1 Significance Test for RQ1

Table 1: Code summarization performance of GPT-4o on codes with different readability.

Group	Readability	BLEU	BERTScore
low-readability	3.44	6.32	18.25
high-readability	4.28	**8.12**	**19.78**
p-value*	<0.0001	<0.01	<0.04(0.0383)

**p-value is calculated with pairwise 2-sample Wilcoxon Signed rank test between the two groups.*

The results in Table 1 show that the discrepancy between the two groups are significant, with all p-values <0.04.

2 Significance Test for RQ3

Table 2: BLEU scores on the cross-obfuscated datasets. The semantic perturbation chooses the one with the greatest impact on the model in each programming language. The values in italic indicate nonsignificant decrease (p>0.05)

Dataset	CodeBERT			CodeT5			CodeLlama		
	Python	**Go**	**Java**	**Python**	**Go**	**Java**	**Python**	**Go**	**Java**
Semantic Perturb.	**IHR**	**IOE**	**IHR**	**IOE**	**FNE**	**IOE**	**IOE**	**IHR**	**IHR**
primary	12.54	11.02	12.84	15.23	16.19	14.18	11.15	11.85	12.63
Cross Perturb.									
Semantic × OOS	*12.52*	*10.93*	*12.87*	14.21	*16.25*	*14.34*	*11.07*	*11.64*	*12.50*
Semantic × HVI	11.57	*10.81*	12.16	13.75	*15.75*	13.58	8.85	*11.82*	*12.51*
Semantic × DBI	**11.49**	*10.74*	**12.13**	**12.70**	**14.58**	**13.31**	**8.15**	*11.41*	**12.07**
average	11.83	10.83	12.39	13.55	15.53	13.74	9.36	11.62	12.36

According to the results in Table 2, most cross-perturbations exhibit a significant descrease of BLEU score (p<0.0001). We particularly notice that Semantic×OOS exhibits nonsignificant decrease (p>0.05). The results also align with our current conclusion: oprand swap does not affect the readability significantly whereas Semantic×DBI is the most significant way for structural obfuscation.

3 Implementation algorithms of perturbation methods

Algorithm 1 IOE perturbation algorithm

Output: AST parsed by tools $Code_{AST}$
Input: The new code string after perturbation$Code_{new}$
$Identifier\ List \leftarrow Code_{AST}.Walk()$
$rate \leftarrow 0.8$
$Identifier\ List \leftarrow List_{Identifier}.Sample(rate)$
for $i, ident \in List_{Identifier}$ **do**
$ident.Name \leftarrow' v'_i$
end for
$Code_{new} \leftarrow Code_{AST}.ReGetString$
return $Code_{new}$

Algorithm 2 IS perturbation algorithm

Output: AST parsed by tools $Code_{AST}$
Input: The new code string after perturbation$Code_{new}$
$Identifier\ List \leftarrow Code_{AST}.Walk()$
$Copy\ List \leftarrow List_{Identifier}[1:] + List_{Identifier}[0]$
for $i, ident \in List_{Identifier}$ **do**
$ident.Name \leftarrow List_{Copy}[i].Name$
end for
$Code_{new} \leftarrow Code_{AST}.ReGetString$
return $Code_{new}$

Algorithm 3 IHR perturbation algorithm

Output: AST parsed by tools $Code_{AST}$
Input: The new code string after perturbation$Code_{new}$
$Identifier\ List \leftarrow Code_{AST}.Walk()$
$Identifier\ List \leftarrow List_{Identifier}.Shuffle()$
for $i, ident \in List_{Identifier}$ **do**
$ident.Name \leftarrow List_{high\ frequence\ words}.Pop()$
end for
$Code_{new} \leftarrow Code_{AST}.ReGetString$
return $Code_{new}$

Algorithm 4 FNE perturbation algorithm

Output: AST parsed by tools $Code_{AST}$
Input: The new code string after perturbation$Code_{new}$
$FunctionDefinition \leftarrow Code_{AST}.Walk()$
$FunctionDefinition.Name \leftarrow' v'_0$
$Code_{new} \leftarrow Code_{AST}.ReGetString()$
return $Code_{new}$

Algorithm 5 OOS perturbation algorithm

Output: AST parsed by tools $Code_{AST}$
Input: The new code string after perturbation$Code_{new}$
$BinaryOperationsList \leftarrow Code_{AST}.Walk()$
for $oper\ node \in List_{Binary\ Operations}$ **do**
$oper\ node \leftarrow Swap\ Binary\ Operations(oper\ node)$
end for
$Code_{new} \leftarrow Code_{AST}.ReGetString()$
return $Code_{new}$

Algorithm 6 HVI perturbation algorithm

Output: AST parsed by tools $Code_{AST}$
Input: The new code string after perturbation$Code_{new}$
$BodyBlockNode \leftarrow Code_{AST}.Walk()$
$Insert\ Position(pos) \leftarrow Randomly\ Select\ From\ Node_{Body\ Block}$
$Sample\ Number \leftarrow 3$
for all $i < Sample\ Number$ **do**
$statement \leftarrow Randomly\ Generated\ Definition\ Statement\ for\ High-frequency\ Variable$
$Code_{new} \leftarrow Code_{new}.InsertStatement(pos, statement)$
$i \leftarrow i + 1$
end for
return $Code_{new}$

Algorithm 7 DBI perturbation algorithm

Output: AST parsed by tools $Code_{AST}$
Input: The new code string after perturbation$Code_{new}$
$BodyBlockNode(Node_{body}) \leftarrow Code_{AST}.Walk()$
$Dead\ Branch\ Node \leftarrow Randomly\ Select\ From\ List_{Branches}$
$Insert\ Node_{body}\ Into\ Live\ Branch\ of\ Dead\ Branch\ Node$
$Code_{AST}.SetNewBodyBlock(Block_{new})$
$Code_{new} \leftarrow Code_{AST}.ReGetString()$
return $Code_{new}$

4 P-value of single Pertub

Table 3: Evaluation results on the obfuscated datasets (BL=BLEU, BS=BERTScore).

Dataset	CodeBERT					
	Python		Go		Java	
	BL	BS	BL	BS	BL	BS
Primary						
	17.95	29.64	17.78	40.11	18.62	31.92
Semantic Perturb.						
IOE	13.89(0.0000)	19.52(0.0000)	**11.02 (0.0000)**	**16.26(0.0000)**	13.85(0.0000)	21.53 (0.0000)
IS	14.70(0.0000)	22.82(0.0000)	13.07(0.0000)	23.28(0.0000)	15.42(0.0000)	25.50 (0.0000)
IHR	**12.54 (0.0000)**	**17.53(0.0000)**	12.10(0.0000)	22.38(0.0000)	**12.84(0.0000)**	**19.61 (0.0000)**
FNE	14.74(0.0000)	22.63(0.0000)	12.95(0.0000)	21.48(0.0000)	15.40(0.0000)	25.17 (0.0000)
Syntactic Perturb.						
OOS	17.94(0.9632)	29.63(0.9511)	17.79(0.9685)	40.14(0.9259)	18.61(0.9782)	31.90 (0.9531)
HVI	17.53(0.0133)	28.69 (0.0001)	17.75(0.9023)	40.08(0.9393)	18.15(0.0436)	30.99 (0.0021)
DBI	17.34(0.0003)	28.27(0.0000)	17.87(0.7555)	40.40 (0.3955)	18.26(0.1269)	31.41 (0.0874)

4.1 Conclusion

4.1.1 Semantic Pertub

In semantic perturb, there are (model)*3(language)*4(perturb)*2(score)=72 sets of data, 5 of which do not meet the range of $pvalue < 0.04$ (marked in red), and these 5 sets all appear in CodeLlma's Go language tasks.

4.1.2 Synatic Pertub

In synatic perturb, there are 3 (model)*3(language)*3(perturb)*2(score)=54 sets of data, 11 of which do not meet the range of $pvalue > 0.04$(marked in pink), 9 of which have $pvalue \in [0.04, 0.5]$ (marked in blue).So overall, 20 out of 54 groups did not meet the $pvalue > 0.5$.

5 P-value of cross Perturb

Table 4: Evaluation results on the obfuscated datasets (BL=BLEU, BS=BERTScore).

Dataset	CodeT5					
	Python		Go		Java	
	BL	BS	BL	BS	BL	BS
Primary						
	20.38	34.41	19.67	43.18	20.66	35.35
Semantic Perturb.						
IOE	**15.23(0.0000)**	**22.05(0.0000)**	**16.90(0.0000)**	37.09(0.0000)	**14.18 (0.0000)**	**20.90 (0.0000)**
IS	16.50(0.0000)	26.97(0.0000)	17.87(0.0000)	37.64(0.0000)	15.88(0.0000)	26.51(0.0000)
IHR	15.84(0.0000)	25.03(0.0000)	16.96(0.0000)	**36.26(0.0000)**	14.82(0.0000)	24.65 (0.0000)
FNE	17.05(0.0000)	26.99(0.0000)	16.19(0.0000)	35.37(0.0000)	15.72(0.0000)	22.60 (0.0000)
Syntactic Perturb.						
OOS	19.34(0.0000)	33.43(0.0000)	19.71(0.8869)	43.27(0.7846)	20.68(0.9417)	35.45 (0.7514)
HVI	19.32(0.0000)	33.19(0.0000)	19.62(0.8724)	43.00(0.5885)	20.65(0.9793)	35.25 (0.7369)
DBI	18.77(0.0000)	31.61(0.0000)	19.15(0.0717)	42.51(0.0461)	20.25(0.1267)	34.87 (0.1257)

Table 5: Evaluation results on the obfuscated datasets (BL=BLEU, BS=BERTScore).

Dataset	CodeLlama					
	Python		Go		Java	
	BL	BS	BL	BS	BL	BS
Primary						
	22.03	38.91	12.78	26.30	15.11	33.68
Semantic Perturb.						
IOE	**11.15(0.0000)**	**13.27(0.0000)**	12.75(0.9115)	26.26(0.9274)	13.34(0.0000)	29.40 (0.0000)
IS	17.30(0.0000)	30.03 (0.0000)	12.40(0.1642)	25.54(0.0561)	13.95(0.0000)	31.43(0.0000)
IHR	11.86(0.0000)	20.02(0.0000)	**11.85(0.0004)**	**24.01(0.0000)**	**12.63(0.0000)**	**27.85 (0.0000)**
FNE	18.54(0.0000)	31.78(0.0000)	12.42(0.1873)	25.18(0.0055)	14.17(0.0000)	31.28 (0.0000)
Syntactic Perturb.						
OOS	22.08(0.9401)	38.86(0.9466)	12.77(0.9682)	26.29 (0.9858)	15.08(0.9196)	33.70 (0.9541)
HVI	21.69(0.5722)	38.53 (0.5994)	12.88(0.7320)	26.46 (0.6856)	15.22(0.7134)	33.67 (0.9909)
DBI	21.69(0.5650)	38.49(0.5690)	12.95(0.5496)	26.28(0.9671)	14.95(0.5721)	33.06 (0.2001)

Table 6: BLEU scores on the cross-obfuscated datasets. The semantic perturbation chooses the one with the greatest impact on the model in each programming language.The value in brackets is P-Value(compare with primary)

Dataset	**CodeBERT**			**CodeT5**	
	Python	**Go**	**Java**	**Python**	**Go**
Semantic Perturb.	**IHR**	**IOE**	**IHR**	**IOE**	**FNE**
primary	12.54	11.02	12.84	15.23	16.19
Cross Perturb.					
Semantic × OOS	12.52(0.8480)	10.93(0.5834)	12.87(0.8665)	14.21(0.0000)	16.25(0.822
Semantic × HVI	11.57(0.0000)	10.81(0.2122)	12.16(0.0000)	13.75(0.0000)	15.75(0.069
Semantic × DBI	**11.49(0.0000)**	**10.74(0.0796)**	**12.13(0.0000)**	**12.70(0.0000)**	**14.58(0.000**
average	11.83(0.0000)	10.83(0.2238)	12.39(0.0029)	13.55(0.0000)	15.53(0.004

Table 7:

Dataset	**CodeLlama**		
	Python	**Go**	**Java**
Semantic Perturb.	**IOE**	**IHR**	**IHR**
primary	11.15	11.85	12.63
Cross Perturb.			
Semantic × OOS	11.07(0.8733))	11.64(0.3911)	12.50(0.5637)
Semantic × HVI	8.85(0.0000)	11.82(0.9266)	12.51(0.5861)
Semantic × DBI	**8.15(0.0000)**	**11.41(0.0782)**	**12.07(0.0101)**
average	9.36(0.0001)	11.62(0.3539)	12.36(0.2058)