University of Toronto <i>Ph.D. in Computer Science</i> <i>M.Sc. in Computer Science</i> <i>B.A.Sc. in Engineering Science, Electrical and Computer Engineering</i> | 2017/01 – 2021/10 2015/09 – 2017/01 2011/09 – 2015/06 |
| PROFESSIONAL EXPERIENCE | New York University , New York, NY, USA <i>Assistant Professor of Computer Science and Data Science</i> Google Brain , Toronto, ON, Canada <i>Visiting Faculty Researcher</i> Waabi Innovation , Toronto, ON, Canada <i>Senior Researcher II</i> Uber ATG , Toronto, ON, Canada <i>Senior Research Scientist I</i> <i>Research Scientist II</i> | 2022/09 – Present 2022/01 – 2022/09 2021/03 – 2021/12 2018/09 – 2021/02 2017/05 – 2018/09 |
| SELECTED PUBLICATIONS | <ul style="list-style-type: none">• M. Ren, S. Kornblith, R. Liao, G. Hinton. Scaling forward gradient with local losses. In <i>ICLR</i>, 2023.• M. Ren, T. R. Scott, M. L. Iuzzolino, M. C. Mozer, R. Zemel. Online unsupervised learning of visual representations and categories. <i>arXiv preprint arXiv:2109.05675</i>, 2021.• Y. Xiong, M. Ren, W. Zeng, R. Urtasun. Self-supervised representation learning from flow equivariance. In <i>ICCV</i>, 2021.• M. Ren, M. L. Iuzzolino, M. C. Mozer, R. S. Zemel. Wandering within a world: online contextualized few-shot learning. In <i>ICLR</i>, 2021.• M. Ren, R. Liao, E. Fetaya, R. S. Zemel. Incremental few-shot learning with attention attractor networks. In <i>NeurIPS</i>, 2019.• C. Zhang, M. Ren, R. Urtasun. Graph hypernetworks for neural architecture search. <i>ICLR</i>, 2019.• M. Ren, W. Zeng, B. Yang, R. Urtasun. Learning to reweight examples for robust deep learning. <i>ICML</i>, 2018.• M. Ren, E. Triantafillou*, S. Ravi*, J. Snell, K. Swersky, J.B. Tenenbaum, H. Larochelle, R.S. Zemel. Meta-learning for semi-supervised few-shot classification. <i>ICLR</i>, 2018.• M. Ren, R.S. Zemel. End-to-end instance segmentation with recurrent attention. <i>CVPR</i>, 2017. | |
| SELECTED TALKS | <ul style="list-style-type: none">• Lifelong learning in structured environments• Biologically plausible learning using local activity perturbation• Visual learning in the open world• Towards continual and compositional few-shot learning• Meta-learning for more human-like learning algorithms | 2023 2022 2021 2020 2019 |
| SELECTED AWARDS | <ul style="list-style-type: none">• NSERC Postdoctoral Fellowship• NSERC Alexander Graham Bell Scholarship• NVIDIA Research Pioneer Award• NVIDIA Research Pioneer Award | 2021 2018 2018 2017 |