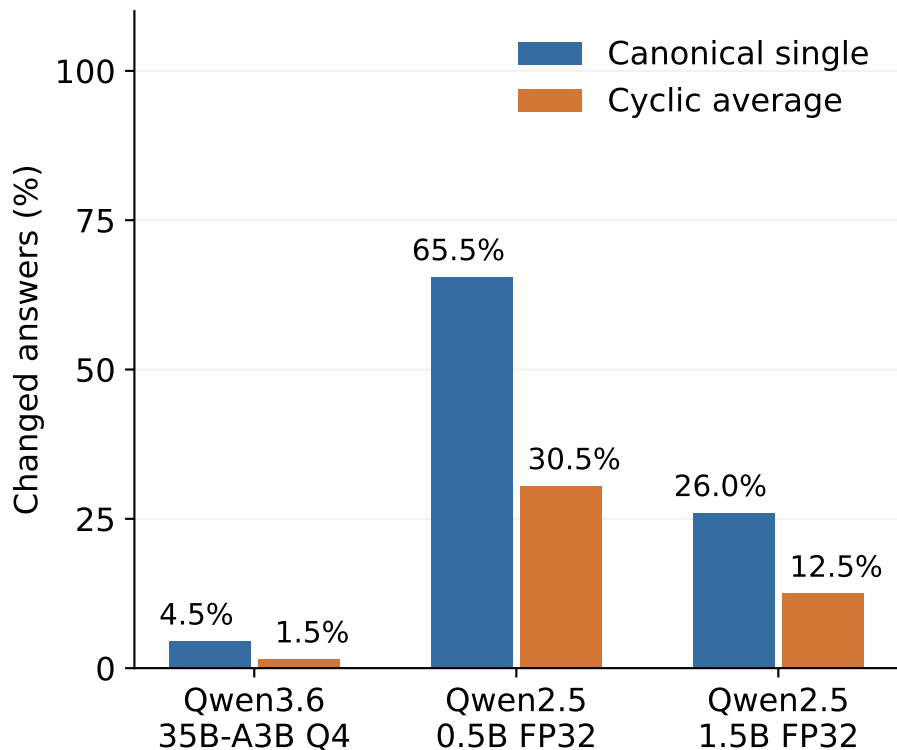
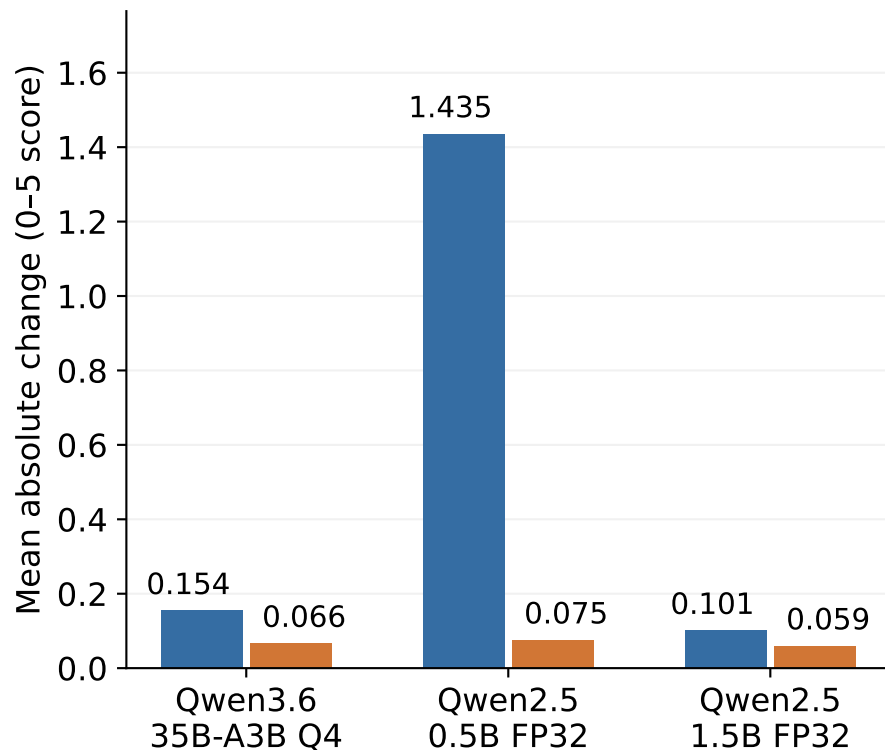


Display-order sensitivity on 600 new evaluation items

JCQA: answer changes after reversal



JSTS: change in expected score



200 items/task. Lower is less sensitivity. Binary JCoLA is excluded: its shared-pool invariance is structural. Cyclic answers use K calls (JCQA 5; JSTS 6). Models also differ in family, precision and backend.