UNLEARNING

1. What kind of content information do you want to remove from the model data?
    a. Biased or discriminatory data
    b. Useful patterns and trends
    c. General public data
    d. Random noise
    e. Personally identifiable information
    f. Valid and accurate data

2. What are the reasons to choose unlearning over retraining?
    a. To improve the overall performance of the model
    b. To add new data to the model
    c. To change the underlying algorithm of the model
    d. To completely overhaul the model's architecture
    e. To save computational resources and time

3. Identify the steps involved in the exact unlearning as discussed in the course.
    a. Isolate the data -> shard the data -> slice the data -> aggregate the data
    b. Aggregate the data -> isolate the data -> slice the data -> shard the data
    c. Shard the data -> Slice the data -> Isolate the data -> Aggregate the data
    d. Shard the data -> Isolate the data -> Slice the data -> Aggregate the data
    e. Isolate the data -> slice the data -> shard the data -> aggregate the data

4. Which model should be retrained in the exact unlearning process?
    a. The constituent model that is trained over the isolated data
    b. The constituent model that is trained over the sharded data
    c. The constituent model that is trained over the aggregated data
    d. The constituent model that is trained over the sliced data

5. How should the original model and the model after the below unlearning methods behave?
    1) exact unlearning
    2) approximate unlearning
    a. 1) distributionally identical 2) distributionally identical
    b. 1) distributionally close 2) distributionally close
    c. 1) distributionally identical 2) distributionally close
    d. 1) distributionally close 2) distributionally identical

6. How does unlearning via differential privacy work?

   a. check whether an adversary can reliably tell apart the models before unlearning and after unlearning
   b. check whether the model can output private and sensitive information before and after unlearning
   c. check whether the model's predictions become more consistent and stable for private information before and after unlearning.
   d. check whether an adversary can identify the differences in the distribution of output data of the model before and after unlearning

7. Identify all the methods for privacy unlearning.

   a. Gradient descent on encountering the forget set
   b. Remove noise from the weights influencing the forget set
   c. Add noise to weights influencing data in forget set
   d. Gradient ascent on encountering the forget set
   e. Increase the learning rate when encountering the forget set
   f. Apply dropout to all layers when encountering the forget set

8. Match the unlearning method to their corresponding concept

1) privacy unlearning	I. data and model architecture is not modified
2) concept unlearning	II. use membership inference attack concept
3) example unlearning	III. forget set is not clearly defined
4) ask for unlearning	IV. forget set is clearly defined

   a. 1-III, 2-I, 3-IV, 4-II
   b. 1-II, 2-III, 3-IV, 4-I
   c. 1-IV, 2-II, 3-I, 4-III
   d. 1-I, 2-IV, 3-II, 4-III
   e. 1-IV, 2-I, 3-III, 4-II

9. The forget set to be unlearned is not known in which of the following:

   a. Example Unlearning
   b. Differential Privacy Unlearning
   c. Privacy unlearning
   d. Concept unlearning

10. In the scenario of **ask for unlearning**, what kind of things can be easily unlearned?

   a. Hate speech
   b. Toxic content
   c. Factual Information
   d. Sensitive information

11. When evaluating the quality of unlearning using Membership Inference Attack, which of the following scenarios implies that the unlearning is successful?
    a. The accuracy increases on the forget set
    b. The accuracy drops on the forget set
    c. The accuracy stays the same on the forget set
    d. The accuracy increases on the test set
    e. The accuracy drops on the test set
    f. The accuracy stays the same on the test set

12. What are some metrics to evaluate the unlearning?
    a. If it was more computationally efficient compared to retraining
    b. Increased size of the original dataset
    c. If the unlearning retains information derived from the concept to be forgotten
    d. If the performance has been maintained before and after unlearning

13. In an interclass confusion scenario where confusion is synthetically added to a dataset by label flipping for some of the concepts, identify the kind of unlearning method that can be used to unlearn the data points that have their labels flipped. Assume that you have the entire data points for which the labels were flipped.
    a. Concept unlearning
    b. Example Unlearning
    c. Differential Privacy Unlearning
    d. Exact unlearning
    e. Ask to forget

14. What idea does the paper Corrective Machine Learning build upon?
    a. Not all poisoned data can be identified for unlearning
    b. Identifying and removing a small subset of poisoned data points is sufficient to ensure the model's integrity
    c. enhancing the model's ability to handle completely new, unseen poisoned data
    d. The accuracy of the model improves proportionally with the amount of data removed, regardless of whether it is poisoned or not
    e. adding redundant data to the dataset to counteract the effects of poisoned data.
    f. Not all poisoned data can be identified for unlearning

15. Identify all the methods that act as the baseline for the TOFU benchmark dataset
    a. Gradient Descent
    b. Gradient Ascent
    c. Gradient Difference
    d. Gradient boosting
    e. Gradient Clipping

16. The WMDP benchmark tests on unlearning what kind of information?

    a. Biosecurity
    b. High-school biology
    c. Hate speech on Twitter
    d. Crime data

17. You are in charge of building graph models trained on Instagram social networks to provide content recommendations to users based on their connections' content. You realize that a particular user in the network is leading to toxic content recommendations. What kind of unlearning would you use in this scenario to prevent the recommendation of toxic content?

    a. Node feature unlearning
    b. Node unlearning
    c. Edge Unlearning
    d. Subgraph unlearning

18. In Representation Engineering, what is the **representation**?

    a. Attention heads affecting the data
    b. Positional embeddings of the data
    c. Activations of the layer affecting the data
    d. The encoder of a transformer
    e. The decoder of a transformer