Implementation Details

- Codebase references
    - Model: GPT-2, LoRA adapters on c_attn, c_proj, c_fc
    - Quantization scheme: QAT-style fake quantization with EMA observers, per-channel weight quant (row-wise) and per-tensor activation quant, switchable per-layer bit-widths at runtime, on-the-fly calibration for eval
    - Downstream dataset: SQuAD v1.1(HF datasets). I evaluate primarily with perplexity(PPL) for stability. EM/F1 in a generative framing is also reported qualitatively.

Results & Analysis

1. Task accuracy under different bit-widths
    a. Gpt-2 fine-tuned briefly (1k steps) and used as generative QA (not span extraction) often yields very low EM/F1 because the model's free-form answers rarely match gold strings exactly. Perplexity is a much more stable, sensitive proxy for quality changes due to quantization.
    b. Final PPL
        i. Evaluated on SQuAD-dev (1k examples), CPU, with per-profile calibration (scripts/03_eval_ppl.py)

Profile	Bits (attn_qkv / mlp_in)	PPL
uniform8	8 / 8	29.71-29.86
mixed8648	6 / 4	32.25-32.27

Gap: mixed8648 is ~ +8-9% PPL vs uniform 8 after proper calibration. (Earlier, without calibration and without true per-channel weight quant, mixed collapsed to very high PPL; fixing per-channel quant and a short calibration pass per profile resolved it.)

    c. EM/F1
        i. Using the GPT-2 model with a simple generative prompt and short max_new_tokens, EM/F1 was near zero on 100-200 samples for both profiles. This is expected since GPT-2 is not span-supervised for extractive QA, so exact string matches are rare. These low EM/F1 scores do not contradict the PPL comparison, which is a more reliable metric in this setting.

2. How I picked an optimal bit-width and insights
    a. Uniform8 gives best accuracy (lowest PPL).
    b. Mixed8648 (6b attention, 4b MLP in) is close (~+8-9% PPL) and offers a compute/precision saving, especially in the large MLPs, so it's a reasonable Pareto point for resource-constrained inference.

c. Calibration matters. If you switch to weight quantization at eval time, run a short calibration pass (dozens of batches) so observers see real range, otherwise both activations and weights can collapse.
d. Per-channel weight quant (row-wise) is crucial. With per-tensor ranges at 4-6 bits, large weight matrices lose signal. Per-channel quant preserves it.
e. MLP tolerates lower precision better than attention, but 4-bit MLP still imposes a small accuracy tax. 6-bit MLP is a safer default when quality is critical.
f. Trade-off
    i. If accuracy is top priority, uniform8
    ii. If you want a leaner setting with minimal loss, try attn 6b, MLP 6b, which should narrow the PPL gap further while still saving bits vs. 8/8.

3. Objectives that could better support switching bit-widths
    a. To make switching precision a first-class capability during training, the loss function can explicitly encourage consistency across profiles. For example, minimizing KL/JS divergence between token distributions of 8-bit and mixed profiles on the same inputs would align their predictions and reduce switching-induced instability. This can be extended into a multi-teacher distillation setup across several bit-width profiles. Training can also incorporate random sampling of layer-wise bit assignments at each step or micro-batch so that the model naturally learns robustness to precision changes. Adding a small stability regularizer on intermediate activations can further smooth transitions. Another promising idea is to learn lightweight gating mechanisms that dynamically select higher or lower precision for certain layers or tokens, while including a bit-budget penalty in the loss to balance accuracy and efficiency. Finally, penalizing excessive variance in learned or EMA-calibrated activation ranges across profiles could make quantization ranges more consistent and reduce abrupt distribution shifts when bits are switched.

4. Does my observation align with CPT (ICLR'21)?
    a. Our results are broadly consistent with CPT's hypothesis that low precision acts as exploration noise and higher precision enables fine-grained convergence. During cyclic precision training, we observed that alternating between uniform8 and mixed8648 led to comparable or lower final perplexities versus static profiles, suggesting convergence toward flatter minima as CPT predicts. But, our EM/F1 remained near zero for GPT-2, likely because the generative evaluation protocol was not tuned for span-extraction. This limits direct accuracy comparison but does not contradict CPT's generalization benefit claims.
    b. Potential reasons:
        i. Vision vs. language gap (attention/casual-LM loss may respond differently to quant noise"
        ii. limited training budget (300-1k steps may be too short)
        iii. LoRA-only finetuning (base weights frozen) and limited profile diversity (only 8/8 and 6/4)

5. Does my observation align with Double-Win-Quant (ICML'21)?
    a. Under character-level perturbations we observed:
        i.

Setting	Clean PPL	Adv PPL	Δ vs. Clean
uniform8	24.79	42.22	+70.3%
mixed8648	26.82	45.67	+70.3%
random precision (batch-level)	25.58	43.82	+71.3%

Our robustness evaluation showed that both uniform8 and mixed8648 experienced ~70% perplexity increase under adversarially perturbed inputs, nearly identical degradation, indicating that switchable quantization did not inherently mitigate adversarial vulnerability. This partially diverges from Double-Win Quant, which attributes improved robustness to random precision inference/training that disrupts adversarial transferability. We did not implement random precision sampling during inference, so our results do not capture the "win-win" effect. Future work could incorporate random precision or switchable batch-norm, as suggested by DWQ, to explore whether quantization noise can shield perturbations for LLMs.

    b. Potential reasons
        i. Our randomization was per-batch between two profiles
        ii. DWQ's defense effect likely needs finer per-layer/per-token randomization and more than two profiles.
        iii. Threat model differs (character-level text vs. gradient-based image attacks)
        iv. GPT-2 with brief fine tuning may not exhibit DWQ's effect. Deterministic calibration may not optimally support stochastic inference.

6. Promising research directions
    a. Our experiments highlight several natural next steps for integrating switchable and dynamic quantization with LLMs. One promising direction is to combine cyclic precision training with random precision switching at inference, using CPT to encourage convergence toward flatter minima while leveraging stochastic precision at runtime to improve robustness. Another one is to explore finer-grained schedules, such as per-payer or per-token bit-width switching, rather than alternating between only two global profiles. Extending beyond LoRA adapters to full-model fine tuning could also allow CPT's benefits to fully manifest. Also, developing calibration techniques that remain consistent under stochastic precision switching could mitigate performance degradation when profiles change dynamically. Finally, scaling these experiments to larger GPT-2

variants and evaluating them under gradient-based adversarial attacks like FGSM, PGD would provide a more complete test of whether dynamic quantization can enhance both generalization and adversarial resilience in LLMs.