

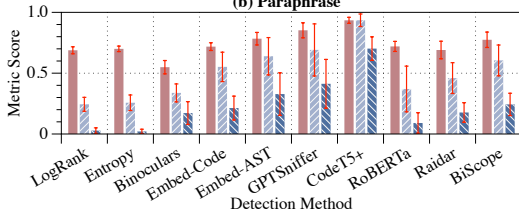
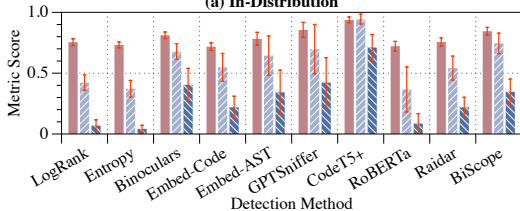
F1 Score

TPR@FPR=10%

TPR@FPR=1%

(a) In-Distribution

(b) Paraphrase



(c) Cross-Model

(d) Cross-Model Paraphrase

