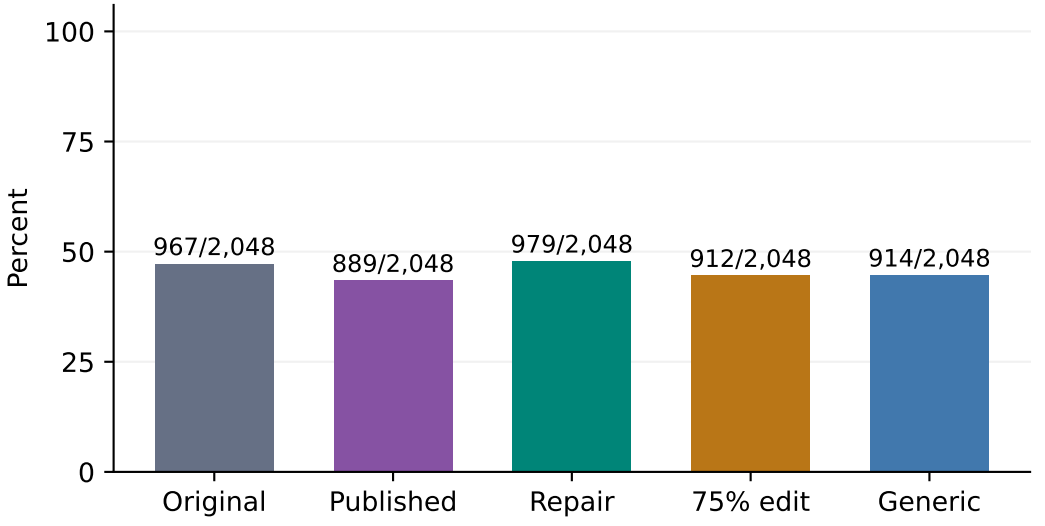
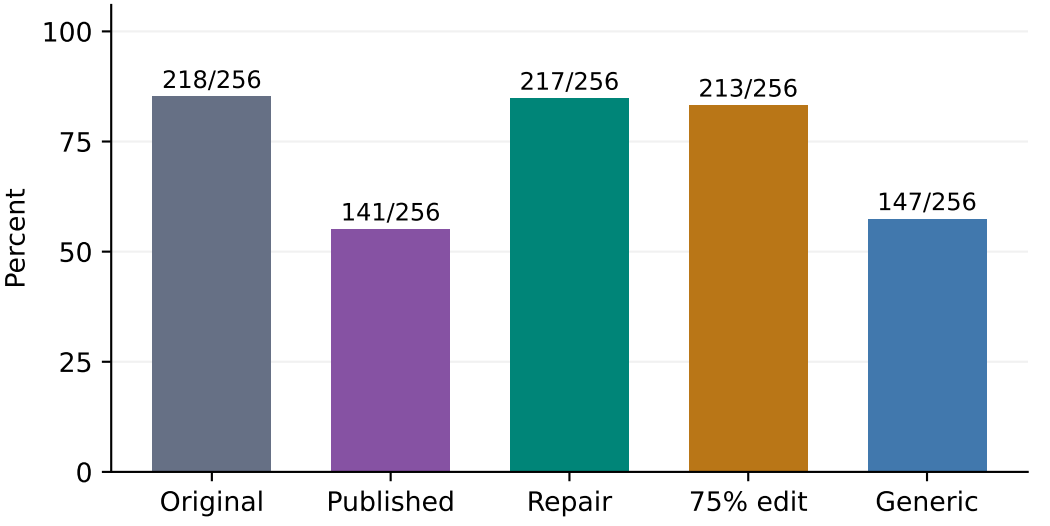
A frozen correction helps on selected tasks, with trade-offs

Related-task screen



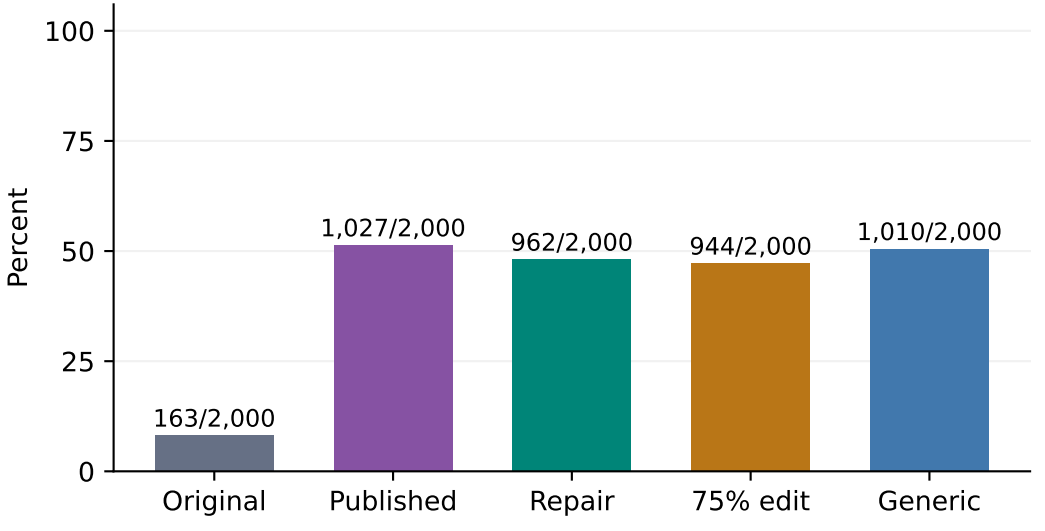
Non-thinking; new non-lookup operators missed the competence gate.

Thinking-mode copying



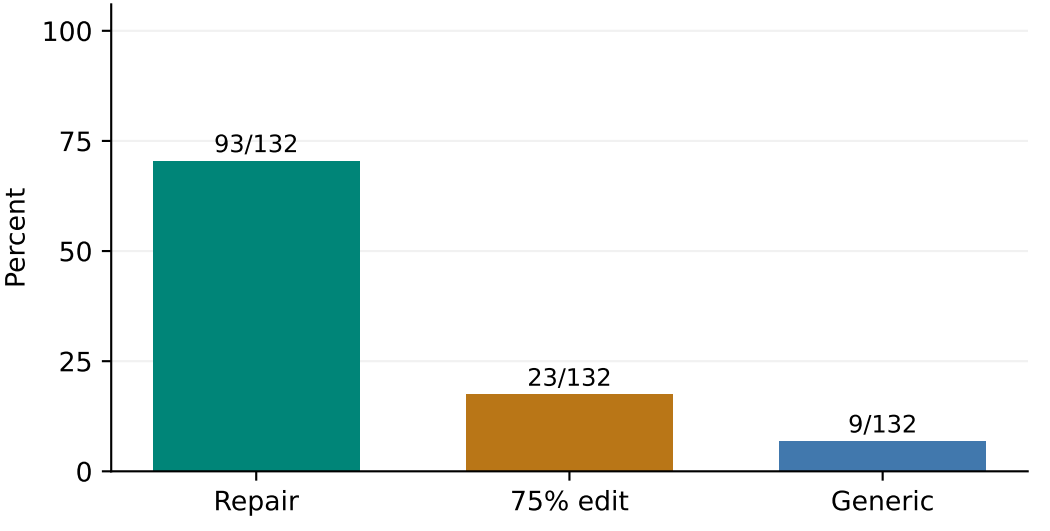
64 strings × 4 seeds; original 218/256 < required 240/256.

Full MATH-500



500 questions × 4 samples; no equivalence/noninferiority claim.

Consistently repaired screen failures



Fixed set of 132 repeat-confirmed losses; selected workload.

Point estimates only. Counts and claims are recomputed from the repository's prompt-level records; see CLAIMS.md.