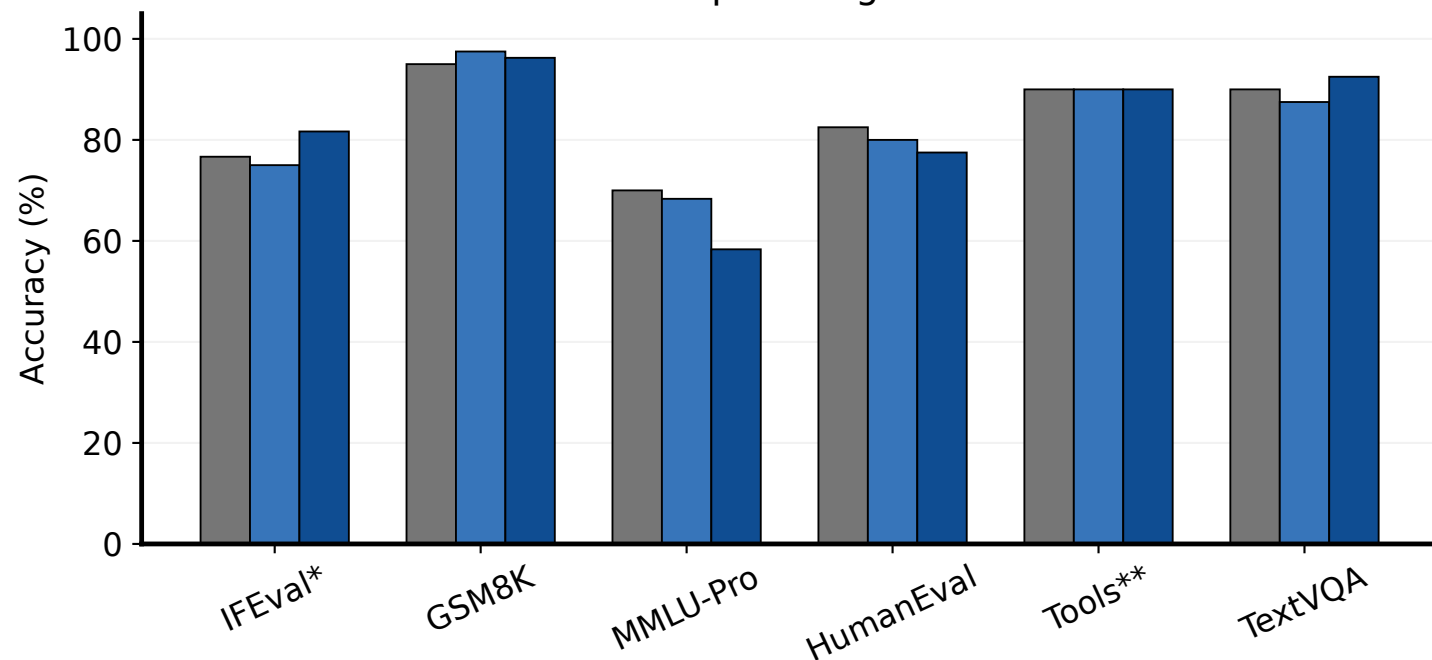


Runtime v2 quality — preserved evidence, explicit trade-offs

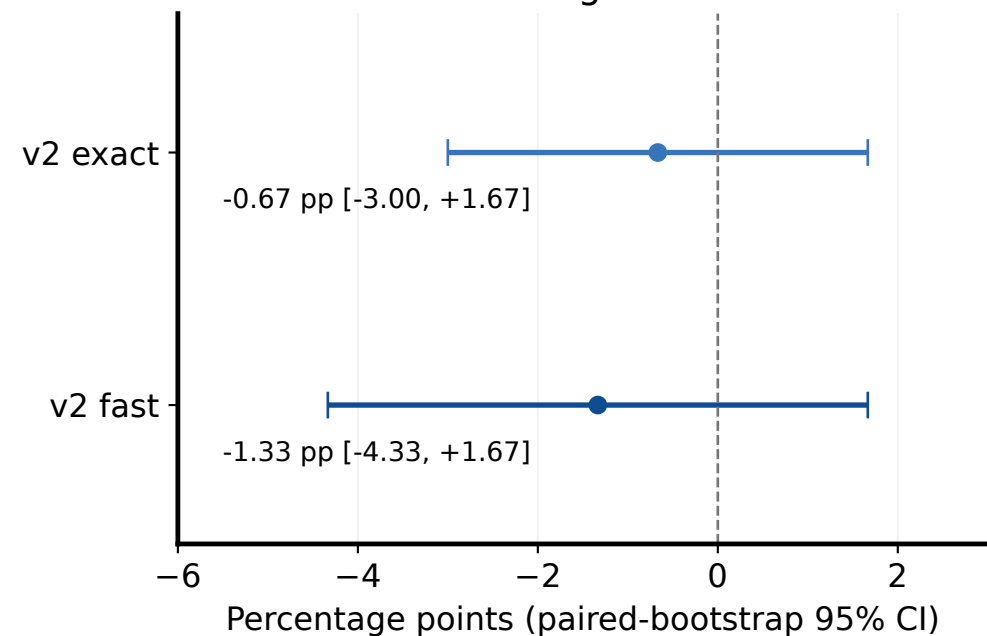
Same 300-item suite • thinking off • greedy • evaluation concurrency 16

Frozen v1 + MTP v2 exact + MTP v2 fast + MTP

A First complete regression run



B Pooled change vs frozen v1



Frozen: 251/300 (83.7%) • exact: 249/300 (83.0%) • fast: 247/300 (82.3%)

Fast MMLU-Pro first run: 58.3% vs 70.0% frozen (−11.7 pp); fast reruns: 71.7%, 65.0%.

Fast truncations: 20 vs 14 frozen. Fast fails original first-run per-task/stability tripwires; not “lossless.”

*Re-implemented IFEval checker. **20 hand-written tool cases. Small subsets, not full-benchmark results.

Source: runtime-v2/runs/final_*/eval; all reruns retained in runtime-v2/runs/var and RESULTS.md.