Deep learning has recently driven tremendous progress in a wide array of applications, ranging from realistic image generation and impressive retrieval systems to language models that can hold human-like conversations.

While this progress is very exciting, the widespread use of deep neural network models requires caution: as guided by Google’s AI Principles, we seek to develop AI technologies responsibly by understanding and mitigating potential risks, such as the propagation and amplification of unfair biases and protecting user privacy.

Fully erasing the influence of the data requested to be deleted is challenging since, aside from simply deleting it from databases where it’s stored, it also requires erasing the influence of that data on other artifacts such as trained machine learning models.

Moreover, recent research [1, 2] has shown that in some cases it may be possible to infer with high accuracy whether an example was used to train a machine learning model using membership inference attacks (MIAs).

This can raise privacy concerns, as it implies that even if an individual's data is deleted from a database, it may still be possible to infer whether that individual's data was used to train a model.