Topic 9: Efficiency and Optimizations

1. LLM Distillation Demystified: A guide explaining how to compress a large language model by training a smaller “student” model using the outputs of a big “teacher” model: https://snorkel.ai/blog/llm-distillation-demystified-a-complete-guide/
2. Train With Mixed Precision: NVIDIA’s guide to mixed-precision training: https://docs.nvidia.com/deeplearning/performance/mixed-precision-training/index.html
3. Mixed Precision Training Overview: https://blog.paperspace.com/mixed-precision-training-overview/
4. Quantization: A practical tutorial for applying quantization to LLMs: https://www.datacamp.com/tutorial/quantization-for-large-language-models
5. Knowledge Distillation: A video walkthrough of how knowledge distillation works for large and small language models: https://www.youtube.com/watch?v=FIOigevZdDU
6. Framework-specific guidance for PyTorch users on how to implement mixed-precision training, alongside tradeoffs and best practices: https://pytorch.org/blog/what-every-user-should-know-about-mixed-precision-training-in-pytorch/
7. A Visual Guide to Quantization: https://maartengrootendorst.com/blog/quantization/