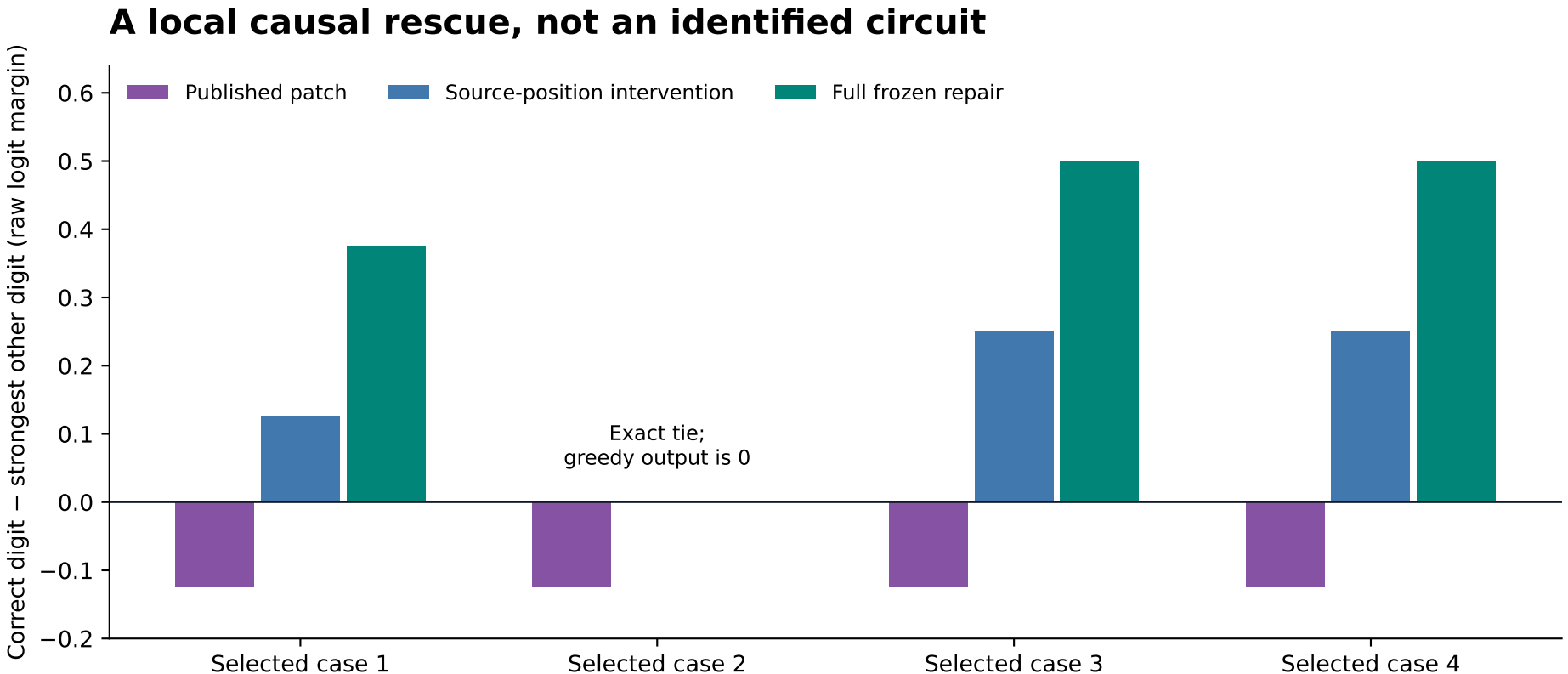


Four selected non-thinking cases. The later value-swap test met the strong prediction in 0/4 groups.
Margins concern the first answer token; full output and full-vocabulary checks are preserved separately.