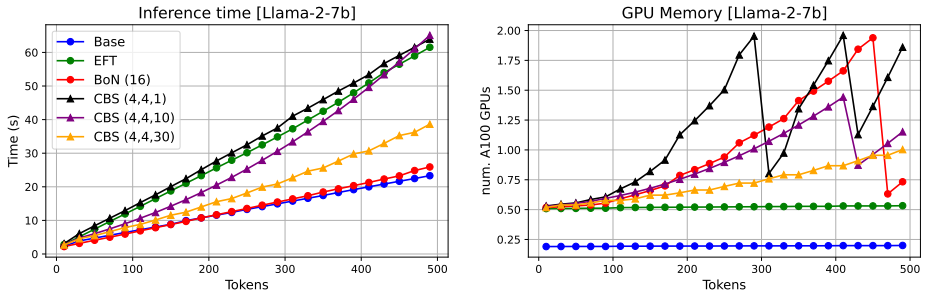
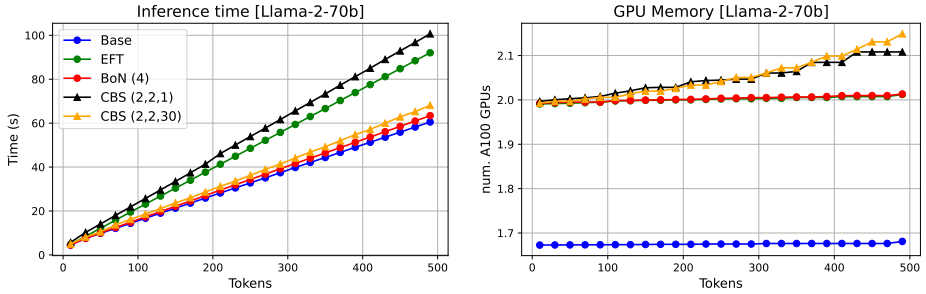
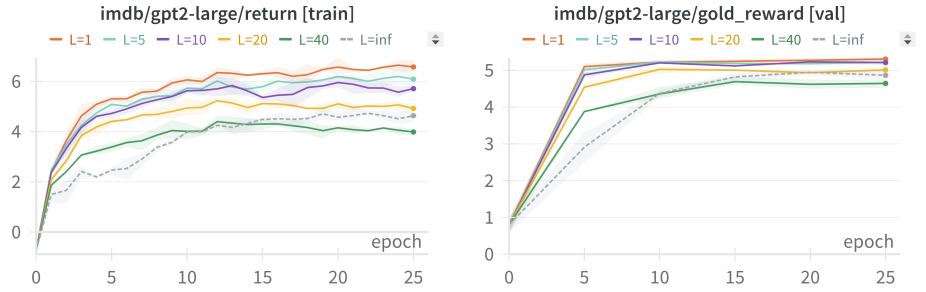


(a) Benchmarking results for guiding `Llama-2-7b`.

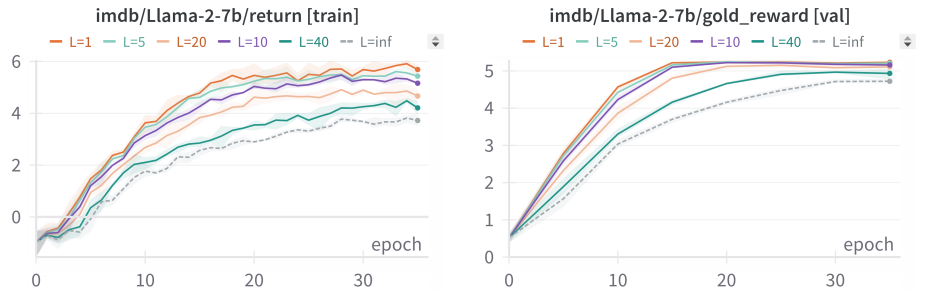


(b) Benchmarking results for guiding `Llama-2-70b`.

Figure 1: **Inference time and GPU memory usage for using 7B models to guide different base models.** Our proposed CBS, with default hyperparameters $(4, 4, 30)$ and $(2, 2, 30)$, is comparable to BoN sampling in both inference time and memory usage but performs better across multiple benchmarks (see Section 5 for details). We use `transformers` and `flash-attn` for benchmarking.



(a) PPO fine-tuning results for `gpt2-large`.



(b) PPO fine-tuning results for `Llama-2-7b` w/ LoRA ($r = 128$, $\alpha = 128$).

Figure 2: **PPO fine-tuning with chunk-level dense rewards for controlled-sentiment generation**. Using dense rewards parameterized by tuned and untuned `gpt2` models (see Section 5 for details), we train larger base models with PPO. The chunk length L controls the reward sparsity. For example, $L = 5$ means rewards are accumulated and emitted every 5 tokens (delayed to the last token of each chunk) while $L = \text{inf}$ corresponds to vanilla PPO with sequence-level sparse rewards. **Denser rewards facilitate credit assignment and accelerate training, improving both the achieved return on training prompts (accumulated dense rewards over the complete responses) (left) and the achieved gold reward on validation prompts (right). These results agree with the chunk length ablations of CBS in Figure 5(a) of the main paper.** We plot mean $\pm$ std over three random seeds.