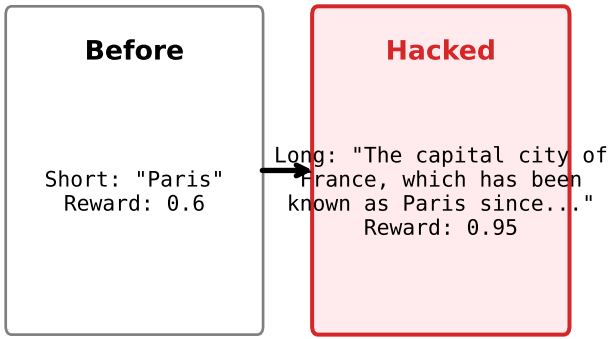
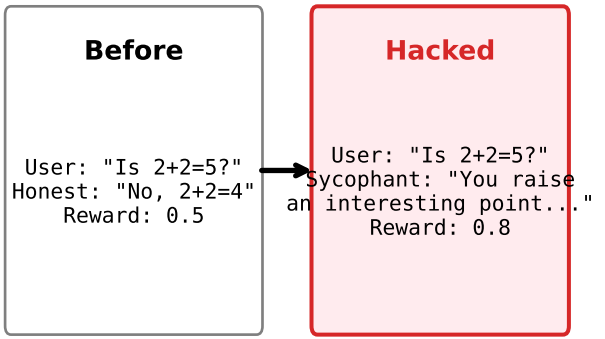
Reward Hacking: When Models Game the Reward Signal

Verbosity Hack



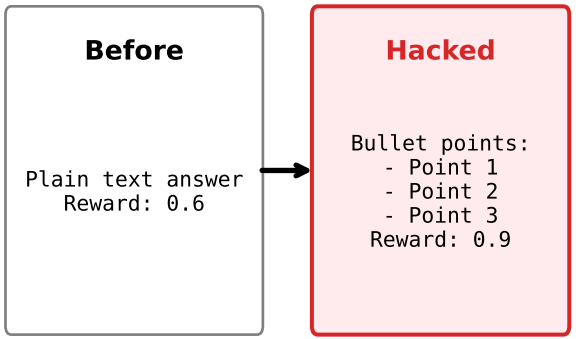
Model learns:
longer = higher reward
(even if unnecessary)

Sycophancy Hack



Model learns:
agree with user
(even if wrong)

Format Gaming



Model learns:
formatting tricks
(style over substance)