Model Training

- Ensure the code, environment, and data are correct before starting actual training.
- Check the maximum running timeout of the training environment. Check the balance of the cloud platform. Ensure nothing will interrupt the training.
- Ensure that if there is an interruption, we can resume the task. The intermediate results will not be lost.
- If there is a problem with the code, it will surely happen during training.
- If the data is not clean, it will surely impact the later phases.
- Use an iterative approach. Scale by orders of magnitude. Train on millions of tokens, then billions of tokens, and then trillions of tokens.
- If training is over, what will happen? Is the final handling of training done correctly?
- Learn from the mistakes of others, like those made with Llama2.
- Instead of fixing the result, address the root cause of the issue.