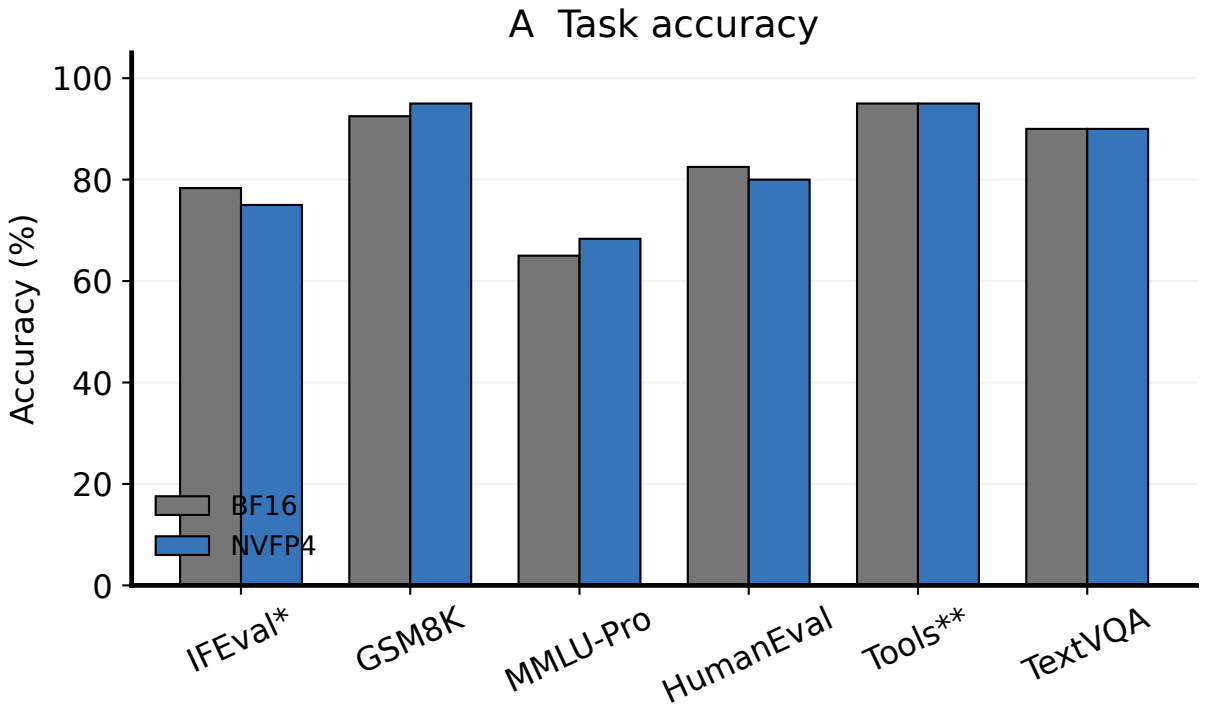
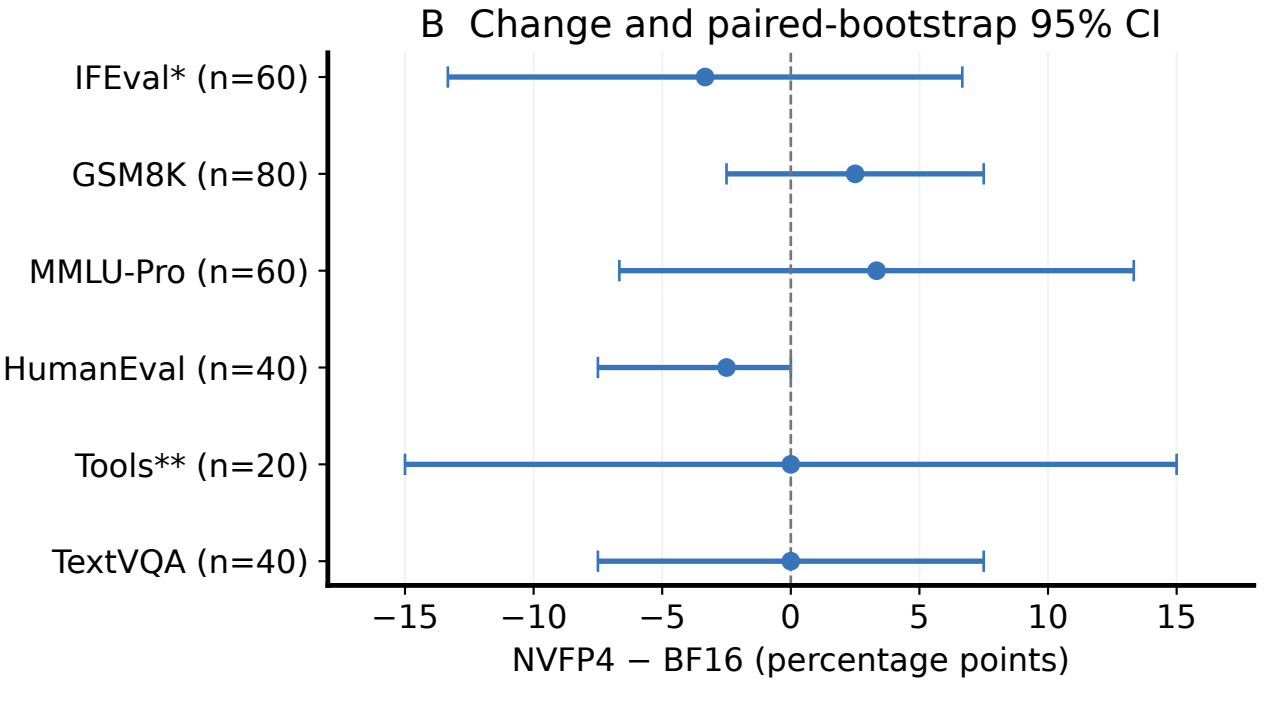
Quality after NVFP4 quantization — small regression suite

300 fixed items • thinking off • paired comparison to BF16 • not full-benchmark scores





Pooled accuracy: BF16 82.7% → NVFP4 83.0% • Δ +0.3 pp [−3.0, +4.0]

*IFEval subset uses a re-implemented checker. **Tools: 20 hand-written BFCL-style cases.

Subsets of 20–80 items/task; uncertainty is wide. No claim of losslessness or intelligence improvement.

Source: evaluation/compare_vs_ref.json; see model card for truncations and methodology.