Compression and Distillation

- are for making LLMs efficient and usable

and help make smaller versions of LLMs Like Open AI's Mini models

1) Model compression (General concept)

- refers to reducing the size and computation requirements of a Large model without significantly hurting performance

- common techniques
  - pruning: Removing weights or neurons that contribute little to performance
  - Quantization: Representing weights and activations with fewer Bits (Float32 → int8)
  - Weight sharing: Reusing certain weights across layers to reduce Storage
  - Low rank factorization: Approximating large Weight matracies using Products of smaller ones

2) Knowlage distillation (specific technique)

- distillation is a type of compression where:
  - you have large model (teacher)
  - you have smaller model (student) to mimic the teachers behaviour/ Output

- how it works
  1. Run inputs through the teacher model
  2. get the soft predictions (probability distributions, not just hard labels)
  3. train student to match those soft predictions (along with true labels sometimes)
  4. the student learns not just answers, but also the confidence and nuance the teacher learned

Why: for smaller models to run of phones while retaining much of the performance of the larger model + used in RAG, MCP, agents

NOTE: Compression: Outputs same model but optimized

distillation: outputs new smaller model