Appendix

We provide supplementary experimental results in this Appendix.

1 HUMAN EVALUATION

To answer the question "which BLEU is correlated with the human perception the most", we conduct the human evaluation. Here we provide the details of the correlation coefficients in this work.

1.1 Human Annotation

Each annotator is asked to assign scores from 0 to 4 to measure the semantic similarity between reference and generated summaries. The meaning of the score is shown in Table 1

Table 1: The meaning of scores in human evaluation[2].

Score	Meaning
0	No similarity between the generation and reference.
1	Have few shared tokens, not semantically similar.
2	Have some shared tokens, probable semantically similar.
3	Much similar in semantic but a few tokens are different.
4	Identical in semantic.

1.2 Correlation Coefficient

Kendall's τ formulation:

$$\tau = \frac{|Concordant - pairs| - |Discordant - pairs|}{|Concordant - pairs| + |Discordant - pairs|} \quad (1)$$

where concordant-pairs are the set where both human scores and automatic metrics suggest the same order for any two generated summaries. Discordant is the set where human assessment disagrees with the order that automatic metrics suggest for any two generated summaries. If two generated summaries are assigned the same BLEU score, then the concordant pair and discordant pair both get a half count.

Pearson correlation coefficient γ is computed as follow:

$$\gamma = \frac{Cov(X,Y)}{\sqrt{s_x^2 s_y^2}} \quad (2)$$

where X and Y are two continuous random variables. Cov(X,Y) is the covariance. s_x^2 and s_y^2 is the variance.

The Spearman correlation coefficient ρ between two variables is equal to the Pearson correlation between the rank values of those two variables.

Before the calculation of correlation coefficients, these human direct scores are converted into relative rankings as Direct Assessment Relative Rankings (DaRR) serve as the golden standard for a corpus.

1.3 Experiment Result

In the main text, we show the values of correlation coefficient under different size corpus when using arithmetic average to aggregate summary-level human score as corpus-level score. Here, we show the values of correlation coefficient under different size corpus when using geometric average to aggregate summary-level human score as corpus-level score in Table 2. The conclusion is that BLEU-DC, sentence-level BLEU with method$_4$, is more correlated with human perception.

2 ENSEMBLE MODEL

Leclair et al. [1] proposed two ensemble architecture and prove that the stacking-based technique is better than bagging-based. Thus, We use the stacking-based technique(Figure 1) to aggregate component models that are trained on the whole training set, but the training set is processed with different pre-processing. In detail, the ensemble technique just simply averages the last softmax output of each model for every time step when generating the summary.

3 RQ2 EVALUATED WITH OTHER METRICS

- Different pre-processing operations:
  Table 3, Table 4, Table 5, Table 6, Table 7, Table 8, Table 9,and Table 10 show the results evaluated with BLEU-DM, BLEU-FC, BLEU-CN, BLEU-NCS, BLEU-RC, Rouge, Meteor, and Cider, respectively.
- Statistical tests:
  Table 11, Table 12, Table 13, Table 14, Table 15, Table 16, Table 17, Table 18, and Table 19 show the results evaluated with BLEU-DM, BLEU-FC, BLEU-DC, BLEU-CN, BLEU-NCS, BLEU-RC, Rouge, Meteor, and Cider, respectively.
- Different pre-processing operation combinations:
  Table 20, Table 21, Tab 22, Table 23, Table 24, Table 25, Table 26, and Table 27 show the results evaluated with BLEU-DM, BLEU-FC, BLEU-CN, BLEU-NCS, BLEU-RC, Rouge, Meteor, and Cider, respectively.
- In summary:
  The findings and conclusions from the above tables are consistent with RQ2 in the main text.

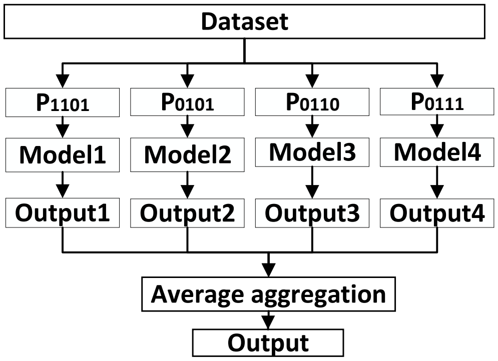

Figure 1: Ensemble Strategy.

Table 2: The values of the correlation coefficients. Here we use geometric average to aggregate summary-level human score as the corpus-level score.

Metric	1		20		40		60		80		100	
	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ
BLEU-DM s, m_0	0.32	0.68	0.45	0.63	0.46	0.64	0.46	0.64	0.46	0.64	0.45	0.64
BLEU-FC c, m_0	0.32	0.68	0.31	0.45	0.3	0.44	0.29	0.43	0.29	0.43	0.29	0.42
BLEU-DC s, m_4	**0.54**	**0.75**	**0.48**	**0.67**	**0.49**	**0.68**	**0.49**	**0.68**	**0.5**	**0.69**	**0.49**	**0.68**
BLEU-CN s, m_2	0.47	0.66	0.46	0.64	0.47	0.66	0.48	0.67	0.48	0.67	0.48	0.67
BLEU-NCS s, m_l	0.37	0.53	0.42	0.6	0.44	0.62	0.44	0.62	0.45	0.63	0.44	0.62
BLEU-RC s, m_0	0.32	0.68	0.45	0.63	0.46	0.64	0.46	0.64	0.46	0.64	0.45	0.64

4 RQ3 EVALUATED WITH OTHER METRICS

- Corpus sizes:
  Table 28, Table 29, Table 31, Table 32, Table 33, Table 34, Table 35, and Table 36 show the results evaluated with BLEU-DM, BLEU-FC, BLEU-CN, BLEU-NCS, BLEU-RC, Rouge, Meteor, and Cider, respectively.
- Data splitting ways:
  Table 37, Table 38, Table 39, Table 40, Table 41, Table 42, Table 43, and Table 44 show the results evaluated with BLEU-DM, BLEU-FC, BLEU-CN, BLEU-NCS, BLEU-RC, Rouge, Meteor, and Cider, respectively.
- Duplication ratios:
  Table 53, Table 54, Table 30, Table 55, Table 56, Table 57, Table 58, Table 59, and Table 60 show the results evaluated with BLEU-DM, BLEU-FC, BLEU-DC, BLEU-CN, BLEU-NCS, BLEU-RC, Rouge, Meteor, and Cider, respectively.
- In summary:
  The findings and conclusions from the above tables are consistent with RQ3 in the main text.

5 OTHERS

- Data difference:
  Experiments result when we control all three factors (splitting methods, duplication ratios, and dataset sizes): Table 61, Table 62, Table 63, Table 64, Table 65, Table 66, Table 67, Table 68, and Table 69 show the results evaluated with BLEU-DM, BLEU-FC, BLEU-DC, BLEU-CN, BLEU-NCS, BLEU-RC, Rouge, Meteor, and Cider, respectively.
- In summary:
  We observe that even when we control all three factors (splitting methods, duplication ratios, and dataset sizes), the performance of the same model still varies greatly between different datasets. This indicates that the differences in training data may also be a factor that affects the performance of code summarization. We leave it to future work to study the impact of data differences

Table 3: The result of four code pre-processing operations. Evaluated with BLEU-DM.

Model	R_0	R_1	S_0	S_1	F_0	F_1	L_0	L_1
CodeNN	4.4	4.38	4.44	4.34	4.39	4.39	4.4	4.38
Astattgru	3.07	3.13	2.95	3.25	3.04	3.16	2.99	3.2
Rencos	19.21	18.9	18.35	19.75	19.17	18.94	18.77	19.34
NCS	8.9	8.74	8.5	9.14	8.74	8.9	8.51	9.13
Avg.	8.90	8.78	8.56	9.12	8.84	8.85	8.67	9.01

Table 4: The result of four code pre-processing operations. Evaluated with BLEU-FC.

Model	R_0	R_1	S_0	S_1	F_0	F_1	L_0	L_1
CodeNN	5.13	5.16	5.12	5.17	5.13	5.16	5.15	5.14
Astattgru	4.48	4.52	4.34	4.66	4.43	4.57	4.38	4.62
Rencos	20.79	20.56	20.01	21.34	20.85	20.5	20.33	21.02
NCS	7.69	7.48	7.32	7.85	7.47	7.7	7.66	7.51
Avg.	9.52	9.43	9.2	9.76	9.47	9.48	9.38	9.57

Table 5: The result of four code pre-processing operations. Evaluated with BLEU-CN.

Model	R_0	R_1	S_0	S_1	F_0	F_1	L_0	L_1
CodeNN	13.50	13.51	13.39	13.62	13.51	13.51	13.52	13.49
Astattgru	12.66	12.64	12.13	13.17	12.55	12.75	12.50	12.80
Rencos	27.6	27.37	26.57	28.41	27.53	27.44	27.24	27.73
NCS	19.52	19.43	18.8	20.15	19.37	19.57	19.16	19.79
Avg.	18.32	18.24	17.72	18.84	18.24	18.32	18.10	18.45

Table 6: The result of four code pre-processing operations. Evaluated with BLEU-NCS.

Model	R_0	R_1	S_0	S_1	F_0	F_1	L_0	L_1
CodeNN	14.45	14.46	14.44	14.48	14.46	14.46	14.47	14.45
Astattgru	13.48	13.49	13.07	13.9	13.41	13.56	13.35	13.62
Rencos	28.24	28.0	27.3	28.94	28.19	28.05	27.88	28.36
NCS	19.95	19.88	19.43	20.4	19.83	20.0	19.63	20.19
Avg.	19.03	18.96	18.56	19.43	18.97	19.02	18.83	19.16

Table 7: The result of four code pre-processing operations. Evaluated with BLEU-RC.

Model	R_0	R_1	S_0	S_1	F_0	F_1	L_0	L_1
CodeNN	4.4	4.38	4.44	4.34	4.39	4.4	4.4	4.38
Astattgru	3.07	3.13	2.95	3.25	3.04	3.16	2.99	3.2
Rencos	19.21	18.9	18.35	19.75	19.17	18.94	18.77	19.34
NCS	8.9	8.74	8.5	9.14	8.74	8.9	8.51	9.13
Avg.	8.9	8.79	8.56	9.12	8.84	8.85	8.67	9.01

Table 8: The result of four code processing operations. Evaluated with Rouge.

Model	R_0	R_1	S_0	S_1	F_0	F_1	L_0	L_1
CodeNN	8.42	8.47	8.1	8.79	8.39	8.5	8.44	8.45
Astattgru	7.4	7.37	6.85	7.93	7.3	7.47	7.21	7.56
Rencos	16.63	16.53	15.78	17.37	16.64	16.51	16.38	16.77
NCS	11.16	11.07	10.37	11.86	11.02	11.21	11.0	11.24
Avg.	10.9	10.86	10.27	11.49	10.84	10.92	10.76	11.0

Table 9: The result of four code pre-processing operations. Evaluated with Meteor.

Model	R_0	R_1	S_0	S_1	F_0	F_1	L_0	L_1
CodeNN	24.8	24.95	24.32	25.44	24.84	24.91	24.87	24.88
Astattgru	24.74	24.63	23.33	26.04	24.45	24.92	24.44	24.93
Rencos	39.68	39.53	38.04	41.17	39.49	39.72	39.36	39.86
NCS	33.68	33.72	32.02	35.37	33.56	33.83	33.34	34.06
Avg.	30.72	30.71	29.43	32.01	30.58	30.84	30.5	30.93

Table 10: The result of four code pre-processing operations. Evaluated with Cider.

Model	R_0	R_1	S_0	S_1	F_0	F_1	L_0	L_1
CodeNN	0.68	0.68	0.67	0.69	0.68	0.68	0.68	0.68
Astattgru	0.5	0.5	0.45	0.55	0.49	0.51	0.48	0.52
Rencos	2.07	2.04	1.96	2.15	2.07	2.04	2.03	2.08
NCS	1.21	1.21	1.12	1.3	1.2	1.22	1.16	1.26
Avg.	1.12	1.11	1.05	1.17	1.11	1.11	1.09	1.14

Table 11: p-values of the *t-test* and *Wilcoxon-Mann-Whitney-test* (*WMW-test*). Evaluated with BLEU-DM.

Model	*t-test*				*WMW-test*			
	R	S	F	L	R	S	F	L
CodeNN	0.7737	0.2626	0.9229	0.7737	0.9581	0.3184	0.7929	0.7132
Astattgru	0.6821	0.0344	0.4273	0.1477	0.7929	0.0661	0.4309	0.1893
Rencos	0.5690	0.0043	0.6780	0.2855	0.7929	0.0054	0.1893	0.4309
NCS	0.5942	0.0188	0.5859	0.0239	0.9581	0.0239	0.9581	0.0406

Table 12: p-values of the *t-test* and *Wilcoxon-Mann-Whitney-test* (*WMW-test*). Evaluated with BLEU-FC.

Model	*t-test*				*WMW-test*			
	R	S	F	L	R	S	F	L
CodeNN	0.7675	0.5573	0.7512	0.9271	0.7132	0.9581	0.8748	0.9581
Astattgru	0.8811	0.1200	0.5286	0.2466	0.7929	0.1563	0.7132	0.2701
Rencos	0.6869	0.0089	0.5410	0.2121	0.7132	0.0136	0.1036	0.2271
NCS	0.4474	0.0309	0.3737	0.5585	0.4309	0.0520	0.4309	0.4309

Table 13: p-values of the *t-test* and *Wilcoxon-Mann-Whitney-test* (*WMW-test*). Evaluated with BLEU-DC.

Model	*t-test*				*WMW-test*			
	R	*S*	*F*	*L*	*R*	*S*	*F*	*L*
CodeNN	0.9272	0.8950	0.9098	0.8370	0.9581	0.7929	0.9581	0.9581
Astattgru	0.7846	0.0003	0.4086	0.2233	0.9581	0.0028	0.4948	0.3184
Rencos	0.6010	0.0014	0.7698	0.3267	0.7929	0.0009	0.4309	0.5635
NCS	0.7581	0.0019	0.5814	0.0731	0.9581	0.0014	0.9581	0.1893

Table 14: p-values of the *t-test* and *Wilcoxon-Mann-Whitney-test* (*WMW-test*). Evaluated with BLEU-CN.

Model	*t-test*				*WMW-test*			
	R	*S*	*F*	*L*	*R*	*S*	*F*	*L*
CodeNN	0.9366	0.0750	0.9999	0.8018	0.7132	0.0406	0.8748	0.9581
Astattgru	0.9669	0.0000	0.5421	0.3448	0.7929	0.0009	0.6365	0.4948
Rencos	0.6953	0.0003	0.8723	0.4093	0.7929	0.0009	0.4948	0.6365
NCS	0.8396	0.0002	0.6579	0.1452	0.9581	0.0009	0.7929	0.2271

Table 15: p-values of the *t-test* and *Wilcoxon-Mann-Whitney-test* (*WMW-test*). Evaluated with BLEU-NCS.

Model	*t-test*				*WMW-test*			
	R	*S*	*F*	*L*	*R*	*S*	*F*	*L*
CodeNN	0.9078	0.6509	0.9581	0.8344	0.8748	0.9581	0.8748	0.9581
Astattgru	0.9720	0.0001	0.5544	0.2888	0.8748	0.0009	0.6365	0.4948
Rencos	0.6681	0.0006	0.8009	0.3774	0.8748	0.0009	0.4309	0.6365
NCS	0.8283	0.0010	0.6285	0.0939	0.9581	0.0009	0.8748	0.2271

Table 16: p-values of the *t-test* and *Wilcoxon-Mann-Whitney-test* (*WMW-test*). Evaluated with BLEU-RC.

Model	*t-test*				*WMW-test*			
	R	*S*	*F*	*L*	*R*	*S*	*F*	*L*
CodeNN	0.7730	0.2609	0.9214	0.7746	0.9581	0.3184	0.7929	0.7132
Astattgru	0.6827	0.0345	0.4270	0.1476	0.7929	0.0661	0.4309	0.1893
Rencos	0.5690	0.0043	0.6780	0.2856	0.7929	0.0054	0.1893	0.4309
NCS	0.5944	0.0188	0.5858	0.0239	0.9581	0.0239	0.9581	0.0406

Table 17: p-values of the *t-test* and *Wilcoxon-Mann-Whitney-test* (*WMW-test*). Evaluated with Rouge.

Model	*t-test*				*WMW-test*			
	R	*S*	*F*	*L*	*R*	*S*	*F*	*L*
CodeNN	0.7348	0.0023	0.8873	0.9746	0.6365	0.0136	0.8748	0.9581
Astattgru	0.8902	0.0000	0.5605	0.5421	0.7929	0.0009	0.5635	0.5635
Rencos	0.8698	0.0000	0.7912	0.5752	0.5635	0.0009	0.9581	0.4948
NCS	0.9716	0.0000	0.7831	0.4644	0.9581	0.0009	0.6365	0.4948

Table 18: p-values of the *t-test* and *Wilcoxon-Mann-Whitney-test* (*WMW-test*). Evaluated with Meteor.

Model	*t-test*				*WMW-test*			
	R	*S*	*F*	*L*	*R*	*S*	*F*	*L*
CodeNN	0.8542	0.0011	0.6720	0.9564	0.7130	0.0117	0.2476	0.9163
Astattgru	0.9256	0.0000	0.6124	0.2802	0.7929	0.0009	0.4948	0.2701
Rencos	0.8363	0.0000	0.7848	0.4200	0.8748	0.0009	0.4309	0.4948
NCS	0.8250	0.0000	0.6639	0.5793	0.8747	0.0009	0.5283	0.5632

Table 19: p-values of the *t-test* and *Wilcoxon-Mann-Whitney-test* (*WMW-test*). Evaluated with Cider.

Model	*t-test*				*WMW-test*			
	R	*S*	*F*	*L*	*R*	*S*	*F*	*L*
CodeNN	1.0000	0.0582	0.8233	0.8233	0.8720	0.0763	1.0000	0.9145
Astattgru	0.9113	0.0002	0.6282	0.2405	1.0000	0.0009	0.5984	0.2466
Rencos	0.6509	0.0002	0.7104	0.3733	0.7525	0.0009	0.4306	0.5283
NCS	0.9018	0.0003	0.6500	0.0785	0.9580	0.0009	0.6733	0.1138

Table 20: Performance of different code pre-processing combinations. Evaluated with BLEU-DM.

Model	P_{0000}	P_{0001}	P_{0010}	P_{0011}	P_{0100}	P_{0101}	P_{0110}	P_{0111}	P_{1000}	P_{1001}	P_{1010}	P_{1011}	P_{1100}	P_{1101}	P_{1110}	P_{1111}	Ensemble
CodeNN	4.29	4.41	4.24	4.55	4.7	4.08	4.56	4.17	4.55	4.45	4.5	4.55	4.18	4.49	4.21	4.3	7.89
Astattgru	3.01	2.83	2.75	2.85	3.3	3.52	3.24	3.53	3.15	3.07	2.66	3.27	2.81	3.57	3.01	2.99	8.60
Rencos	17.54	17.78	18.78	18.47	18.73	20.64	19.53	19.71	18.34	18.38	18.72	18.83	18.58	21.52	19.91	19.37	21.52
NCS	7.99	8.85	7.94	8.96	8.53	9.69	8.69	9.25	8.62	8.41	8.66	8.55	8.95	10.14	8.67	9.16	16.84

Table 21: Performance of different code pre-processing combinations. Evaluated with BLEU-FC.

Model	P_{0000}	P_{0001}	P_{0010}	P_{0011}	P_{0100}	P_{0101}	P_{0110}	P_{0111}	P_{1000}	P_{1001}	P_{1010}	P_{1011}	P_{1100}	P_{1101}	P_{1110}	P_{1111}	Ensemble
CodeNN	5.12	5.18	4.9	5.14	5.48	5.03	5.34	5.06	5.15	4.98	5.11	5.38	5.0	5.31	5.08	5.04	8.28
Astattgru	4.56	4.01	3.84	4.13	4.95	4.8	4.73	5.1	4.76	4.63	3.78	4.97	4.11	4.73	4.27	4.62	10.32
Rencos	19.22	19.39	20.34	20.36	20.06	22.36	21.22	21.57	19.76	19.92	20.49	20.59	20.17	23.16	21.42	20.79	23.18
NCS	7.02	7.26	7.32	7.11	7.68	8.11	7.52	7.84	8.17	6.9	7.95	6.83	7.85	8.63	7.81	7.37	16.35

Table 22: Performance of different code pre-processing combinations, evaluated with BLEU-CN.

Model	P_{0000}	P_{0001}	P_{0010}	P_{0011}	P_{0100}	P_{0101}	P_{0110}	P_{0111}	P_{1000}	P_{1001}	P_{1010}	P_{1011}	P_{1100}	P_{1101}	P_{1110}	P_{1111}	Ensemble
CodeNN	13.21	13.29	13.21	13.36	14.10	13.44	13.84	13.64	13.36	13.41	13.46	13.83	13.53	13.71	13.47	13.23	16.82
Astattgru	11.98	12.40	12.08	11.98	12.74	13.77	13.10	13.09	12.25	12.43	11.99	11.92	12.74	13.68	13.12	13.13	17.36
Rencos	26.04	26.08	26.93	26.61	27.54	29.20	28.16	28.40	26.59	26.64	26.78	26.87	27.44	29.98	28.46	28.06	30.01
NCS	18.47	19.09	18.23	19.11	19.62	20.83	19.65	20.42	19.06	18.68	18.97	18.76	19.70	21.11	19.56	20.30	26.15

Table 23: Performance of different code pre-processing combinations, Evaluated with BLEU-NCS.

Model	P_{0000}	P_{0001}	P_{0010}	P_{0011}	P_{0100}	P_{0101}	P_{0110}	P_{0111}	P_{1000}	P_{1001}	P_{1010}	P_{1011}	P_{1100}	P_{1101}	P_{1110}	P_{1111}	Ensemble
CodeNN	14.27	14.43	14.37	14.38	14.84	14.25	14.67	14.5	14.43	14.51	14.47	14.64	14.37	14.58	14.31	14.3	17.65
Astattgru	12.93	13.32	13.02	12.97	13.52	14.42	13.83	13.91	13.17	13.3	12.98	12.9	13.5	14.34	13.83	13.83	18.19
Rencos	26.73	26.8	27.71	27.4	28.07	29.67	28.68	28.93	27.31	27.31	27.51	27.64	27.97	30.52	29.03	28.61	30.54
NCS	19.1	19.71	18.95	19.78	19.9	21.01	19.94	20.61	19.62	19.28	19.59	19.41	20.05	21.29	19.91	20.44	26.52

Table 24: Performance of different code pre-processing combinations. Evaluated with BLEU-RC.

Model	P_{0000}	P_{0001}	P_{0010}	P_{0011}	P_{0100}	P_{0101}	P_{0110}	P_{0111}	P_{1000}	P_{1001}	P_{1010}	P_{1011}	P_{1100}	P_{1101}	P_{1110}	P_{1111}	Ensemble
CodeNN	4.29	4.41	4.24	4.55	4.71	4.08	4.56	4.17	4.55	4.45	4.5	4.55	4.18	4.49	4.21	4.3	7.89
Astattgru	3.01	2.83	2.75	2.85	3.3	3.52	3.24	3.53	3.15	3.07	2.66	3.27	2.81	3.57	3.01	2.99	8.60
Rencos	17.54	17.78	18.78	18.47	18.73	20.64	19.53	19.71	18.34	18.38	18.72	18.83	18.58	21.52	19.91	19.37	21.52
NCS	7.99	8.85	7.94	8.96	8.53	9.69	8.69	9.25	8.62	8.41	8.66	8.55	8.95	10.14	8.67	9.16	16.84

Table 25: Performance of different code pre-processing combinations. Evaluated with Rouge.

Model	P_{0000}	P_{0001}	P_{0010}	P_{0011}	P_{0100}	P_{0101}	P_{0110}	P_{0111}	P_{1000}	P_{1001}	P_{1010}	P_{1011}	P_{1100}	P_{1101}	P_{1110}	P_{1111}	Ensemble
CodeNN	24.21	24.11	24.02	24.14	26.19	25.53	25.81	25.58	24.0	24.05	24.2	25.8	25.49	25.67	25.03	24.18	28.92
Astattgru	22.92	23.98	23.13	23.07	25.11	27.26	26.15	25.41	23.71	23.96	23.19	22.66	25.47	26.93	25.84	26.16	29.42
Rencos	37.85	37.9	38.18	37.87	40.48	41.99	40.76	41.23	38.13	38.37	38.02	38.04	40.5	42.57	40.93	40.87	42.61
NCS	31.86	32.14	31.21	32.0	34.9	36.55	34.83	36.24	32.72	31.96	32.52	31.78	34.2	36.34	34.49	35.44	40.34

Table 26: Performance of different code pre-processing combinations. Evaluated with Meteor.

Model	P_{0000}	P_{0001}	P_{0010}	P_{0011}	P_{0100}	P_{0101}	P_{0110}	P_{0111}	P_{1000}	P_{1001}	P_{1010}	P_{1011}	P_{1100}	P_{1101}	P_{1110}	P_{1111}	Ensemble
CodeNN	8.14	8.09	7.82	7.99	9.06	8.96	8.84	8.85	8.01	7.92	7.95	8.91	8.92	8.89	8.77	8.01	10.11
Astattgru	6.82	6.91	6.5	6.77	7.71	8.43	8.15	7.68	6.99	7.14	6.52	7.13	7.48	8.29	7.51	8.16	11.10
Rencos	15.39	15.51	15.97	15.87	16.72	17.92	17.29	17.54	15.69	15.81	15.95	16.06	16.71	18.33	17.34	17.12	18.35
NCS	10.2	10.45	10.19	10.38	11.63	12.31	11.37	12.01	10.73	10.28	10.64	10.1	11.6	12.48	11.6	11.88	15.79

Table 27: Performance of different code pre-processing combinations. Evaluated with Cider.

Model	P_{0000}	P_{0001}	P_{0010}	P_{0011}	P_{0100}	P_{0101}	P_{0110}	P_{0111}	P_{1000}	P_{1001}	P_{1010}	P_{1011}	P_{1100}	P_{1101}	P_{1110}	P_{1111}	Ensemble
CodeNN	0.65	0.66	0.65	0.67	0.72	0.67	0.7	0.69	0.67	0.66	0.67	0.7	0.69	0.7	0.67	0.65	0.99
Astattgru	0.44	0.46	0.44	0.43	0.52	0.62	0.56	0.54	0.45	0.47	0.42	0.46	0.49	0.61	0.52	0.56	1.07
Rencos	1.89	1.9	1.99	1.97	2.06	2.23	2.13	2.15	1.95	1.96	1.98	2.01	2.04	2.31	2.17	2.13	2.31
NCS	1.08	1.16	1.05	1.17	1.22	1.4	1.22	1.35	1.12	1.13	1.11	1.12	1.25	1.43	1.22	1.33	1.97

Table 28: Performance in Different datasets. Evaluated with BLEU-DM.

Model	Dataset		
	TLC	FCM	CSN
CodeNN	26.23±0.15	7.26±0.14	1.56±0.11
Deepcom	13.38±2.00	4.95±0.04	0.99±0.25
Astattgru	24.09±0.74	9.92±0.06	5.14±4.05
Rencos	40.40±0.01	10.40±0.00	4.76±0.06
NCS	37.30±0.23	12.26±0.54	3.13±0.53

Table 29: Performance in Different datasets. Evaluated with BLEU-FC.

Model	Dataset		
	TLC	FCM	CSN
CodeNN	25.63±0.47	11.99±0.21	2.75±0.02
Deepcom	11.19±2.89	8.73±0.09	1.52±0.45
Astattgru	25.21±0.96	15.94±0.09	4.87±0.91
Rencos	41.20±0.17	16.34±0.00	7.29±0.16
NCS	34.74±0.83	17.72±0.74	2.11±1.65

Table 30: Performance in Different datasets. Evaluated with BLEU-DC

Model	Dataset		
	TLC	FCM	CSN
CodeNN	28.24±0.19	12.64±0.13	3.32±0.09
Deepcom	15.65±2.12	9.12±0.03	1.98±0.30
Astattgru	25.90±0.79	15.58±0.11	6.86±3.07
Rencos	42.46±0.05	15.47±0.00	6.65±0.05
NCS	39.50±0.23	18.07±0.46	6.66±0.51

Table 31: Performance on different datasets. Evaluated with BLEU-CN.

Model	Dataset		
	TLC	FCM	CSN
CodeNN	33.03±0.20	25.26±0.01	8.58±0.15
Deepcom	20.54±2.57	20.80±0.02	6.12±0.64
Astattgru	30.19±0.86	27.63±0.24	11.73±0.41
Rencos	46.81±0.06	25.82±0.00	11.19±0.09
NCS	44.25±0.21	30.69±0.12	11.80±0.94

Table 32: Performance in Different datasets. Evaluated with BLEU-NCS.

Model	Dataset		
	TLC	FCM	CSN
CodeNN	33.73±0.19	26.43±0.06	11.40±0.32
Deepcom	21.63±2.43	23.44±0.06	9.18±0.30
Astattgru	30.92±0.81	28.44±0.24	12.44±1.14
Rencos	47.23±0.05	28.23±0.00	14.76±0.08
NCS	44.55±0.22	31.18±0.34	14.68±0.62

Table 33: Performance in Different datasets. Evaluated with BLEU-RC.

Model	Dataset		
	TLC	FCM	CSN
CodeNN	26.23±0.15	7.28±0.14	1.56±0.11
Deepcom	13.38±2.00	4.96±0.04	0.99±0.25
Astattgru	24.09±0.74	9.94±0.06	5.14±4.05
Rencos	40.40±0.01	10.41±0.00	4.76±0.06
NCS	37.30±0.23	12.30±0.53	3.13±0.53

Table 34: Performance in Different datasets. Evaluated with Rouge.

Model	Dataset		
	TLC	FCM	CSN
CodeNN	42.67±0.34	34.98±0.05	16.01±0.38
Deepcom	30.44±3.48	28.13±0.05	10.86±0.23
Astattgru	39.07±0.99	38.59±0.22	22.20±1.76
Rencos	56.45±0.08	35.78±0.00	20.78±0.10
NCS	54.68±0.22	40.61±0.19	21.91±2.34

Table 35: Performance in Different datasets. Evaluated with Meteor.

Model	Dataset		
	TLC	FCM	CSN
CodeNN	19.24±0.23	14.84±0.21	5.36±0.26
Deepcom	12.10±1.33	10.42±0.01	3.72±0.35
Astattgru	17.52±0.60	17.72±0.10	8.68±1.01
Rencos	28.63±0.05	16.55±0.00	8.94±0.08
NCS	26.22±0.30	19.04±0.25	6.39±1.19

Table 36: Performance in Different datasets. Evaluated with Cider.

Model	Dataset		
	TLC	FCM	CSN
CodeNN	2.74±0.02	1.36±0.01	0.32±0.02
Deepcom	1.46±0.22	0.76±0.01	0.15±0.03
Astattgru	2.44±0.08	1.78±0.02	0.36±0.19
Rencos	4.12±0.01	1.63±0.00	0.58±0.01
NCS	3.89±0.02	2.21±0.03	0.63±0.06

Table 37: The result in different corpus sizes. Evaluated with BLEU-DM.

Model	FCM$_{\text{Method-Small}}$	FCM$_{\text{Method-Medium}}$	FCM$_{\text{Method-Large}}$	CSN$_{\text{Method-Small}}$	CSN$_{\text{Method-Medium}}$
CodeNN	5.31±0.17	9.10±0.15	12.60±0.25	3.28±0.06	9.26±0.25
Deepcom	4.86±0.05	6.35±0.16	6.96±0.36	5.76±0.75	5.78±0.78
Astattgru	7.80±0.51	12.41±0.06	15.76±0.08	3.93±0.08	12.31±0.22
Rencos	9.48±0.17	16.78±0.08	18.07±0.05	5.61±0.07	16.71±0.04
NCS	8.98±0.09	17.04±0.31	22.93±0.27	5.16±0.13	18.66±0.28

Table 38: The result in different corpus sizes. Evaluated with BLEU-FC.

Model	FCM$_{\text{Method-Small}}$	FCM$_{\text{Method-Medium}}$	FCM$_{\text{Method-Large}}$	CSN$_{\text{Method-Small}}$	CSN$_{\text{Method-Medium}}$
CodeNN	9.70±0.23	14.54±0.27	19.01±0.30	3.95±0.05	9.88±0.18
Deepcom	8.23±0.09	10.41±0.28	11.34±0.54	3.34±0.03	3.77±0.75
Astattgru	12.93±0.71	18.84±0.06	22.91±0.08	4.79±0.45	11.85±0.28
Rencos	14.86±0.15	23.65±0.10	25.57±0.05	7.35±0.52	16.26±0.16
NCS	13.66±0.40	23.35±0.56	29.64±0.54	2.61±0.14	11.39±0.75

Table 39: The result in different corpus sizes. Evaluated with BLEU-CN.

Model	FCM$_{\text{Method-Small}}$	FCM$_{\text{Method-Medium}}$	FCM$_{\text{Method-Large}}$	CSN$_{\text{Method-Small}}$	CSN$_{\text{Method-Medium}}$
CodeNN	22.85±0.12	27.38±0.04	31.19±0.17	9.38±0.14	20.13±0.34
Deepcom	20.49±0.16	22.78±0.12	23.72±0.32	12.30±0.64	12.64±1.07
Astattgru	23.94±0.67	29.83±0.20	33.36±0.16	11.38±0.42	24.11±0.25
Rencos	24.02±0.03	31.47±0.04	33.95±0.03	11.73±0.16	25.03±0.02
NCS	**27.89±0.37**	**35.41±0.20**	**40.73±0.16**	**12.74±0.13**	**30.12±0.27**
OOV Ratio of Deepcom	91.90%	88.94%	88.32%	91.49%	85.81%
OOV Ratio of Others	63.36%	53.09%	48.60%	60.99%	34.00%

Table 40: The result in different corpus sizes. Evaluated with BLEU-NCS.

Model	FCM$_{\text{Method-Small}}$	FCM$_{\text{Method-Medium}}$	FCM$_{\text{Method-Large}}$	CSN$_{\text{Method-Small}}$	CSN$_{\text{Method-Medium}}$
CodeNN	24.45±0.09	28.40±0.10	31.96±0.19	12.57±0.13	21.63±0.17
Deepcom	23.18±0.06	24.92±0.20	25.62±0.37	14.49±0.23	14.96±0.95
Astattgru	25.53±0.64	30.60±0.14	33.92±0.14	12.99±0.13	24.05±0.34
Rencos	26.83±0.08	33.98±0.05	35.91±0.02	15.04±0.16	27.64±0.00
NCS	28.29±0.27	35.55±0.29	40.71±0.28	16.57±0.16	32.55±0.27

Table 41: The result in different corpus sizes. Evaluated with BLEU-RC.

Model	FCM$_{\text{Method-Small}}$	FCM$_{\text{Method-Medium}}$	FCM$_{\text{Method-Large}}$	CSN$_{\text{Method-Small}}$	CSN$_{\text{Method-Medium}}$
CodeNN	5.32±0.17	9.12±0.15	12.64±0.25	3.28±0.06	9.28±0.25
Deepcom	4.87±0.05	6.36±0.16	6.97±0.36	5.76±0.75	5.79±0.78
Astattgru	7.82±0.51	12.44±0.06	15.80±0.09	3.93±0.08	12.32±0.22
Rencos	9.49±0.17	16.82±0.08	18.11±0.04	5.62±0.07	16.73±0.04
NCS	9.02±0.09	17.10±0.31	23.00±0.27	5.16±0.13	18.69±0.28

Table 42: The result in different corpus sizes. Evaluated with Rouge.

Model	$\text{FCM}_{\text{Method-Small}}$	$\text{FCM}_{\text{Method-Medium}}$	$\text{FCM}_{\text{Method-Large}}$	$\text{CSN}_{\text{Method-Small}}$	$\text{CSN}_{\text{Method-Medium}}$
CodeNN	32.17±0.22	37.83±0.16	42.43±0.29	16.53±0.15	30.24±0.29
Deepcom	27.68±0.15	30.79±0.36	32.04±0.61	18.64±0.35	19.00±0.87
Astattgru	34.08±1.02	41.13±0.23	44.94±0.15	18.37±0.11	34.21±0.64
Rencos	33.56±0.03	41.86±0.05	45.37±0.04	19.87±0.20	37.04±0.11
NCS	37.83±0.86	45.67±0.31	50.89±0.33	21.22±0.29	40.97±0.52

Table 43: The result in different corpus sizes. Evaluated with Meteor.

Model	$\text{FCM}_{\text{Method-Small}}$	$\text{FCM}_{\text{Method-Medium}}$	$\text{FCM}_{\text{Method-Large}}$	$\text{CSN}_{\text{Method-Small}}$	$\text{CSN}_{\text{Method-Medium}}$
CodeNN	12.93±0.20	16.30±0.15	19.01±0.21	5.95±0.04	11.53±0.12
Deepcom	9.97±0.01	11.67±0.16	12.48±0.28	7.47±0.25	7.47±0.25
Astattgru	14.87±0.64	19.04±0.09	21.40±0.07	6.37±0.20	13.20±0.04
Rencos	15.00±0.03	20.00±0.05	21.84±0.03	7.67±0.21	15.48±0.11
NCS	16.71±0.33	21.91±0.22	25.25±0.23	6.97±0.07	15.62±0.24

Table 44: The result in different corpus sizes. Evaluated with Cider.

Model	$\text{FCM}_{\text{Method-Small}}$	$\text{FCM}_{\text{Method-Medium}}$	$\text{FCM}_{\text{Method-Large}}$	$\text{CSN}_{\text{Method-Small}}$	$\text{CSN}_{\text{Method-Medium}}$
CodeNN	0.99±0.02	1.50±0.02	1.95±0.03	0.44±0.01	1.21±0.03
Deepcom	0.69±0.02	0.91±0.02	1.01±0.04	0.67±0.08	0.68±0.09
Astattgru	1.25±0.10	2.00±0.02	2.41±0.02	0.45±0.00	1.46±0.03
Rencos	1.37±0.01	2.21±0.01	2.48±0.00	0.64±0.01	1.82±0.01
NCS	1.82±0.01	2.70±0.02	3.27±0.02	0.75±0.02	2.20±0.03

Table 45: The result in different data splitting methods. Evaluated with BLEU-DM.

Model	$\text{CSN}_{\text{Project-Medium}}$	$\text{CSN}_{\text{Class-Medium}}$	$\text{CSN}_{\text{Method-Medium}}$	$\text{FCM}_{\text{Project-Large}}$	$\text{FCM}_{\text{Method-Large}}$
CodeNN	1.56±0.11	6.82±0.08	9.26±0.25	7.26±0.14	12.60±0.25
Deepcom	0.99±0.25	4.31±0.22	5.78±0.78	4.95±0.04	6.96±0.36
Astattgru	5.14±4.05	8.55±0.41	12.31±0.22	9.92±0.06	15.76±0.08
Rencos	4.76±0.06	11.85±0.04	16.71±0.04	10.40±0.00	18.07±0.05
NCS	3.13±0.53	12.25±0.14	18.66±0.28	12.26±0.54	22.93±0.27

Table 46: The result in different data splitting methods. Evaluated with BLEU-FC.

Model	$\text{CSN}_{\text{Project-Medium}}$	$\text{CSN}_{\text{Class-Medium}}$	$\text{CSN}_{\text{Method-Medium}}$	$\text{FCM}_{\text{Project-Large}}$	$\text{FCM}_{\text{Method-Large}}$
CodeNN	2.75±0.02	7.59±0.17	9.88±0.18	11.99±0.21	19.01±0.30
Deepcom	1.52±0.45	3.37±0.14	3.77±0.75	8.73±0.09	11.34±0.54
Astattgru	4.87±0.91	8.60±0.65	11.85±0.28	15.94±0.09	22.91±0.08
Rencos	7.29±0.16	12.20±0.14	16.26±0.16	16.34±0.00	25.57±0.05
NCS	2.11±1.65	7.46±0.30	11.39±0.75	17.72±0.74	29.64±0.54

Table 47: The result of different data splitting ways. Evaluated with BLEU-CN.

Model	$\text{CSN}_{\text{Project-Medium}}$	$\text{CSN}_{\text{Class-Medium}}$	$\text{CSN}_{\text{Method-Medium}}$	$\text{FCM}_{\text{Project-Large}}$	$\text{FCM}_{\text{Method-Large}}$
CodeNN	8.58±0.15	16.16±0.20	20.13±0.34	25.26±0.00	31.19±0.17
Deepcom	6.12±0.64	11.29±0.21	12.64±1.07	20.80±0.02	23.72±0.32
Astattgru	11.73±0.41	20.22±0.39	24.11±0.25	27.63±0.24	33.36±0.16
Rencos	11.19±0.09	19.75±0.10	25.03±0.02	25.82±0.00	33.95±0.03
NCS	**11.80±0.94**	**23.25±0.13**	**30.12±0.27**	**30.69±0.12**	**40.73±0.16**
OOV Ratio	48.74%	35.38%	34.00%	57.56%	48.60%

Table 48: The result in different data splitting methods. Evaluated with BLEU-NCS.

Model	$\text{CSN}_{\text{Project-Medium}}$	$\text{CSN}_{\text{Class-Medium}}$	$\text{CSN}_{\text{Method-Medium}}$	$\text{FCM}_{\text{Project-Large}}$	$\text{FCM}_{\text{Method-Large}}$
CodeNN	11.40±0.32	18.06±0.18	21.63±0.17	26.43±0.06	31.96±0.19
Deepcom	9.18±0.30	14.04±0.34	14.96±0.95	23.44±0.06	25.62±0.37
Astattgru	12.44±1.14	20.07±0.70	24.05±0.34	28.44±0.24	33.92±0.14
Rencos	14.76±0.08	22.62±0.05	27.64±0.00	28.23±0.00	35.91±0.02
NCS	14.68±0.62	25.77±0.17	32.55±0.27	31.18±0.34	40.71±0.28

Table 49: The result in different data splitting methods. Evaluated with BLEU-RC.

Model	$\text{CSN}_{\text{Project-Medium}}$	$\text{CSN}_{\text{Class-Medium}}$	$\text{CSN}_{\text{Method-Medium}}$	$\text{FCM}_{\text{Project-Large}}$	$\text{FCM}_{\text{Method-Large}}$
CodeNN	1.56±0.11	6.83±0.08	9.28±0.25	7.28±0.14	12.64±0.25
Deepcom	0.99±0.25	4.31±0.22	5.79±0.78	4.96±0.04	6.97±0.36
Astattgru	5.14±4.05	8.55±0.41	12.32±0.22	9.94±0.06	15.80±0.09
Rencos	4.76±0.06	11.86±0.04	16.73±0.04	10.41±0.00	18.11±0.04
NCS	3.13±0.53	12.28±0.14	18.69±0.28	12.30±0.53	23.00±0.27

Table 50: The result in different data splitting methods. Evaluated with Rouge.

Model	$\text{CSN}_{\text{Project-Medium}}$	$\text{CSN}_{\text{Class-Medium}}$	$\text{CSN}_{\text{Method-Medium}}$	$\text{FCM}_{\text{Project-Large}}$	$\text{FCM}_{\text{Method-Large}}$
CodeNN	16.01±0.38	24.77±0.24	30.24±0.29	34.98±0.05	42.43±0.29
Deepcom	10.86±0.23	17.10±0.30	19.00±0.87	28.13±0.05	32.04±0.61
Astattgru	22.20±1.76	29.61±0.81	34.21±0.64	38.59±0.22	44.94±0.15
Rencos	20.78±0.10	30.57±0.14	37.04±0.11	35.78±0.00	45.37±0.04
NCS	21.91±2.34	33.48±0.15	40.97±0.52	40.61±0.19	50.89±0.33

Table 51: The result in different data splitting methods. Evaluated with Meteor.

Model	$\text{CSN}_{\text{Project-Medium}}$	$\text{CSN}_{\text{Class-Medium}}$	$\text{CSN}_{\text{Method-Medium}}$	$\text{FCM}_{\text{Project-Large}}$	$\text{FCM}_{\text{Method-Large}}$
CodeNN	5.36±0.26	9.74±0.14	11.53±0.12	14.84±0.21	19.01±0.21
Deepcom	3.72±0.35	6.76±0.23	7.47±0.25	10.42±0.01	12.48±0.28
Astattgru	8.68±1.01	11.47±0.31	13.20±0.04	17.72±0.10	21.40±0.07
Rencos	8.94±0.08	12.82±0.14	15.48±0.11	16.55±0.00	21.84±0.03
NCS	6.39±1.19	12.86±0.16	15.62±0.24	19.04±0.25	25.25±0.23

Table 52: The result in different data splitting methods. Evaluated with Cider.

Model	$\text{CSN}_{\text{Project-Medium}}$	$\text{CSN}_{\text{Class-Medium}}$	$\text{CSN}_{\text{Method-Medium}}$	$\text{FCM}_{\text{Project-Large}}$	$\text{FCM}_{\text{Method-Large}}$
CodeNN	0.32±0.02	0.90±0.01	1.21±0.03	1.36±0.01	1.95±0.03
Deepcom	0.15±0.03	0.53±0.01	0.68±0.09	0.76±0.01	1.01±0.04
Astattgru	0.36±0.19	1.06±0.06	1.46±0.03	1.78±0.02	2.41±0.02
Rencos	0.58±0.01	1.30±0.01	1.82±0.01	1.63±0.00	2.48±0.00
NCS	0.63±0.06	1.56±0.01	2.20±0.03	2.21±0.03	3.27±0.02

Table 53: The result of different duplication ratios. Evaluated with BLEU-DM.

Model	duplication ratio					
	0.0	0.1	0.3	0.5	0.7	0.9
CodeNN	4.49	6.01	9.61	13.51	17.00	20.60
Deepcom	3.51	5.10	8.27	11.33	14.43	18.13
Astattgru	4.15	7.79	15.23	22.34	29.98	37.21
Rencos	21.52	29.25	44.60	59.94	75.26	90.64
NCS	10.14	17.10	26.04	35.22	43.77	53.51

Table 54: The result of different duplication ratios. Evaluated with BLEU-FC.

Model	duplication ratio					
	0.0	0.1	0.3	0.5	0.7	0.9
CodeNN	5.31	6.69	9.59	12.42	15.21	17.73
Deepcom	4.03	5.20	7.51	9.65	11.59	14.08
Astattgru	6.39	9.85	15.96	22.14	28.13	33.73
Rencos	23.16	30.05	44.22	57.51	71.27	84.52
NCS	8.63	13.70	19.74	26.33	32.51	39.59

Table 55: The result of different duplication ratios. Evaluated with BLEU-CN.

Model	duplication ratio					
	0.0	0.1	0.3	0.5	0.7	0.9
CodeNN	13.71	15.63	19.49	23.79	27.57	31.59
Deepcom	11.26	13.21	17.03	20.84	24.56	28.88
Astattgru	12.28	15.64	22.5	29.22	36.19	43.0
Rencos	29.98	36.86	50.52	64.18	77.82	91.5
NCS	21.11	26.95	34.66	42.7	50.06	58.56

Table 56: The result of different duplication ratios. Evaluated with BLEU-NCS.

Model	duplication ratio					
	0.0	0.1	0.3	0.5	0.7	0.9
CodeNN	14.58	16.32	20.04	24.18	27.81	31.70
Deepcom	12.68	14.50	18.04	21.61	25.04	29.11
Astattgru	13.16	16.45	23.15	29.70	36.53	43.19
Rencos	30.52	37.34	50.89	64.44	77.97	91.54
NCS	21.29	27.11	34.64	42.57	49.77	58.13

REFERENCES

[1] Alexander LeClair, Aakash Bansal, and Collin McMillan. 2021. Ensemble Models for Neural Source Code Summarization of Subroutines. *CoRR* abs/2107.11423 (2021).

[2] Wei Tao, Yanlin Wang, Ensheng Shi, Lun Du, Shi Han, Hongyu Zhang, Dongmei Zhang, and Wenqiang Zhang. 2021. On the Evaluation of Commit Message Generation Models: An Experimental Study. *CoRR* abs/2107.05373 (2021).

Table 57: The result of different duplication ratios. Evaluated with BLEU-RC.

Model	duplication ratio					
	0.0	0.1	0.3	0.5	0.7	0.9
CodeNN	4.49	6.01	9.62	13.51	17.00	20.60
Deepcom	3.51	5.10	8.27	11.33	14.43	18.13
Astattgru	4.15	7.79	15.23	22.34	29.98	37.21
Rencos	21.52	29.25	44.60	59.94	75.26	90.64
NCS	10.14	17.10	26.04	35.22	43.77	53.51

Table 58: The result of different duplication ratios. Evaluated with Rouge.

Model	duplication ratio					
	0.0	0.1	0.3	0.5	0.7	0.9
CodeNN	25.67	28.55	34.02	39.87	45.35	50.93
Deepcom	19.99	22.95	28.61	34.32	40.10	46.31
Astattgru	23.38	26.79	33.65	40.60	47.61	54.56
Rencos	42.57	48.23	59.51	70.78	82.05	93.34
NCS	36.34	41.34	48.78	56.38	63.66	71.63

Table 59: The result of different duplication ratios. Evaluated with Meteor.

Model	duplication ratio					
	0.0	0.1	0.3	0.5	0.7	0.9
CodeNN	8.89	10.23	12.73	15.33	17.87	20.22
Deepcom	6.87	8.23	10.90	13.58	16.30	19.16
Astattgru	8.44	10.01	13.28	16.86	20.53	24.12
Rencos	18.33	22.17	30.69	39.69	50.27	63.46
NCS	12.48	15.38	19.39	23.62	27.70	32.23

Table 60: The result of different duplication ratios. Evaluated with Cider.

Model	duplication ratio					
	0.0	0.1	0.3	0.5	0.7	0.9
CodeNN	0.70	0.89	1.28	1.69	2.04	2.37
Deepcom	0.51	0.70	1.07	1.44	1.81	2.22
Astattgru	0.57	0.90	1.60	2.25	2.95	3.60
Rencos	2.31	3.05	4.52	5.97	7.44	8.91
NCS	1.43	2.04	2.84	3.67	4.41	5.25

Table 61: Performance (BLEU-DM) in Different datasets with same data splitting way, corpus size, and duplication ratio.

Model	$\text{TLC}_{\text{Dedup}}$	$\text{FCM}_{\text{Method-Small}}$	$\text{CSN}_{\text{Method-Small}}$
CodeNN	7.12±0.14	5.31±0.17	3.28±0.06
Deepcom	3.46±0.86	4.86±0.05	5.76±0.75
Astattgru	4.83±0.39	7.80±0.51	3.93±0.08
Rencos	21.48±0.03	9.48±0.17	5.61±0.07
NCS	18.42±0.10	8.98±0.09	5.16±0.13

Table 62: Performance (BLEU-FC) in Different datasets with same data splitting way, corpus size, and duplication ratio.

Model	TLC$_{\text{Dedup}}$	FCM$_{\text{Method-Small}}$	CSN$_{\text{Method-Small}}$
CodeNN	8.64±0.21	9.70±0.23	3.95±0.05
Deepcom	3.69±1.21	8.23±0.09	3.34±0.03
Astattgru	6.91±0.33	12.93±0.71	4.79±0.45
Rencos	23.56±0.13	14.86±0.15	7.35±0.52
NCS	18.33±0.34	13.66±0.40	2.61±0.14

Table 63: Performance (BLEU-DC) in Different datasets with same data splitting way, corpus size, and duplication ratio.

Model	TLC$_{\text{Dedup}}$	FCM$_{\text{Method-Small}}$	CSN$_{\text{Method-Small}}$
CodeNN	9.69±0.10	10.37±0.17	5.20±0.01
Deepcom	5.65±1.14	8.99±0.06	7.57±0.74
Astattgru	7.07±0.50	12.86±0.64	5.89±0.12
Rencos	24.36±0.00	14.24±0.12	7.36±0.08
NCS	21.67±0.09	14.70±0.19	9.07±0.20

Table 64: Performance(BLEU-CN) in Different dataset with same split way, corpus size, and duplication ratio.

Model	TLC$_{\text{Dedup}}$	FCM$_{\text{Method-Small}}$	CSN$_{\text{Method-Small}}$
CodeNN	15.59±0.05	22.85±0.12	9.38±0.14
Deepcom	10.77±1.73	20.49±0.16	12.30±0.64
Astattgru	12.49±0.59	23.94±0.67	11.38±0.42
Rencos	29.85±0.03	24.02±0.03	11.73±0.16
NCS	27.38±0.14	28.17±0.00	12.67±0.07

Table 65: Performance (BLEU-NCS) in Different datasets with same data splitting way, corpus size, and duplication ratio.

Model	TLC$_{\text{Dedup}}$	FCM$_{\text{Method-Small}}$	CSN$_{\text{Method-Small}}$
CodeNN	16.69±0.07	24.45±0.09	12.57±0.13
Deepcom	12.06±1.88	23.18±0.06	14.49±0.23
Astattgru	13.36±0.65	25.53±0.64	12.99±0.13
Rencos	30.92±0.03	26.83±0.08	15.04±0.16
NCS	28.96±0.10	28.29±0.27	16.57±0.16

Table 66: Performance (BLEU-RC) in Different datasets with same data splitting way, corpus size, and duplication ratio.

Model	TLC$_{\text{Dedup}}$	FCM$_{\text{Method-Small}}$	CSN$_{\text{Method-Small}}$
CodeNN	7.12±0.14	5.32±0.17	3.28±0.06
Deepcom	3.46±0.86	4.87±0.05	5.76±0.75
Astattgru	4.83±0.39	7.82±0.51	3.93±0.08
Rencos	21.48±0.03	9.49±0.17	5.62±0.07
NCS	18.42±0.10	8.98±0.09	5.16±0.13

Table 67: Performance (Rouge) in Different datasets with same data splitting way, corpus size, and duplication ratio.

Model	$\text{TLC}_{\text{Dedup}}$	$\text{FCM}_{\text{Method-Small}}$	$\text{CSN}_{\text{Method-Small}}$
CodeNN	26.64±0.35	32.17±0.22	16.53±0.15
Deepcom	18.92±3.22	27.68±0.15	18.64±0.35
Astattgru	22.77±0.98	34.08±1.02	18.37±0.11
Rencos	42.88±0.02	33.56±0.03	19.87±0.20
NCS	41.76±0.34	37.83±0.86	21.22±0.29

Table 68: Performance (Meteor) in Different datasets with same data splitting way, corpus size, and duplication ratio.

Model	$\text{TLC}_{\text{Dedup}}$	$\text{FCM}_{\text{Method-Small}}$	$\text{CSN}_{\text{Method-Small}}$
CodeNN	11.99±0.18	12.93±0.20	5.95±0.04
Deepcom	5.79±1.77	9.97±0.01	7.47±0.25
Astattgru	8.86±0.47	14.87±0.64	6.37±0.20
Rencos	22.97±0.16	15.00±0.03	7.67±0.21
NCS	23.12±0.22	16.71±0.33	6.97±0.07

Table 69: Performance (Cider) in Different datasets with same data splitting way, corpus size, and duplication ratio.

Model	$\text{TLC}_{\text{Dedup}}$	$\text{FCM}_{\text{Method-Small}}$	$\text{CSN}_{\text{Method-Small}}$
CodeNN	0.99±0.01	0.99±0.02	0.44±0.01
Deepcom	0.57±0.11	0.69±0.02	0.67±0.08
Astattgru	0.61±0.06	1.25±0.10	0.45±0.00
Rencos	2.30±0.00	1.37±0.01	0.64±0.01
NCS	2.13±0.01	1.82±0.01	0.75±0.02