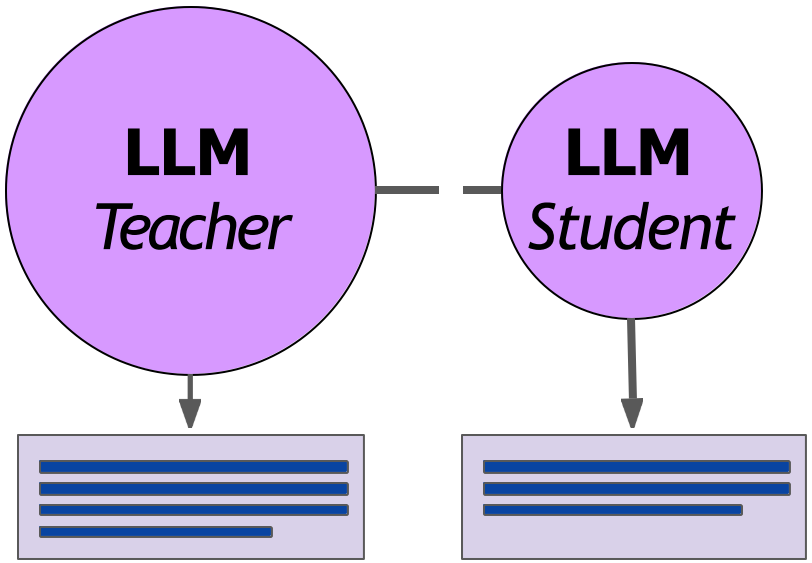
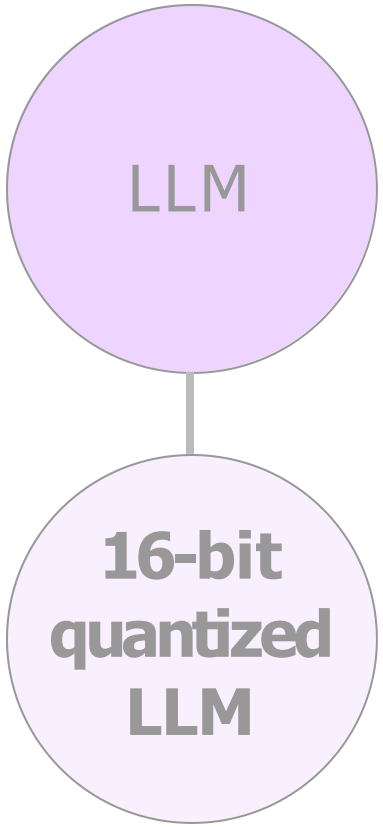
LLM optimization techniques

LLM optimization techniques

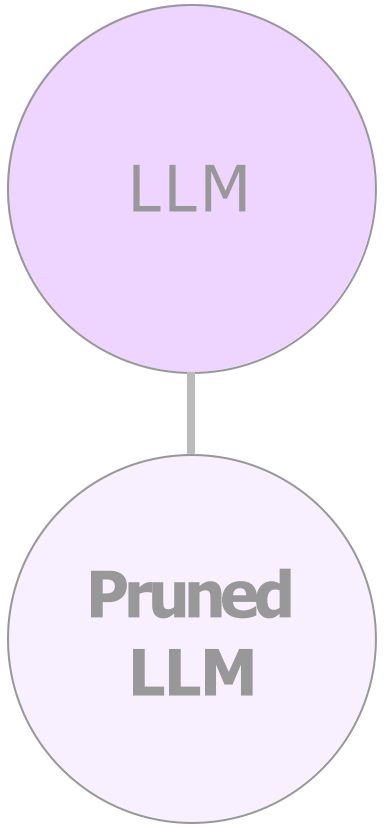
Distillation



Quantization

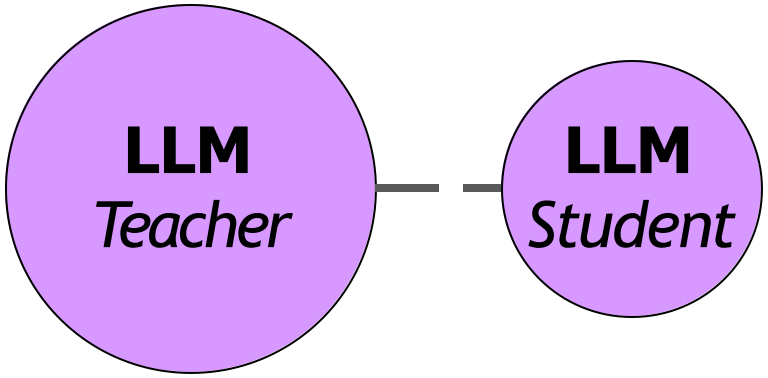


Pruning

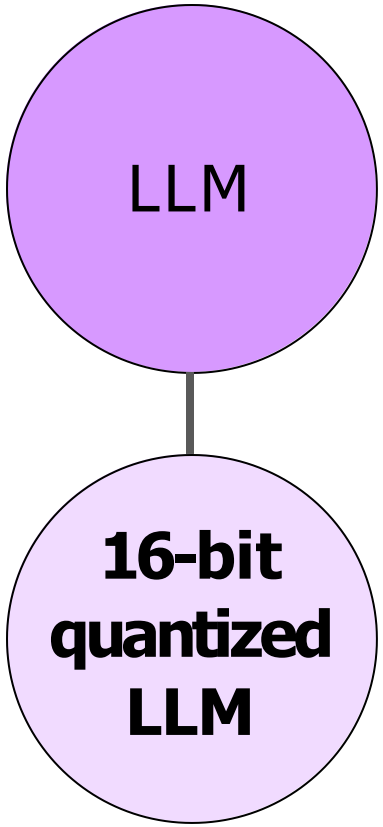


LLM optimization techniques

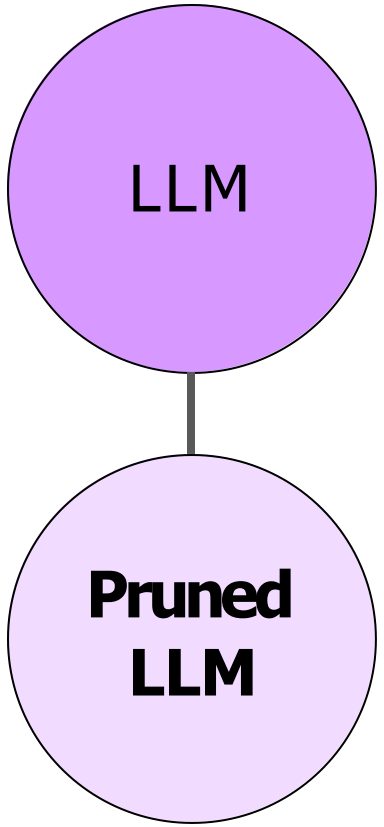
Distillation



Quantization

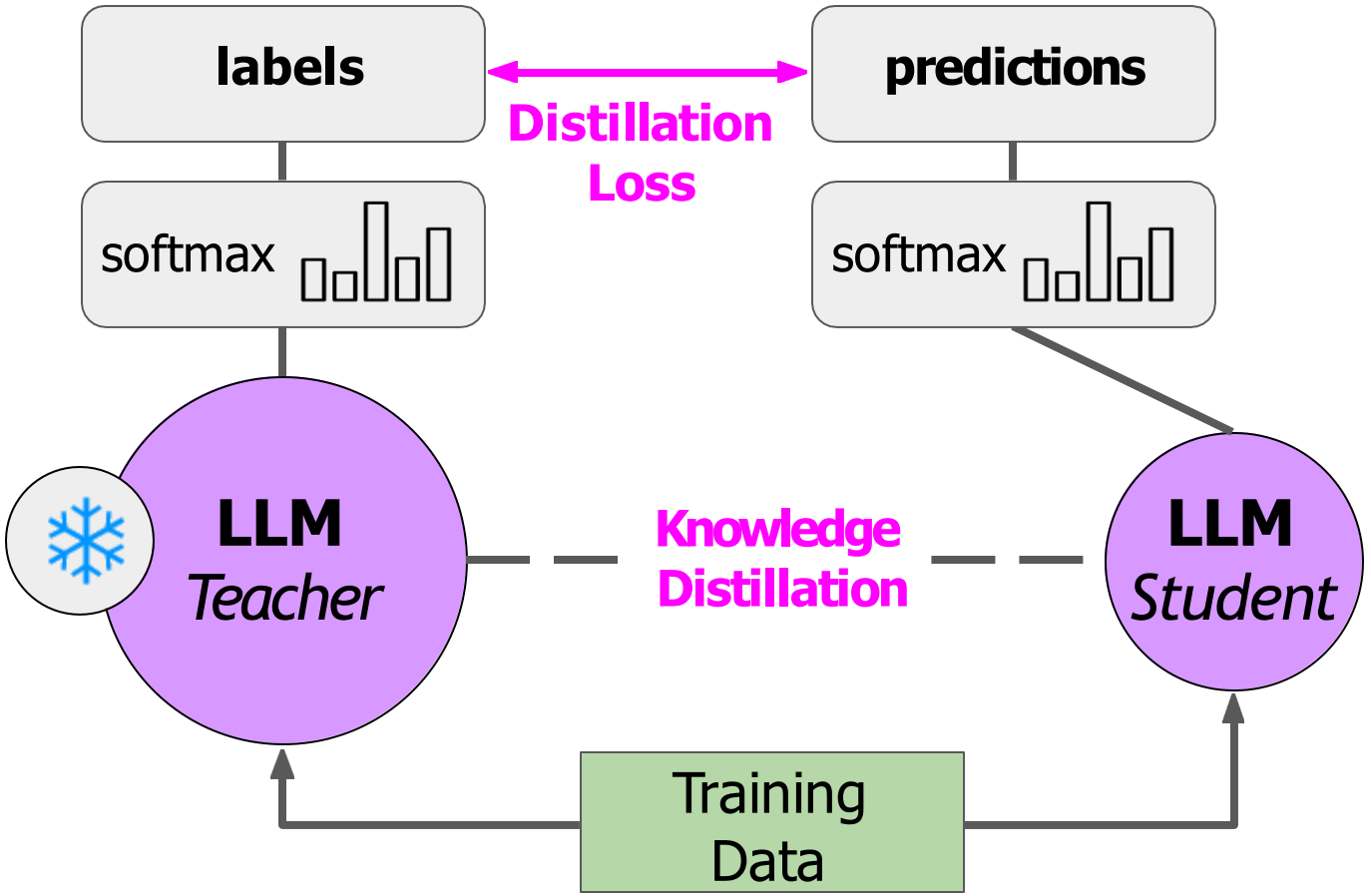


Pruning



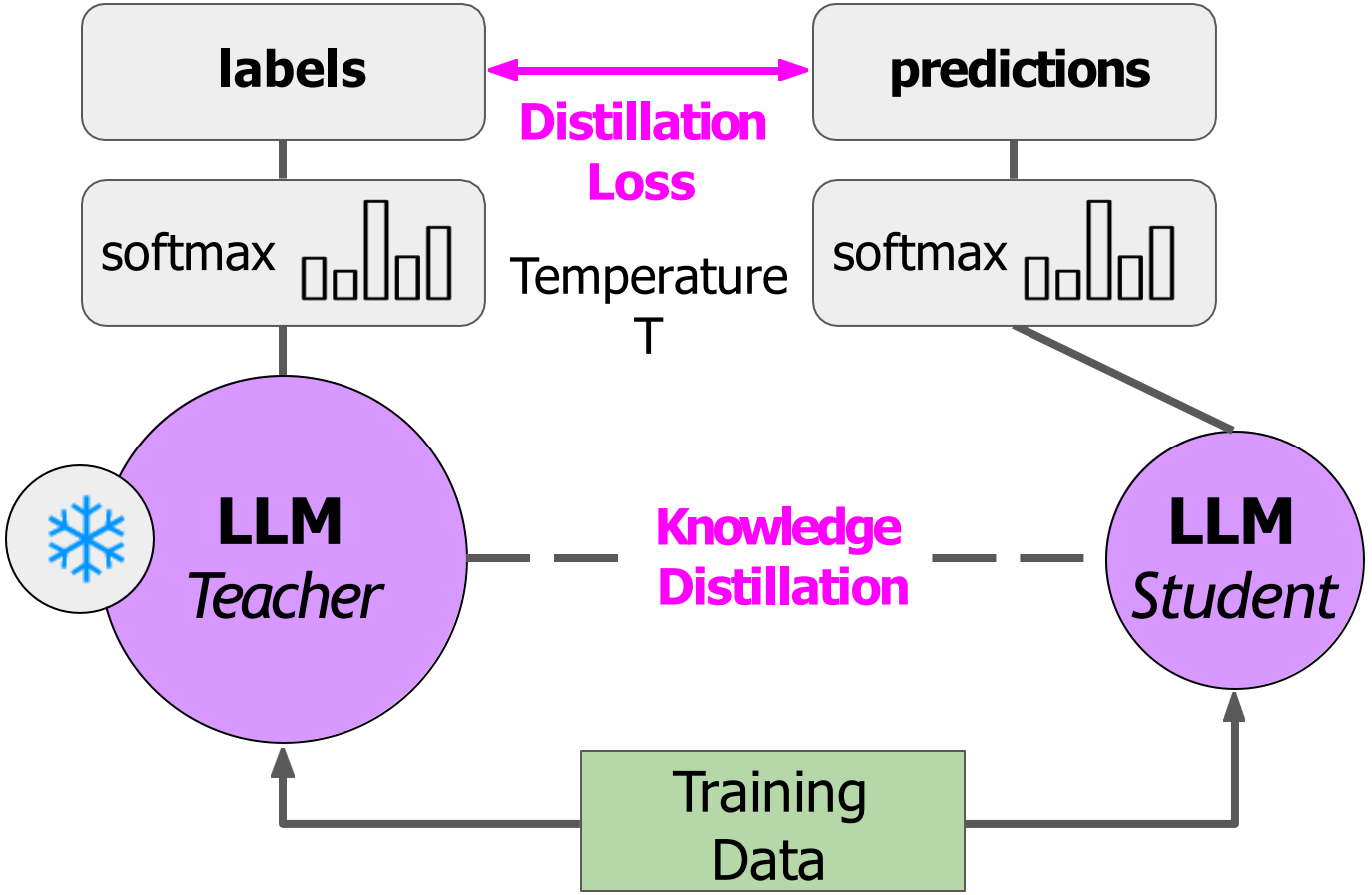
Distillation

Train a smaller student model from a larger teacher model



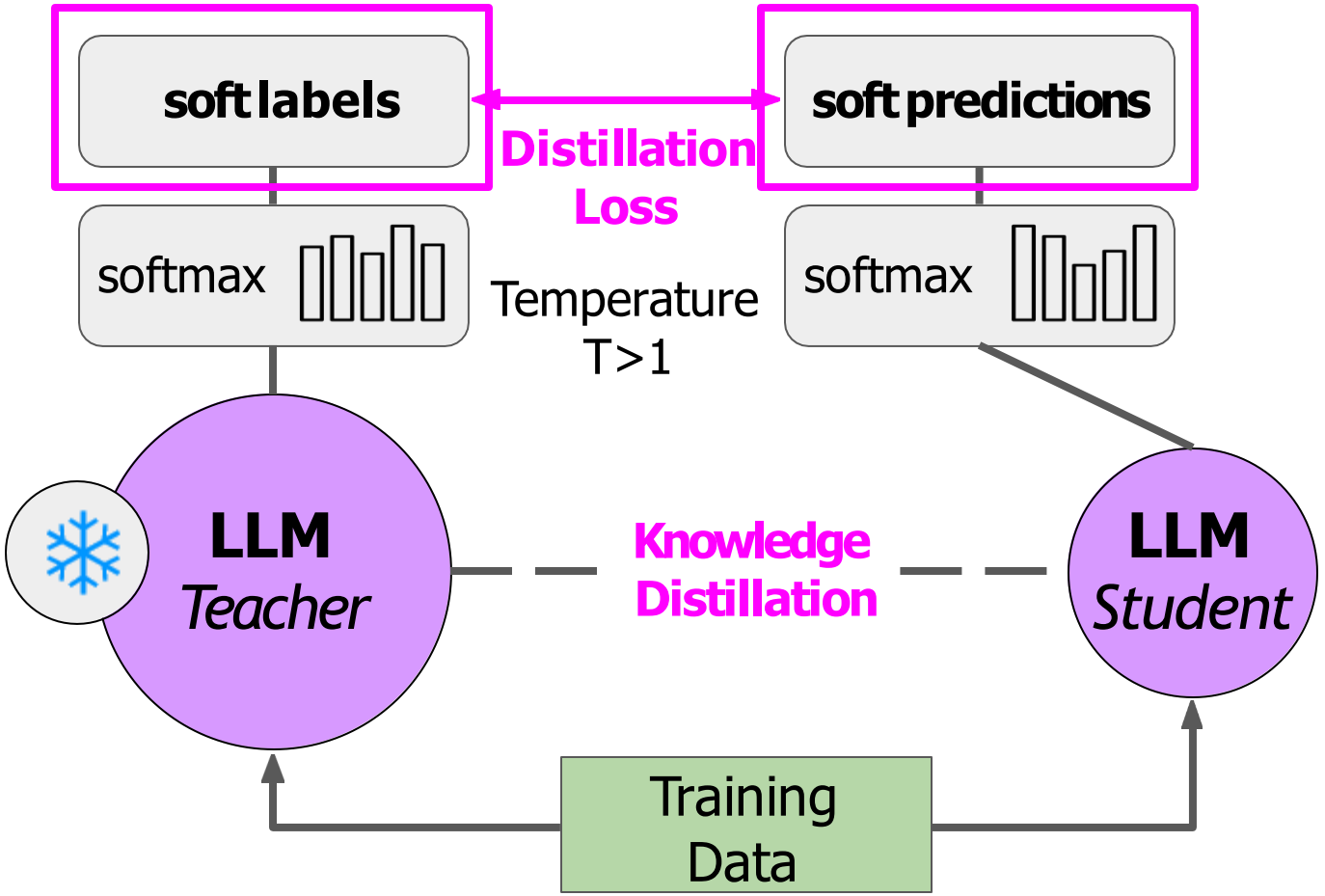
Distillation

Train a smaller student model from a larger teacher model



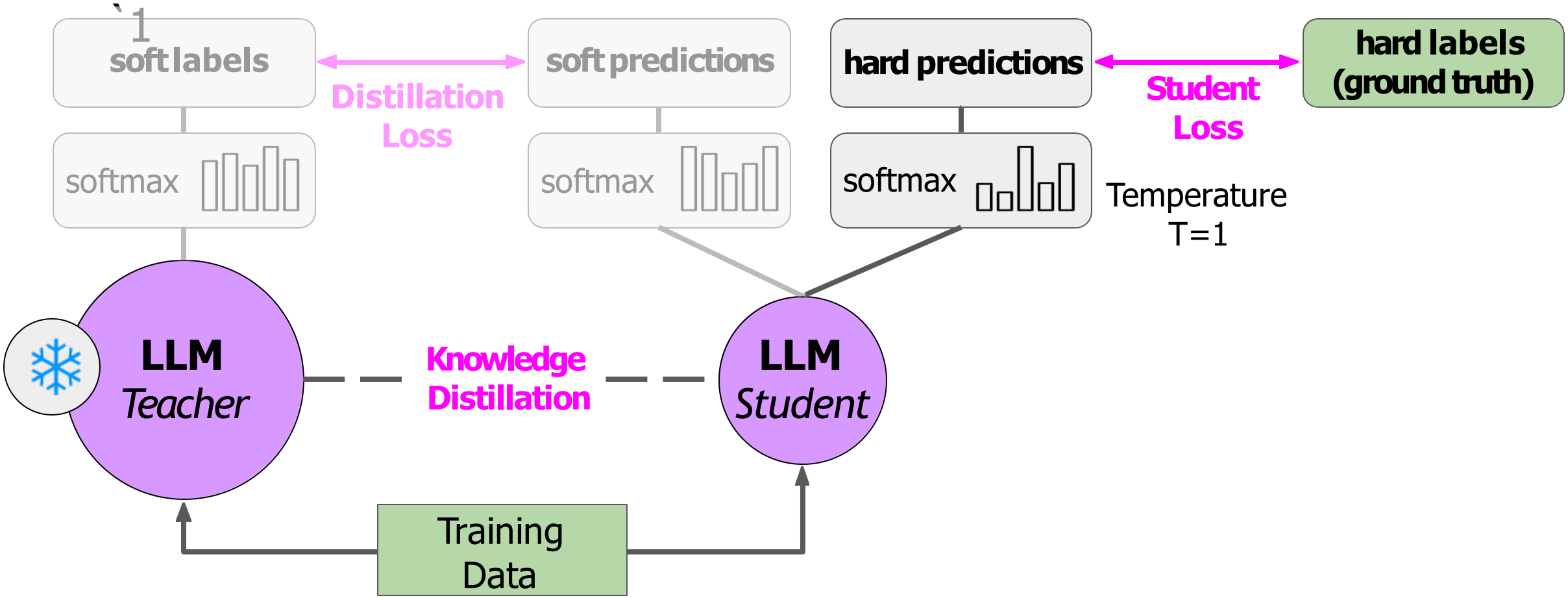
Distillation

Train a smaller student model from a larger teacher model



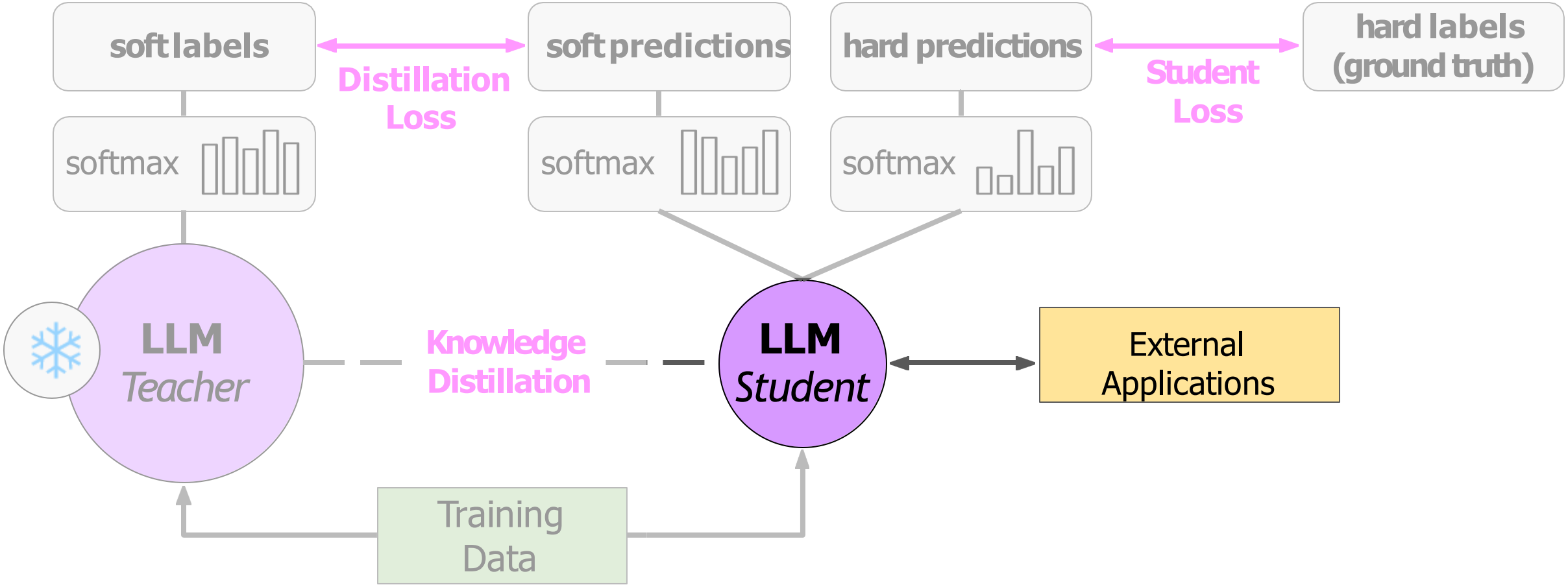
Distillation

Train a smaller student model from a larger teacher model



Distillation

Train a smaller student model from a larger teacher model



Backprogation

Backprogation update weights of LLM Students