

Prompt Adaptation Methods: Cost & VRAM Cheat Sheet

Method	Trainable Params	VRAM (T4 16GB)	1 Epoch Time (SST-2)
Soft Prompt (20 tokens)	≈15k	~1.5GB	~1min
Prefix Tuning (100 tokens)	≈75k	~2GB	~3min
LoRA r=8 (GPT-2)	≈4.5M	~4GB	~8min
Full Fine-Tune GPT-2	124M	~8GB	~25min

**Approx. wall-clock on free Colab T4, batch=4, fp16 when possible.
Use as rule-of-thumb to pick the quickest viable technique.*