

PRIVACY ERASE MODEL TRAINING METHOD AND APPARATUS

AND PRIVACY ERASE METHOD AND APPARATUS

ABSTRACT

Implementations of the present specification disclose a privacy erase model training method.

5 In the training method, a large language model is separately trained based on original training data and anonymized data of the original training data by using completely same training methods, and weights of the same large language model on the original training data and the anonymized data are recorded, to form a new erase data training set. A privacy erase model can be trained by using the erase data training set, to erase weight data related to privacy data from the large language

10 model. In the method, the privacy erase model is trained to directly modify a parameter of the large language model, so that the privacy data in memory of the large language model is fundamentally deleted, thereby achieving extremely high security. In the method, the large language model does not need to be retrained, so that optimization costs of the large language model are greatly reduced. A privacy erase model training apparatus and a privacy erase method and apparatus in the

15 implementations of the present specification also have the above effects.

FIG. 1