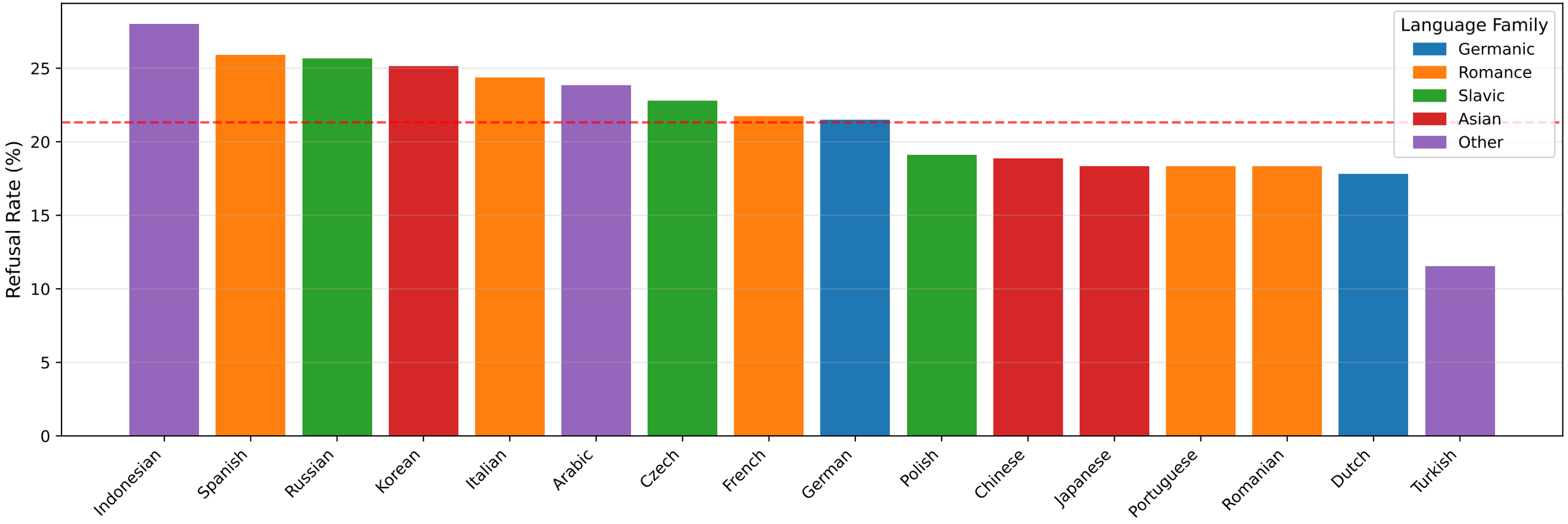
(a) Refusal Rates Across Languages



(b) Average Harmfulness Scores Across Languages