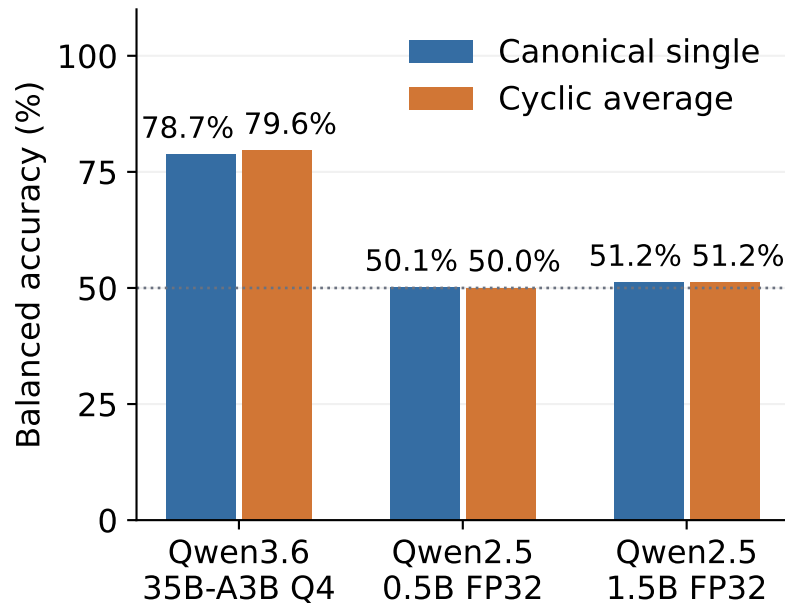
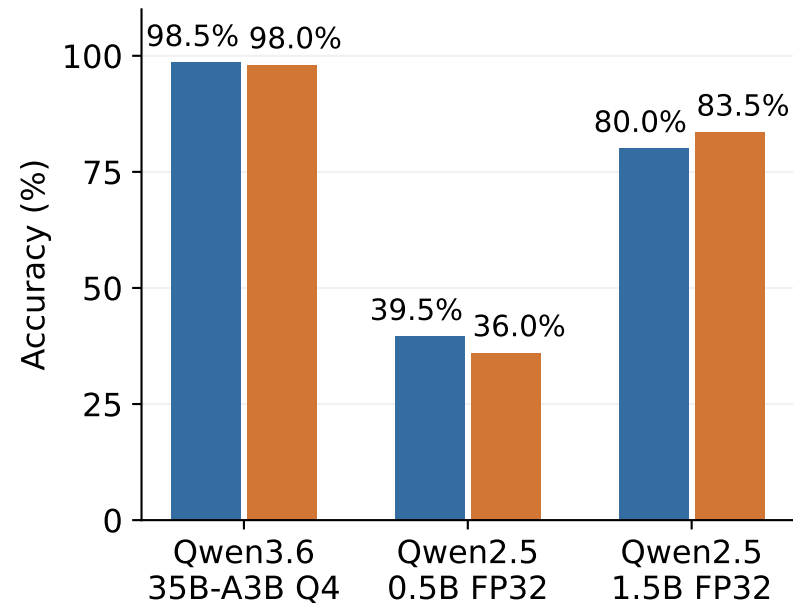


Quality must accompany stability

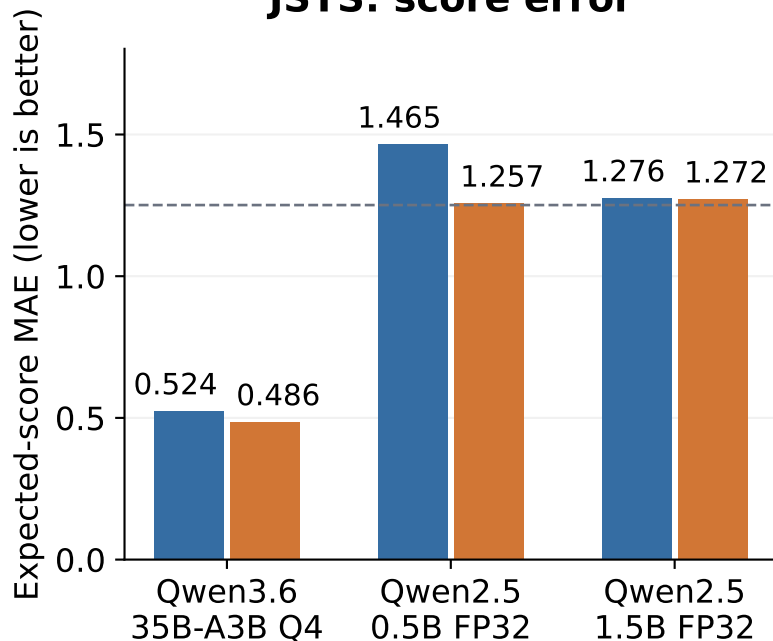
JCoLA: class discrimination



JCQA: answer quality



JSTS: score error



Same 200 items per task.
Cyclic answers require 2 / 5 / 6 calls.

JSTS dashed line: always predict 2.5.
Added post hoc after 0.5B collapse.

35B reversed single: MAE 0.475.
Cyclic averaging is not best among
all measured display orders.

JCoLA source groups partly overlap
earlier evaluations. Paired intervals
and class marginals are in the tables.