# Understanding the Effects of Representation Misdirection for Large Language Model Unlearning

s2310417 Huu-Tien Dang

Modern large language models (LLMs) are pre-trained on massive text corpora from the web, including copyrighted material, toxic and sexual content, sensitive and private information, gender and political bias, and dangerous documents such as cybersecurity attacks and bioweapon development. As a result, LLMs can exhibit harmful and unwanted behaviors. The right to be forgotten emerged for LLMs as a tool to ensure their and our safety. Machine unlearning is an approach that aims to *remove* or *suppress* the target forget knowledge and capabilities from a pre-trained model while maintaining the model's other knowledge and capabilities. Representation Misdirection for Unlearning (RMU)—an approach that performs unlearning by manipulating the latent representations of the forget-samples in the pre-trained models—establishes LLM unlearning methods with state-of-the-art performance. Yet, the underlying causes and explanations remain underexplored.

In this thesis, we show the following: (1) a theoretical analysis that demonstrates steering the forget-representations to a target random representation of RMU's objective reduces forget-token confidence, causing LLMs to generate wrong or nonsense responses. (2) How the coefficient influences the alignment of forget-representations with the random direction and hints at the optimal coefficient values for effective unlearning across different network layers. (3) Analyzing RMU's robustness against white-box knowledge recovery attacks through the lens of an attack–defense game: RMU acts as a defender, impeding the adversary's ability to determine optimal updates for generating adversarial samples, thus improving the adversarial robustness of unlearned models. (4) RMU's forget-loss, which minimizes the mean squared error between forget-representation and a fixed scaled random vector, fails to converge when the norm of the forget-representation is larger than the scaling coefficient, making RMU less effective when applied to middle and deep layers in LLMs. To overcome this limitation, we introduce *Adaptive RMU*, a simple yet effective alternative to RMU, that adaptively adjusts the coefficient based on the norm of forget-representations. Extensive experiments demonstrate that Adaptive RMU achieves higher drop-in accuracy for forgetting knowledge, maintains high performance on general knowledge, and enables effective unlearning for most layers without incurring additional computational overhead. Our implementation is available at https://github.com/huutiendang/llm-unlearning.