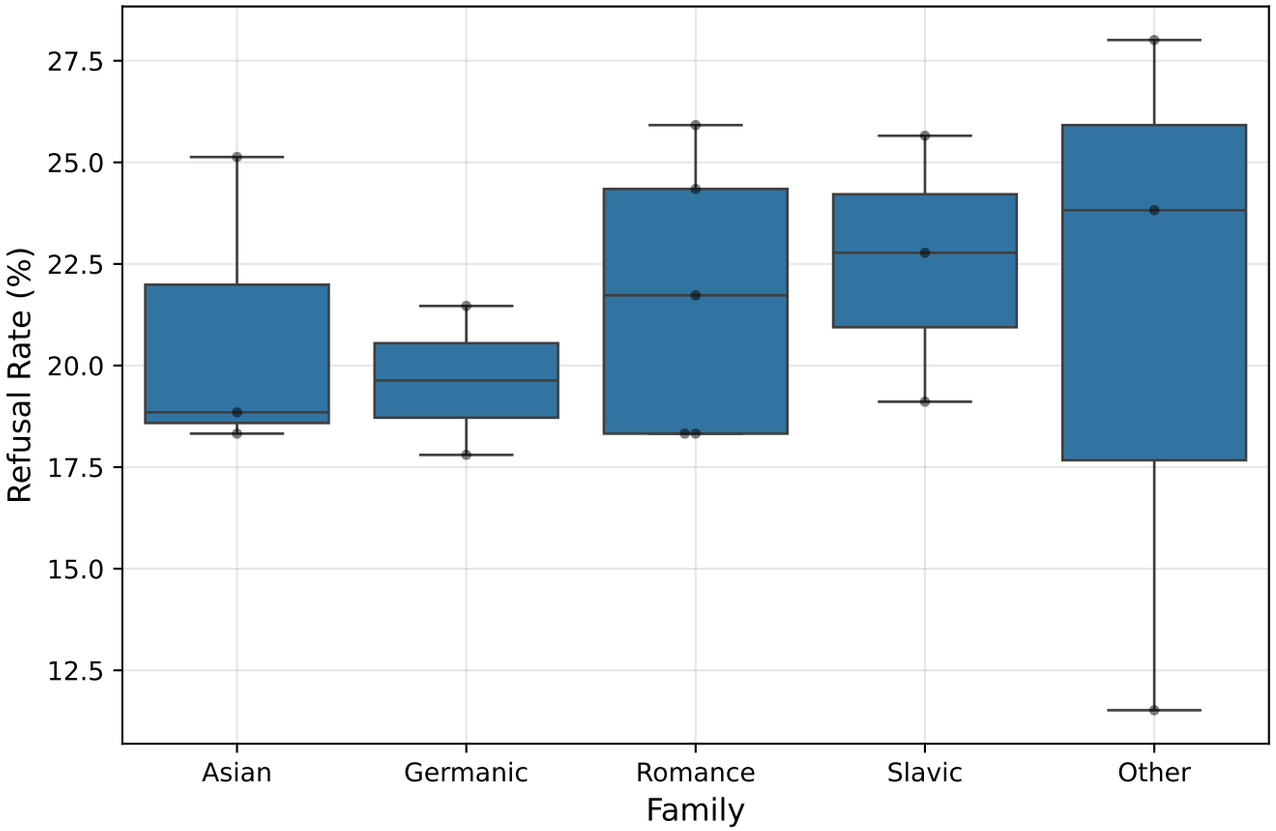
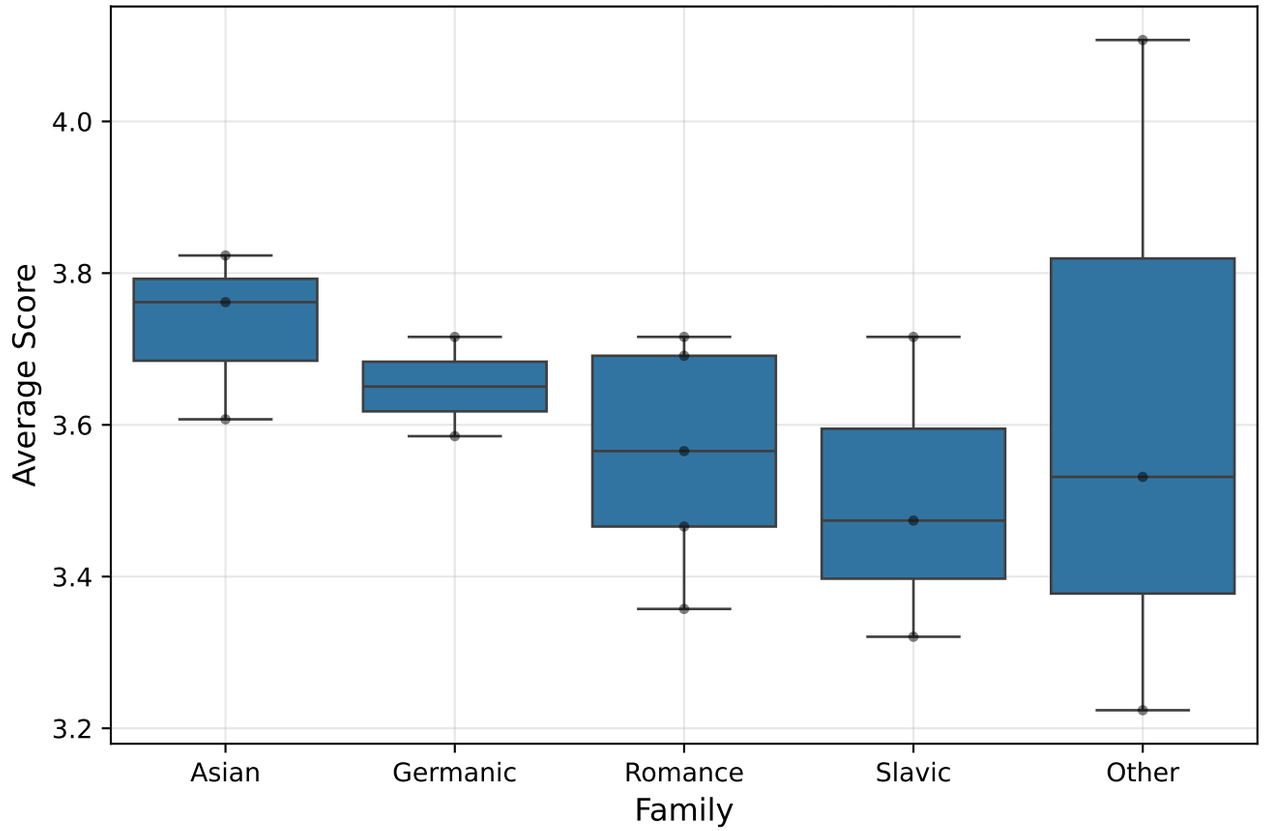
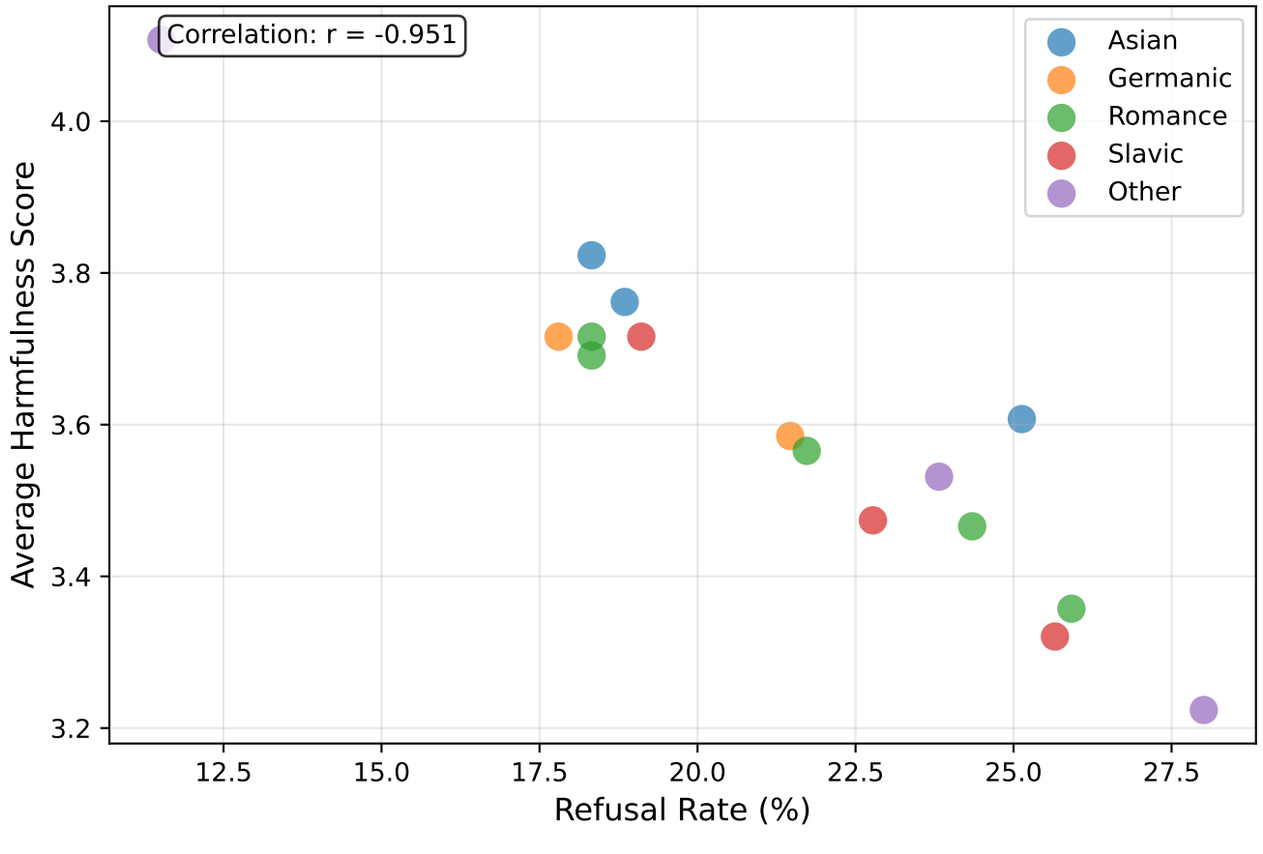
(a) Refusal Rate by Language Family



(b) Harmfulness Score by Language Family



(c) Safety-Harmfulness Trade-off



(d) Average Performance by Family