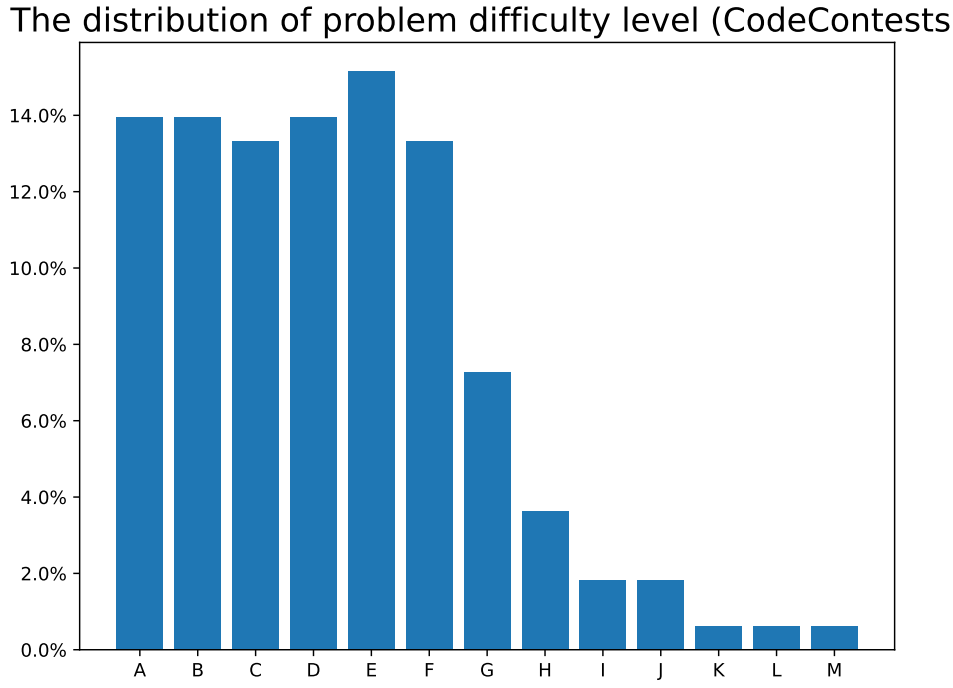

Fig. 1. The distribution of problem difficulty level (CodeContests)

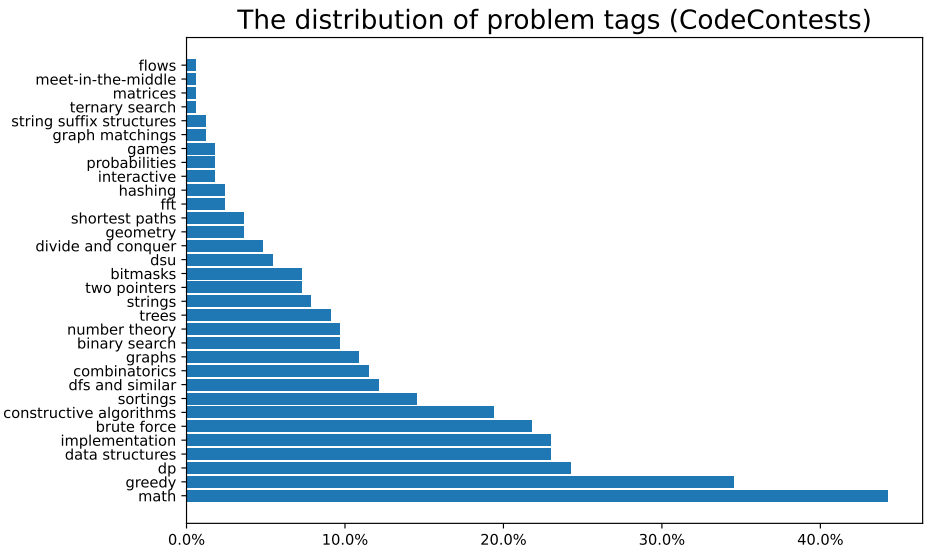

Fig. 2. The distribution of problem tags (CodeContests)

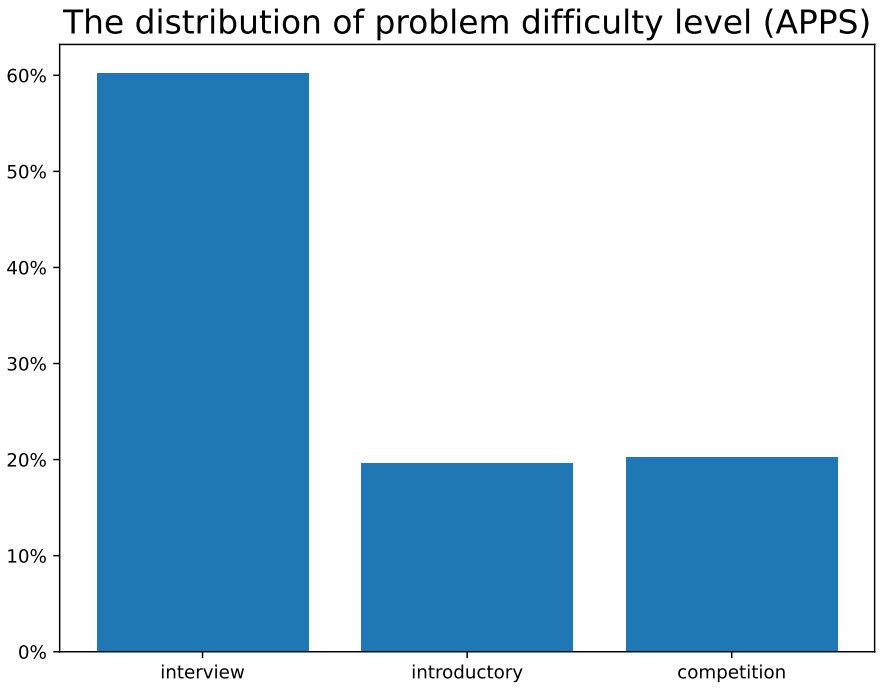

Fig. 3. The distribution of problem difficulty level (APPS)

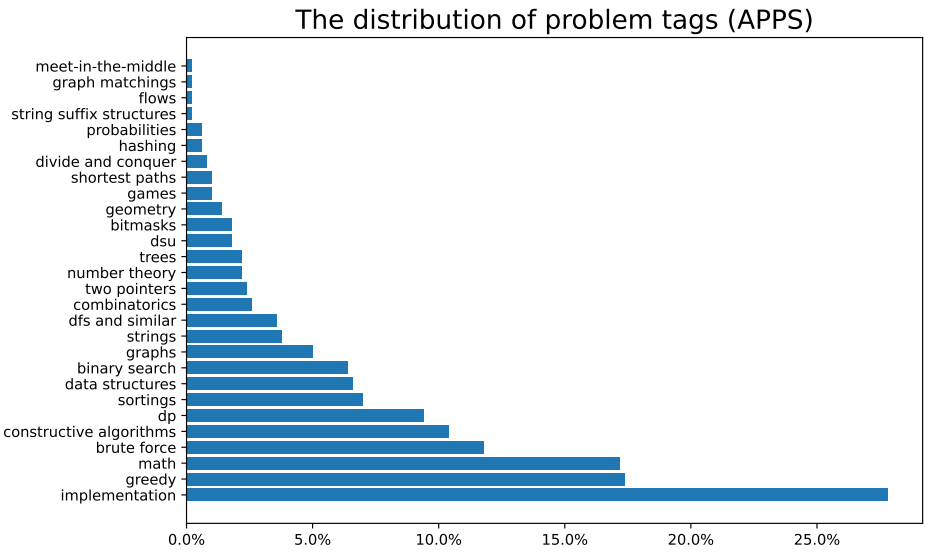

Fig. 4. The distribution of problem tags (APPS)

Table 1. RQ2: Influence of temperature (CodeContests).

Temperature	Test Pass Rate					
	Mean value	Mean variance	Mean max diff	Max diff	Ratio of worst cases	
0	0.15	0.01	0.11	1.00	1.82%	
0.5	0.16	0.02	0.15	1.00	2.42%	
1	0.16	0.03	0.24	1.00	3.64%	
Temperature	OER			OER (no ex.)		
	Mean value	Ratio of worst cases	Pair mean value	Mean value	Ratio of worst cases	Pair mean value
0	0.37	43.64%	0.59	0.27	54.55%	0.46
0.5	0.18	62.42%	0.37	0.13	68.48%	0.28
1	0.09	75.76%	0.27	0.06	81.21%	0.19
Temperature	LCS			LED		
	Mean value	Mean worst value	Pair mean value	Mean value	Mean worst value	Pair mean value
0	0.61	0.44	0.62	23.45	35.87	22.31
0.5	0.33	0.23	0.34	44.48	62.02	44.89
1	0.22	0.16	0.23	58.80	77.46	58.86
Temperature	United_Diff			Tree_Diff		
	Mean value	Mean worst value	Pair mean value	Mean value	Mean worst value	Pair mean value
0	0.41	0.39	0.67	0.50	0.46	0.74
0.5	0.61	0.49	0.63	0.69	0.58	0.71
1	0.33	0.27	0.46	0.41	0.33	0.56

Table 2. RQ2: Influence of temperature (APPS).

Temperature	Test Pass Rate					
	Mean value	Mean variance	Mean max diff	Max diff	Ratio of worst cases	
0	0.43	0.01	0.14	1.00	1.80%	
0.5	0.42	0.03	0.27	1.00	6.20%	
1	0.42	0.04	0.35	1.00	10.40%	
Temperature	OER			OER (no ex.)		
	Mean value	Ratio of worst cases	Pair mean value	Mean value	Ratio of worst cases	Pair mean value
0	0.56	27.4%	0.73	0.50	32.8%	0.65
0.5	0.36	42.20%	0.56	0.33	46.20%	0.50
1	0.27	51.0%	0.47	0.25	53.4%	0.42
Temperature	LCS			LED		
	Mean value	Mean worst value	Pair mean value	Mean value	Mean worst value	Pair mean value
0	0.65	0.50	0.66	18.18	28.41	17.40
0.5	0.37	0.26	0.37	35.00	48.37	34.86
1	0.23	0.16	0.24	47.37	61.55	46.94
Temperature	United_Diff			Tree_Diff		
	Mean value	Mean worst value	Pair mean value	Mean value	Mean worst value	Pair mean value
0	0.49	0.46	0.70	0.60	0.57	0.77
0.5	0.67	0.55	0.69	0.75	0.65	0.77
1	0.43	0.35	0.52	0.54	0.47	0.63

Table 3. RQ2: Influence of temperature (HumanEval).

Temperature	Test Pass Rate					
	Mean value	Mean variance	Mean max diff	Max diff	Ratio of worst cases	
0	0.65	0.03	0.17	1.00	14.02%	
0.5	0.62	0.05	0.30	1.00	20.73%	
1	0.63	0.09	0.53	1.00	39.63%	
Temperature	OER			OER (no ex.)		
	Mean value	Ratio of worst cases	Pair mean value	Mean value	Ratio of worst cases	Pair mean value
0	0.77	18.29%	0.89	0.72	23.17%	0.82
0.5	0.62	26.83%	0.80	0.58	30.49%	0.74
1	0.39	47.56%	0.67	0.35	51.22%	0.61
Temperature	LCS			LED		
	Mean value	Mean worst value	Pair mean value	Mean value	Mean worst value	Pair mean value
0	0.80	0.68	0.81	7.80	14.73	7.67
0.5	0.59	0.42	0.57	17.57	29.75	18.11
1	0.42	0.25	0.41	26.56	43.91	27.10
Temperature	United_Diff			Tree_Diff		
	Mean value	Mean worst value	Pair mean value	Mean value	Mean worst value	Pair mean value
0	0.67	0.63	0.81	0.70	0.65	0.83
0.5	0.82	0.71	0.81	0.86	0.75	0.84
1	0.60	0.47	0.67	0.62	0.48	0.70

Table 4. RQ3: Similarity for different request ways (CodeContests), where t represents the temperature setting.

Request	Test Pass Rate					
Way	Mean value	Mean variance	Mean max diff	Max diff	Ratio of worst cases	
R1 (t=1)	0.17	0.03	0.28	1.00	8.70%	
R2 (t=1)	0.16	0.03	0.24	1.00	3.64%	
R1 (t=0)	0.18	0.00	0.00	0.00	1.20%	
R2 (t=0)	0.15	0.01	0.11	1.00	1.82%	
Request	OER			OER (no ex.)		
	Mean value	Ratio of worst cases	Pair mean value	Mean value	Ratio of worst cases	Pair mean value
R1 (t=1)	0.09	76.09%	0.27	0.04	83.7%	0.18
R2 (t=1)	0.09	75.76%	0.27	0.06	81.21%	0.19
R1 (t=0)	1.00	1.20%	1.00	0.81	12.05%	0.81
R2 (t=0)	0.37	43.64%	0.59	0.27	54.55%	0.46
Request	LCS			LED		
Way	Mean value	Mean worst value	Pair mean value	Mean value	Mean worst value	Pair mean value
R1 (t=1)	0.21	0.15	0.20	61.30	82.73	63.09
R2 (t=1)	0.22	0.16	0.23	58.80	77.46	58.86
R1 (t=0)	1.00	1.00	1.00	0.00	0.00	0.00
R2 (t=0)	0.61	0.44	0.62	23.45	35.87	22.31
Request	United_Diff			Tree_Diff		
Way	Mean value	Mean worst value	Pair mean value	Mean value	Mean worst value	Pair mean value
R1 (t=1)	0.98	0.98	0.98	0.98	0.98	0.98
R2 (t=1)	0.33	0.27	0.46	0.41	0.33	0.56
R1 (t=0)	1.00	1.00	1.00	1.00	1.00	1.00
R2 (t=0)	0.41	0.39	0.67	0.50	0.46	0.74

Table 5. RQ3: Similarity for different request ways (APPS), where t represents the temperature setting.

Request Way	Test Pass Rate					
	Mean value	Mean variance	Mean max diff	Max diff	Ratio of worst cases	
R1 (t=1)	0.41	0.04	0.35	1.00	10.40%	
R2 (t=1)	0.42	0.04	0.35	1.00	10.40%	
R1 (t=0)	0.42	0.00	0.00	0.00	100.00%	
R2 (t=0)	0.43	0.01	0.14	1.00	1.80%	
Request Way	OER			OER (no ex.)		
	Mean value	Ratio of worst cases	Pair mean value	Mean value	Ratio of worst cases	Pair mean value
R1 (t=1)	0.26	55.0%	0.46	0.24	57.0%	0.41
R2 (t=1)	0.27	51.0%	0.47	0.25	53.4%	0.42
R1 (t=0)	1.00	0.2%	1.00	0.90	6.8%	0.90
R2 (t=0)	0.56	27.4%	0.73	0.50	32.8%	0.65
Request Way	LCS			LED		
	Mean value	Mean worst value	Pair mean value	Mean value	Mean worst value	Pair mean value
R1 (t=1)	0.24	0.16	0.24	47.50	62.84	47.58
R2 (t=1)	0.23	0.16	0.24	47.37	61.55	46.94
R1 (t=0)	1.00	1.00	1.00	0.00	0.00	0.00
R2 (t=0)	0.65	0.50	0.66	18.18	28.41	17.40
Request Way	United_Diff			Tree_Diff		
	Mean value	Mean worst value	Pair mean value	Mean value	Mean worst value	Pair mean value
R1 (t=1)	0.98	0.98	0.98	0.98	0.98	0.98
R2 (t=1)	0.43	0.35	0.52	0.54	0.47	0.63
R1 (t=0)	0.99	0.99	0.99	0.99	0.99	0.99
R2 (t=0)	0.49	0.46	0.70	0.60	0.57	0.77

Table 6. RQ3: Similarity for different request ways (HumanEval), where t represents the temperature setting.

Request Way	Test Pass Rate					
	Mean value	Mean variance	Mean max diff	Max diff	Ratio of worst cases	
R1 (t=1)	0.65	0.07	0.44	1.00	32.32%	
R2 (t=1)	0.63	0.09	0.53	1.00	39.63%	
R1 (t=0)	0.63	0.00	0.00	0.00	100.00%	
R2 (t=0)	0.65	0.03	0.17	1.00	14.02%	
Request Way	OER			OER (no ex.)		
	Mean value	Ratio of worst cases	Pair mean value	Mean value	Ratio of worst cases	Pair mean value
R1 (t=1)	0.48	40.24%	0.71	0.45	43.9%	0.65
R2 (t=1)	0.39	47.56%	0.67	0.35	51.22%	0.61
R1 (t=0)	0.99	0.61%	0.99	0.92	7.32%	0.92
R2 (t=0)	0.77	18.29%	0.89	0.72	23.17%	0.82
Request Way	LCS			LED		
	Mean value	Mean worst value	Pair mean value	Mean value	Mean worst value	Pair mean value
R1 (t=1)	0.43	0.26	0.41	27.73	43.86	27.74
R2 (t=1)	0.42	0.25	0.41	26.56	43.91	27.10
R1 (t=0)	0.98	0.98	0.98	0.00	0.00	0.00
R2 (t=0)	0.80	0.68	0.81	7.80	14.73	7.67
Request Way	United_Diff			Tree_Diff		
	Mean value	Mean worst value	Pair mean value	Mean value	Mean worst value	Pair mean value
R1 (t=1)	0.93	0.93	0.93	0.93	0.93	0.93
R2 (t=1)	0.60	0.47	0.67	0.62	0.48	0.70
R1 (t=0)	0.97	0.97	0.97	0.97	0.97	0.97
R2 (t=0)	0.67	0.63	0.81	0.70	0.65	0.83

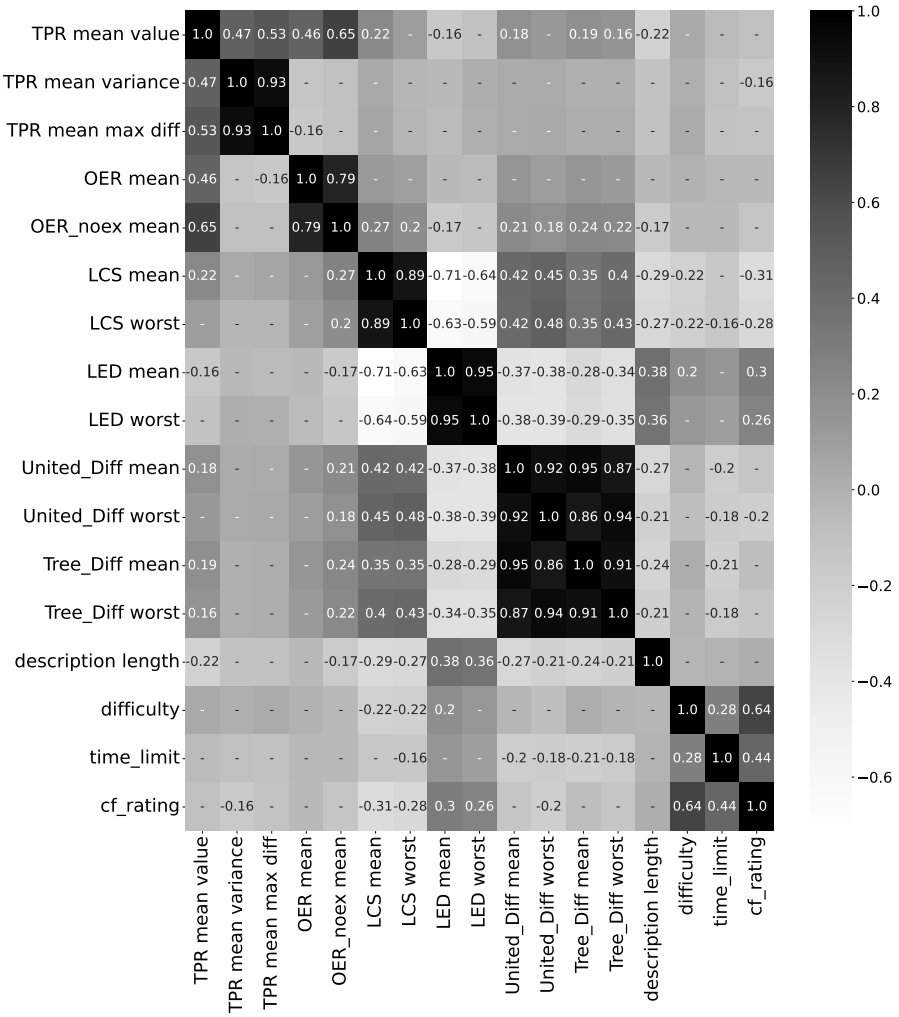

Fig. 5. RQ4: Correlations between coding tasks and non-determinism (CodeContests, temperature=1). Only significant correlations will be displayed on the heatmap, while the insignificant correlations (i.e. p-value > 0.05) are masked by '-'.

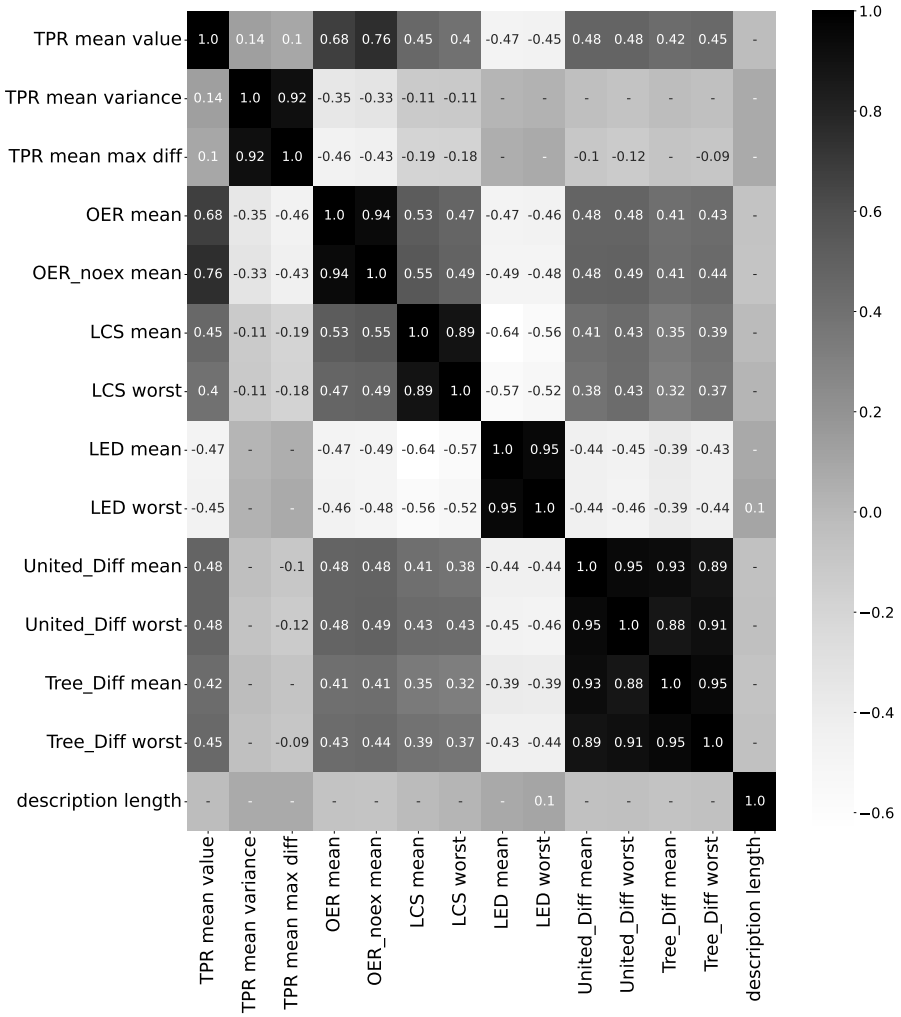

Fig. 6. RQ4: Correlations between coding tasks and non-determinism (APPS, temperature=1). Only significant correlations will be displayed on the heatmap, while the insignificant correlations (i.e. p-value > 0.05) are masked by '-'.

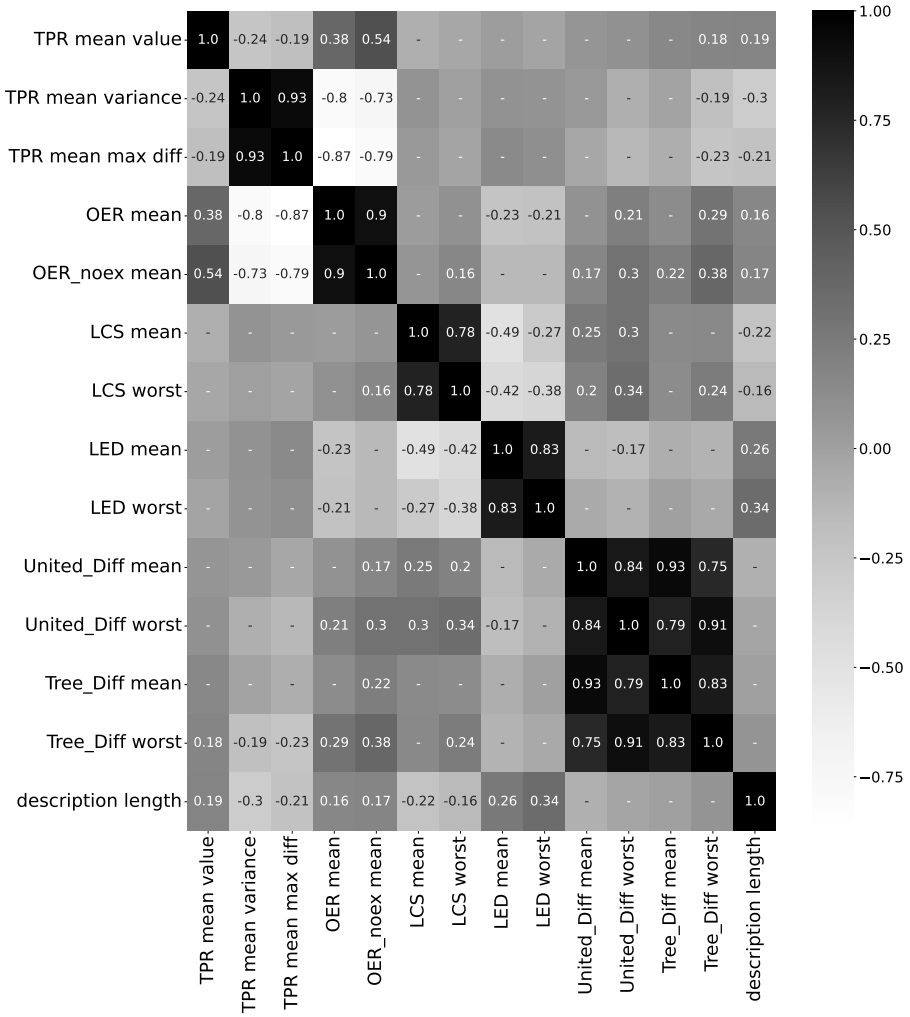

Fig. 7. RQ4: Correlations between coding tasks and non-determinism (HumanEval, temperature=1). Only significant correlations will be displayed on the heatmap, while the insignificant correlations (i.e. p-value > 0.05) are masked by '-'.

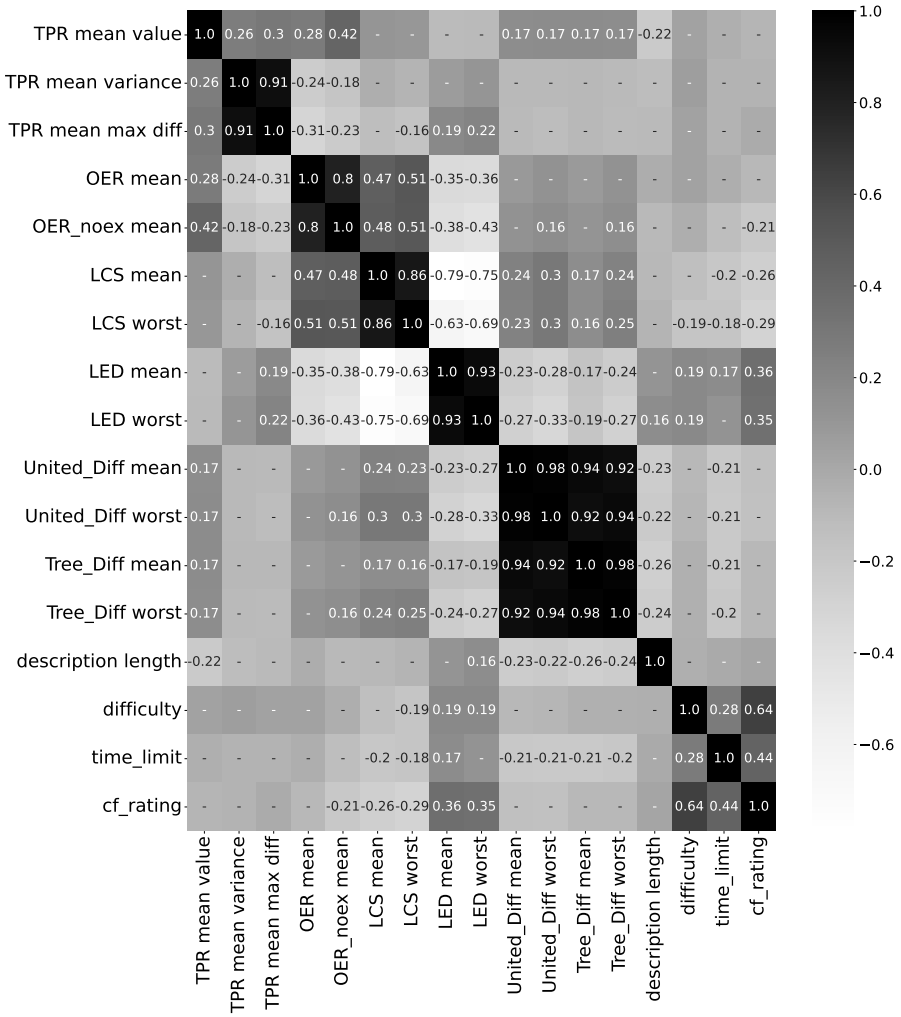

Fig. 8. RQ4: Correlations between coding tasks and non-determinism (CodeContests, temperature=0). Only significant correlations will be displayed on the heatmap, while the insignificant correlations (i.e. p-value > 0.05) are masked by '-'.

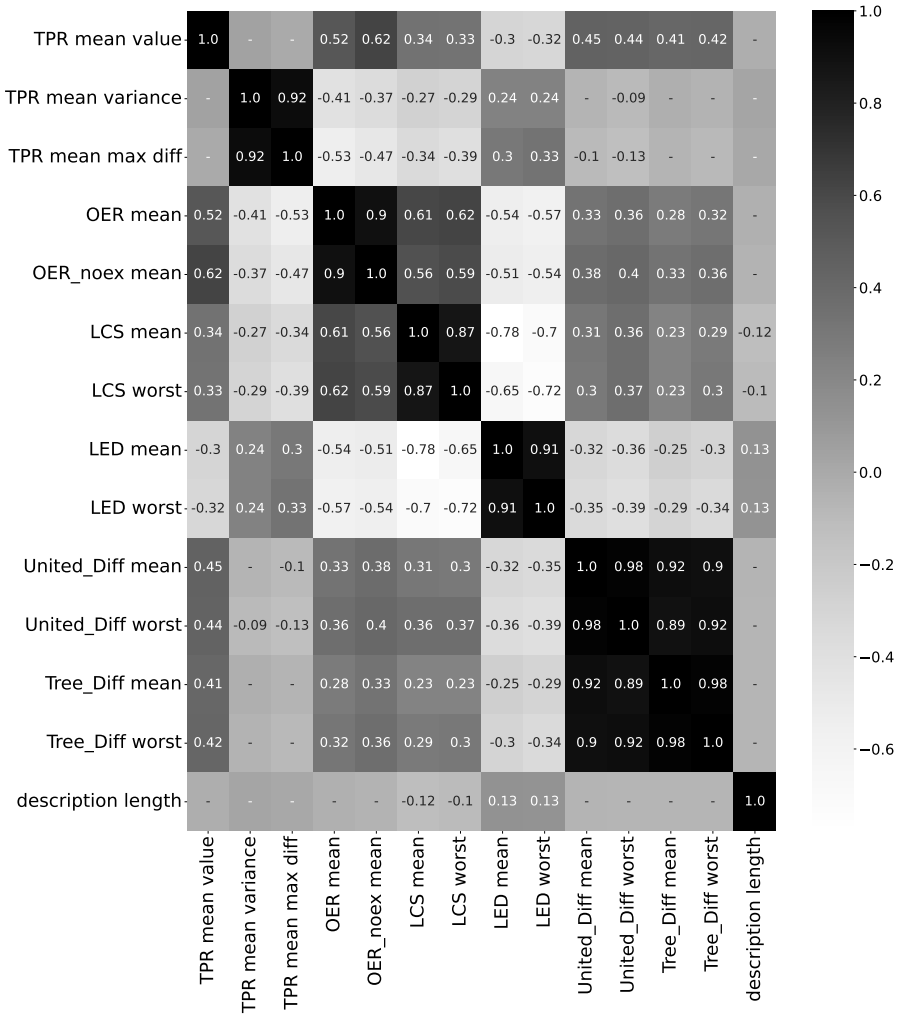

Fig. 9. RQ4: Correlations between coding tasks and non-determinism (APPS, temperature=0). Only significant correlations will be displayed on the heatmap, while the insignificant correlations (i.e. p-value > 0.05) are masked by '-'.

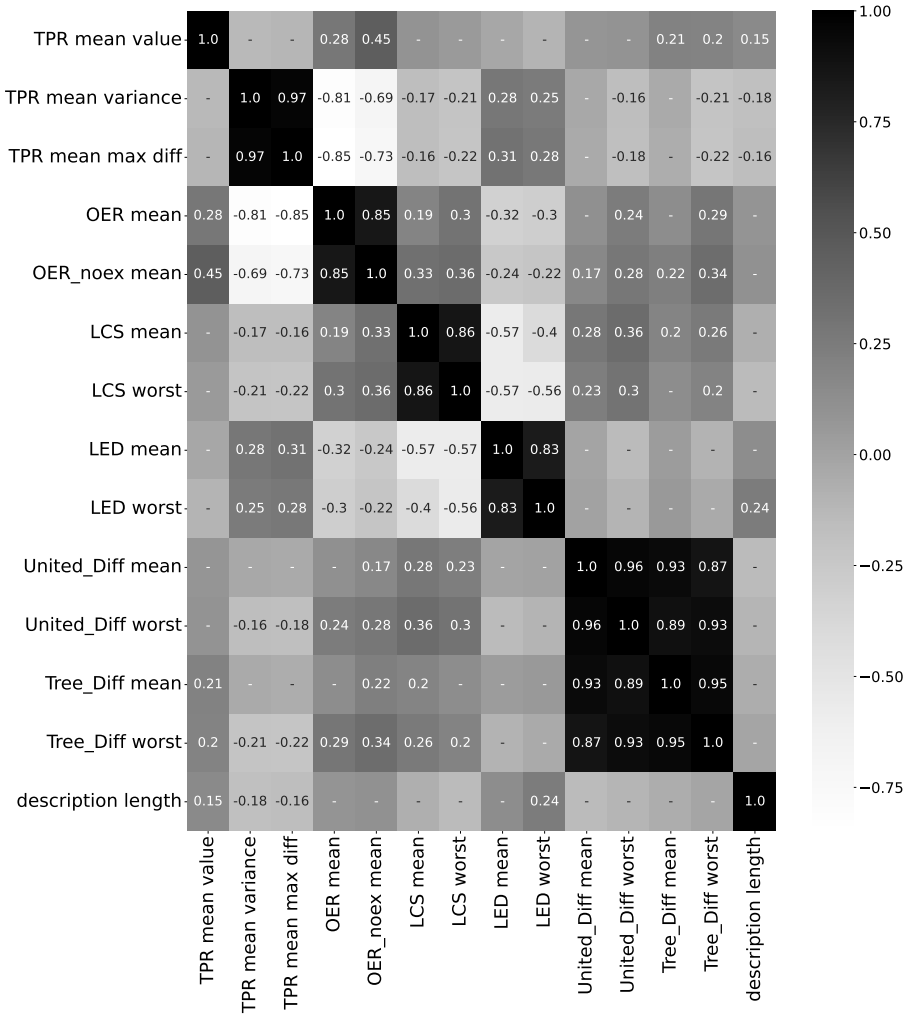

Fig. 10. RQ4: Correlations between coding tasks and non-determinism (HumanEval, temperature=0). Only significant correlations will be displayed on the heatmap, while the insignificant correlations (i.e. p-value > 0.05) are masked by '-'.

Table 7. RQ5: Non-determinism of GPT-4 v.s. GPT-3.5 (CodeContests).

Model	Test Pass Rate					
	Mean value	Mean variance	Mean max diff	Max diff	Ratio of worst cases	
GPT-4 (t=1)	0.14	0.01	0.09	1.00	1.21%	
GPT-3.5 (t=1)	0.16	0.03	0.24	1.00	3.64%	
GPT-4 (t=0)	0.14	0.01	0.08	1.00	1.21%	
GPT-3.5 (t=0)	0.15	0.01	0.11	1.00	1.82%	
Model	OER			OER (no ex.)		
	Mean value	Ratio of worst cases	Pair mean value	Mean value	Ratio of worst cases	Pair mean value
GPT-4 (t=1)	0.35	46.06%	0.58	0.25	55.76%	0.46
GPT-3.5 (t=1)	0.09	75.76%	0.27	0.06	81.21%	0.19
GPT-4 (t=0)	0.37	41.21%	0.59	0.27	52.73%	0.46
GPT-3.5 (t=0)	0.37	43.64%	0.59	0.27	54.55%	0.46
Model	LCS			LED		
	Mean value	Mean worst value	Pair mean value	Mean value	Mean worst value	Pair mean value
GPT-4 (t=1)	0.61	0.45	0.62	24.54	39.74	24.81
GPT-3.5 (t=1)	0.22	0.16	0.23	58.80	77.46	58.86
GPT-4 (t=0)	0.61	0.44	0.61	24.45	40.14	24.12
GPT-3.5 (t=0)	0.61	0.44	0.62	23.45	35.87	22.31
Model	United_Diff			Tree_Diff		
	Mean value	Mean worst value	Pair mean value	Mean value	Mean worst value	Pair mean value
GPT-4 (t=1)	0.78	0.68	0.79	0.82	0.74	0.84
GPT-3.5 (t=1)	0.33	0.27	0.46	0.41	0.33	0.56
GPT-4 (t=0)	0.78	0.68	0.79	0.83	0.75	0.84
GPT-3.5 (t=0)	0.41	0.39	0.67	0.50	0.46	0.74

Table 8. RQ5: Non-determinism of GPT-4 v.s. GPT-3.5 (APPS).

Model	Test Pass Rate					
	Mean value	Mean variance	Mean max diff	Max diff	Ratio of worst cases	
GPT-4 (t=1)	0.43	0.01	0.14	1.00	2.60%	
GPT-3.5 (t=1)	0.42	0.04	0.35	1.00	10.40%	
GPT-4 (t=0)	0.43	0.02	0.15	1.00	2.20%	
GPT-3.5 (t=0)	0.43	0.01	0.14	1.00	1.80%	
Model	OER			OER (no ex.)		
	Mean value	Ratio of worst cases	Pair mean value	Mean value	Ratio of worst cases	Pair mean value
GPT-4 (t=1)	0.54	27.6%	0.72	0.48	32.4%	0.65
GPT-3.5 (t=1)	0.27	51.0%	0.47	0.25	53.4%	0.42
GPT-4 (t=0)	0.57	25.2%	0.74	0.51	29.6%	0.66
GPT-3.5 (t=0)	0.56	27.4%	0.73	0.50	32.8%	0.65
Model	LCS			LED		
	Mean value	Mean worst value	Pair mean value	Mean value	Mean worst value	Pair mean value
GPT-4 (t=1)	0.65	0.49	0.65	19.54	30.62	18.60
GPT-3.5 (t=1)	0.23	0.16	0.24	47.37	61.55	46.94
GPT-4 (t=0)	0.67	0.51	0.67	17.05	27.95	17.04
GPT-3.5 (t=0)	0.65	0.50	0.66	18.18	28.41	17.40
Model	United_Diff			Tree_Diff		
	Mean value	Mean worst value	Pair mean value	Mean value	Mean worst value	Pair mean value
GPT-4 (t=1)	0.82	0.73	0.83	0.87	0.79	0.88
GPT-3.5 (t=1)	0.43	0.35	0.52	0.54	0.47	0.63
GPT-4 (t=0)	0.83	0.74	0.83	0.87	0.81	0.88
GPT-3.5 (t=0)	0.49	0.46	0.70	0.60	0.57	0.77

Table 9. RQ5: Non-determinism of GPT-4 v.s. GPT-3.5 (HumanEval).

Model	Test Pass Rate					
	Mean value	Mean variance	Mean max diff	Max diff	Ratio of worst cases	
GPT-4 (t=1)	0.66	0.03	0.16	1.00	11.59%	
GPT-3.5 (t=1)	0.63	0.09	0.53	1.00	39.63%	
GPT-4 (t=0)	0.65	0.02	0.13	1.00	9.15%	
GPT-3.5 (t=0)	0.65	0.03	0.17	1.00	14.02%	
Model	OER			OER (no ex.)		
	Mean value	Ratio of worst cases	Pair mean value	Mean value	Ratio of worst cases	Pair mean value
GPT-4 (t=1)	0.78	16.46%	0.89	0.73	21.34%	0.83
GPT-3.5 (t=1)	0.39	47.56%	0.67	0.35	51.22%	0.61
GPT-4 (t=0)	0.81	13.41%	0.90	0.75	18.9%	0.84
GPT-3.5 (t=0)	0.77	18.29%	0.89	0.72	23.17%	0.82
Model	LCS			LED		
	Mean value	Mean worst value	Pair mean value	Mean value	Mean worst value	Pair mean value
GPT-4 (t=1)	0.78	0.65	0.79	8.95	17.85	9.23
GPT-3.5 (t=1)	0.42	0.25	0.41	26.56	43.91	27.10
GPT-4 (t=0)	0.81	0.69	0.82	8.28	14.79	8.30
GPT-3.5 (t=0)	0.80	0.68	0.81	7.80	14.73	7.67
Model	United_Diff			Tree_Diff		
	Mean value	Mean worst value	Pair mean value	Mean value	Mean worst value	Pair mean value
GPT-4 (t=1)	0.89	0.83	0.90	0.91	0.85	0.91
GPT-3.5 (t=1)	0.60	0.47	0.67	0.62	0.48	0.70
GPT-4 (t=0)	0.91	0.86	0.91	0.92	0.87	0.92
GPT-3.5 (t=0)	0.67	0.63	0.81	0.70	0.65	0.83

Table 10. RQ6: Prompt engineering techniques (CodeContests).

Prompt	Test Pass Rate					
	Mean value	Mean variance	Mean max diff	Max diff	Ratio of worst cases	
Concise (t=1)	0.15	0.02	0.19	1.00	3.64%	
Base (t=1)	0.16	0.03	0.24	1.00	3.64%	
CoT (t=1)	0.15	0.02	0.19	1.00	3.64%	
Concise (t=0)	0.16	0.01	0.10	1.00	0.61%	
Base (t=0)	0.15	0.01	0.11	1.00	1.82%	
CoT (t=0)	0.19	0.02	0.15	1.00	1.82%	
Prompt	OER			OER (no ex.)		
	Mean value	Ratio of worst cases	Pair mean value	Mean value	Ratio of worst cases	Pair mean value
Concise (t=1)	0.10	76.36%	0.26	0.06	81.82%	0.17
Base (t=1)	0.09	75.76%	0.27	0.06	81.21%	0.19
CoT (t=1)	0.10	73.94%	0.26	0.08	80.0%	0.19
Concise (t=0)	0.39	41.82%	0.63	0.31	49.09%	0.54
Base (t=0)	0.37	43.64%	0.59	0.27	54.55%	0.46
CoT (t=0)	0.28	46.06%	0.50	0.19	54.55%	0.36
Prompt	LCS			LED		
	Mean value	Mean worst value	Pair mean value	Mean value	Mean worst value	Pair mean value
Concise (t=1)	0.22	0.16	0.22	61.53	83.01	62.52
Base (t=1)	0.22	0.16	0.23	58.80	77.46	58.86
CoT (t=1)	0.23	0.15	0.23	59.55	77.68	57.05
Concise (t=0)	0.70	0.53	0.71	11.77	20.55	12.14
Base (t=0)	0.61	0.44	0.62	23.45	35.87	22.31
CoT (t=0)	0.38	0.24	0.39	39.31	58.28	39.81
Prompt	United_Diff			Tree_Diff		
	Mean value	Mean worst value	Pair mean value	Mean value	Mean worst value	Pair mean value
Concise (t=1)	0.44	0.34	0.48	0.54	0.42	0.59
Base (t=1)	0.33	0.27	0.46	0.41	0.33	0.56
CoT (t=1)	0.45	0.35	0.51	0.55	0.43	0.61
Concise (t=0)	0.83	0.74	0.84	0.88	0.82	0.89
Base (t=0)	0.41	0.39	0.67	0.50	0.46	0.74
CoT (t=0)	0.71	0.58	0.72	0.78	0.67	0.79

Table 11. RQ6: Prompt engineering techniques (APPS).

Request Complexity	Test Pass Rate					
	Mean value	Mean variance	Mean max diff	Max diff	Ratio of worst cases	
Concise (t=1)	0.41	0.04	0.35	1.00	10.00%	
Base (t=1)	0.42	0.04	0.35	1.00	10.40%	
CoT (t=1)	0.42	0.04	0.33	1.00	8.40%	
Concise (t=0)	0.38	0.01	0.13	1.00	2.60%	
Base (t=0)	0.43	0.01	0.14	1.00	1.80%	
CoT (t=0)	0.43	0.02	0.21	1.00	4.20%	
Request Complexity	OER			OER (no ex.)		
	Mean value	Ratio of worst cases	Pair mean value	Mean value	Ratio of worst cases	Pair mean value
Concise (t=1)	0.26	54.8%	0.46	0.23	57.0%	0.41
Base (t=1)	0.27	51.0%	0.47	0.25	53.4%	0.42
CoT (t=1)	0.27	51.2%	0.47	0.25	53.8%	0.42
Concise (t=0)	0.58	24.4%	0.75	0.51	31.4%	0.66
Base (t=0)	0.56	27.4%	0.73	0.50	32.8%	0.65
CoT (t=0)	0.43	34.4%	0.62	0.37	39.4%	0.54
Request Complexity	LCS			LED		
	Mean value	Mean worst value	Pair mean value	Mean value	Mean worst value	Pair mean value
Concise (t=1)	0.24	0.16	0.23	48.84	63.96	48.58
Base (t=1)	0.23	0.16	0.24	47.37	61.55	46.94
CoT (t=1)	0.24	0.16	0.24	47.12	61.19	46.77
Concise (t=0)	0.73	0.58	0.73	10.17	17.15	10.14
Base (t=0)	0.65	0.50	0.66	18.18	28.41	17.40
CoT (t=0)	0.40	0.25	0.40	35.21	52.66	35.75
Request Complexity	United_Diff			Tree_Diff		
	Mean value	Mean worst value	Pair mean value	Mean value	Mean worst value	Pair mean value
Concise (t=1)	0.54	0.42	0.56	0.65	0.53	0.67
Base (t=1)	0.43	0.35	0.52	0.54	0.47	0.63
CoT (t=1)	0.55	0.43	0.57	0.65	0.53	0.68
Concise (t=0)	0.83	0.74	0.84	0.87	0.81	0.89
Base (t=0)	0.49	0.46	0.70	0.60	0.57	0.77
CoT (t=0)	0.73	0.61	0.73	0.81	0.71	0.81

Table 12. RQ6: Prompt engineering techniques (HumanEval).

Request Complexity	Test Pass Rate				
	Mean value	Mean variance	Mean max diff	Max diff	Ratio of worst cases
Concise (t=1)	0.63	0.08	0.47	1.00	34.15%
Base (t=1)	0.63	0.09	0.53	1.00	39.63%
CoT (t=1)	0.65	0.08	0.48	1.00	38.41%
Concise (t=0)	0.69	0.02	0.11	1.00	6.10%
Base (t=0)	0.65	0.03	0.17	1.00	14.02%
CoT (t=0)	0.84	0.01	0.11	1.00	4.27%

Request Complexity	OER			OER (no ex.)		
	Mean value	Ratio of worst cases	Pair mean value	Mean value	Ratio of worst cases	Pair mean value
Concise (t=1)	0.43	45.73%	0.67	0.40	48.17%	0.62
Base (t=1)	0.39	47.56%	0.67	0.35	51.22%	0.61
CoT (t=1)	0.43	48.78%	0.68	0.40	51.83%	0.63
Concise (t=0)	0.85	10.37%	0.92	0.77	17.07%	0.84
Base (t=0)	0.77	18.29%	0.89	0.72	23.17%	0.82
CoT (t=0)	0.84	8.54%	0.92	0.83	9.15%	0.90

Request Complexity	LCS			LED		
	Mean value	Mean worst value	Pair mean value	Mean value	Mean worst value	Pair mean value
Concise (t=1)	0.38	0.24	0.40	28.45	44.69	28.27
Base (t=1)	0.42	0.25	0.41	26.56	43.91	27.10
CoT (t=1)	0.40	0.25	0.40	29.31	44.91	29.31
Concise (t=0)	0.88	0.80	0.89	1.65	3.69	1.81
Base (t=0)	0.80	0.68	0.81	7.80	14.73	7.67
CoT (t=0)	0.67	0.52	0.70	12.55	21.18	12.03

Request Complexity	United_Diff			Tree_Diff		
	Mean value	Mean worst value	Pair mean value	Mean value	Mean worst value	Pair mean value
Concise (t=1)	0.69	0.56	0.70	0.72	0.61	0.74
Base (t=1)	0.60	0.47	0.67	0.62	0.48	0.70
CoT (t=1)	0.65	0.53	0.69	0.69	0.56	0.73
Concise (t=0)	0.93	0.88	0.93	0.94	0.91	0.95
Base (t=0)	0.67	0.63	0.81	0.70	0.65	0.83
CoT (t=0)	0.91	0.83	0.91	0.93	0.87	0.93