ROBUSTNESS, RLHF, AI Alignment - WEEK 2

1. **A robust model provides unreliable predictions when met with adversaries. Which all of the following are common adversaries in this context?**
    a. Distribution Shift
    b. Overfitting
    c. Noisy Data
    d. Model Compression
    e. Gradient Descent
    f. Data Augmentation

2. **In the context of AI research, which of the following events could be considered a black swan?**
    a. Incremental improvements in natural language processing algorithms.
    b. The consistent performance of AI models on standard benchmarks.
    c. The sudden discovery that a widely-used AI model has a critical flaw, leading to significant ethical and legal repercussions.
    d. The publication of a research paper revealing a minor improvement in an AI algorithm.
    e. A gradual increase in the accuracy of AI models over time.

3. **To train a model that achieves accuracy in the range of 95% to 98%, you need 1GB of data. To get 100% accuracy, you need 120GB of data. This idea is similar to which of the following principles:**
    a. Sigmoid Distribution
    b. Power law distribution
    c. Uniform distribution
    d. Gaussian Distribution
    e. Long-tailed distribution

4. **Identify the equations that can lead to a long-tailed distribution.**
    a. Idea * student * resources * time
    b. Idea * student + resources * time
    c. Idea + student + resource + time
    d. Idea - student * resource - time

5. **Black Swan lies in which of the following categories?**
    a. Known Knowns
    b. Known Unknowns
    c. Unknown Knowns
    d. Unknown Unknowns

6. **Match the items below with their corresponding descriptions.**

Column A	**Column b**
I. Known Knowns	**A. Close-ended questions**
II. Known Unknowns	**B. Recollection**
III. Unknown Knowns	**C. Open-ended exploration**
IV. Unknown Unknowns	**D. Self-Analysis**

    a. I-C, II-D, III-A, IV-B
    b. I-B, II-A, III-D, IV-C
    c. I-A, II-B, III-C, IV-D
    d. I-D, III-C, III-B, IV-A

7. **Why is Black Swan and Long-tailed distribution important?**
    a. Understand small things that have a small but useful effect
    b. Understand large things that have a small but useful effect
    c. Understand small things that have a catastrophic effect
    d. Understand large things that have a catastrophic effect

8. **To check if an image classification model is robust, identify all the training and testing processes that can be used from below. The three datasets are ImageNet, AugMix, Mixup**
    a. Train on AugMix and test on AugMix
    b. Train on AugMix and test on ImageNet
    c. Train on ImageNet and test on AugMix
    d. Train on Mixup and test on ImageNet
    e. Train on ImageNet and test on ImageNet
    f. Train on ImageNet and test on Mixup

9. **Identify all the conditions to check if a model is robust?**
    a. Models with larger parameters
    b. Models with small parameters
    c. Models that can generalise better
    d. Models trained to perform the best on a specific type of data

10. **Which of the following are some data augmentation methods?**

a. DataShrink
b. AugMix
c. Label Flipping
d. Mixup

11. **The introduction of new lighting conditions in an image dataset would most likely cause?**
    a. Distribution Shift
    b. Concept Shift
    c. Model Decay
    d. Feature Extraction

12. Identify the goal(s) of a model when training with RLHF is as follows:
    a. Maximize the penalty
    b. Maximize the reward
    c. Minimize the penalty
    d. Minimize the reward

13. Identify the step(s) involved in RLHF pipeline:
    a. Supervised fine-tuning
    b. Unsupervised fine-tuning
    c. Reward model training
    d. Penalty model training
    e. Proximal Policy optimization
    f. Convex Policy Optimization

14. Identify issue(s) associated with RLHF from below:
    a. It does not perform as well as supervised learning
    b. Performance sensitive to hyperparameters
    c. It does not perform as well as unsupervised learning
    d. Fitting and optimization of the reward function is computationally expensive
    e. Pretrained models easily outperform them in tasks like summarization

15. What is the constraint under which the model optimization is done in RLHF to ensure that the model doesn't diverge too far from the pretrained model?
    a. KL Divergence
    b. L2 Regularization
    c. Entropy Maximization
    d. Gradient Clipping

16. What are the issues with reward modelling?

a. Reward shrinking - gradually decreasing rewards over time
b. Reward misalignment - reward signals do not align with the desired outcomes
c. Reward saturation - model stops learning after a certain reward threshold is reached
d. Reward consistency - ensuring rewards are uniformly distributed
e. Reward hacking - maximise reward with imperfect proxy and forget the goal

17. Direct Preference of Optimization works in which one of the following ways:

a. RLHF without rewards model
b. RLHF without human feedback
c. RLHF without reinforcement learning
d. RLHF without KL divergence

18. Identify the issues with human feedback in RLHF

a. Overabundance of feedback
b. Consistent and uniform feedback
c. Biased and harmful feedback
d. Feedback redundancy

19. Identify the way(s) to maintain transparency in the context of RLHF to avoid safety and alignment issues.

a. Quality-assurance measures for human feedback
b. Minimize the involvement of humans to reduce biases
c. Use black-box algorithms to simplify the process
d. Avoid documenting the feedback process to save time
e. Limit the diversity of human feedback to ensure consistency
f. Have a powerful loss function when optimizing the reward model

20. What is distribution shift in machine learning?

a. The training distribution is not similar to the test distribution
b. The model's parameters change during training
c. The target variable's distribution changes over time
d. The model's prediction accuracy improves on new data

21. What is data poisoning in the context of machine learning?

a. Removing important features from the dataset

b. Oversampling minority classes in imbalanced datasets
c. Adding hidden functionalities to control the model behaviour
d. Encrypting sensitive information in the training data

22. When is a data poisoning attack considered successful?

a. The model's overall accuracy decreases significantly
b. The model fails to converge during training
c. The model becomes computationally inefficient
d. The model outputs the specific behaviour when it encounters the trigger

23. **Consider the following scenario:**
You have a dataset with 10000 samples and a model trained over it has a test accuracy of almost 94%. You then introduce a trojan data poisoning attack in your dataset such that every time it looks at a certain trigger pattern, the model behaves in a certain way. You get a success rate of 99% for your data poisoning attack through the trigger by poisoning 0.01% of your samples. What is the new test accuracy of your model? Identify the correct range.

a. 40 to 50 percent
b. 50 to 60 percent
c. 60 to 70 percent
d. 70 to 80 percent
e. 80 to 90 percent
f. 90 to 100 percent

24. **What are the different defence mechanism(s) against poisoning attacks?**

a. Biasing
b. Filtering
c. Unlearning
d. Representation engineering
e. AutoDebias