A key distinction of GSPO compared to GRPO is its practice of clipping entire responses rather than individual tokens. Particularly, as shown in Figure 2, we observe a difference of two orders of magnitude in the fractions of clipped tokens between GSPO and GRPO (while adjusting the clipping ranges does not alter the disparity in magnitude). However, despite clipping significantly more tokens and consequently using fewer for training (or gradient estimation), GSPO still achieves higher training efficiency than GRPO. This counter-intuitive finding — that clipping a much larger fraction of tokens leads to superior training efficiency — further indicates that GRPO's token-level gradient estimates are inherently noisy and inefficient for sample exploitation. In contrast, GSPO's sequence-level approach provides a more reliable and effective learning signal.

Compared to the RL training of dense models, the sparse activation nature of MoE models introduces unique stability challenges. In particular, we found that when adopting the GRPO algorithm, the expert-activation volatility of MoE models can prevent RL training from converging properly. To be

specific, after one or more gradient updates, the experts activated for the same response can change significantly. For example, with the 48-layer Qwen3-30B-A3B-Base model, after each RL gradient update
and for the same rollout sample, there are roughly 10% of the experts activated under the new policy.