1 EVALUATION

1.1 *RQ4: Working with Various LLMs*

Besides the underlying LLMs, we also evaluated how the length of functional descriptions and the length of method names would influence the performance of the proposed approach. Our evaluation results suggest that the proposed approach worked well on various methods, regardless of the length of method names or functional descriptions. Tables 1 and 2 illustrate the performance of ContextCraft across different lengths of *JavaData* method names and functional descriptions, respectively. Tables 3 and 4 illustrate the performance of ContextCraft across different lengths of *PythonData* method names and functional descriptions, respectively.

Table 1. Impact of Method Name's Length of *JavaData* (with different underlying LLMs)

Length of Method Name	**ChatGPT-4o**			**Gemini**			**Llama3**		
	EM	**SSI**	**ED**	**EM**	**SSI**	**ED**	**EM**	**SSI**	**ED**
1	76.23%	80.79%	2.23	70.16%	80.40%	2.54	69.28%	79.68%	2.26
2	78.54%	84.80%	2.20	68.83%	80.68%	3.87	73.21%	77.72%	3.22
3	77.53%	84.85%	2.72	74.87%	77.10%	3.38	71.71%	81.31%	2.73
4	74.62%	80.83%	2.55	68.97%	75.82%	3.34	67.16%	76.48%	3.57
5	74.23%	83.94%	2.78	80.68%	80.03%	2.15	66.31%	74.64%	3.48
6+	72.28%	82.80%	3.05	71.26%	72.62%	3.55	69.30%	79.73%	3.22

Table 2. Impact of Length of Functional Descriptions of *JavaData* (with different underlying LLMs)

Length of Description	**ChatGPT-4o**			**Gemini**			**Llama3**		
	EM	**SSI**	**ED**	**EM**	**SSI**	**ED**	**EM**	**SSI**	**ED**
0-5	76.37%	84.52%	3.10	72.88%	79.58%	3.62	69.23%	80.15%	3.27
6-9	74.89%	84.07%	3.10	72.09%	79.47%	3.65	71.12%	79.96%	3.23
10-13	75.51%	84.20%	3.12	72.23%	79.61%	3.70	70.03%	80.65%	3.25
14-17	75.62%	84.10%	3.08	72.45%	79.85%	3.69	70.01%	80.70%	3.24
18-21	75.00%	84.12%	3.09	72.30%	79.65%	3.68	70.30%	80.55%	3.26
22-25	74.79%	84.25%	3.11	72.12%	79.48%	3.66	70.70%	80.75%	3.24
26-29	75.65%	84.32%	3.15	72.50%	79.70%	3.68	69.90%	80.60%	3.27
30-33	75.41%	84.10%	3.11	72.55%	79.72%	3.66	69.80%	80.70%	3.26
34-37	75.10%	84.00%	3.08	72.40%	79.60%	3.68	70.20%	80.70%	3.27
40+	75.18%	84.25%	3.12	72.33%	79.75%	3.67	69.80%	80.80%	3.26

Table 3. Impact of Method Name's Length of *PythonData* (with different underlying LLMs)

Name's Length	ChatGPT-4o			Gemini			Llama3		
	EM	SSI	ED	EM	SSI	ED	EM	SSI	ED
1	75.15%	79.87%	2.30	71.41%	80.67%	2.62	68.67%	78.50%	2.23
2	77.45%	83.53%	2.28	69.66%	79.26%	3.80	70.99%	76.92%	3.18
3	76.44%	83.63%	2.77	74.09%	77.68%	3.40	71.89%	80.24%	2.70
4	73.67%	79.90%	2.59	69.25%	75.66%	3.41	66.48%	76.28%	3.61
5	72.94%	82.96%	2.85	81.57%	79.00%	2.18	65.58%	74.09%	3.45
6+	71.29%	81.83%	3.10	71.00%	72.92%	3.62	68.06%	78.56%	3.22

Table 4. Impact of length of Functional Descriptions of *PythonData* (with different underlying LLMs)

Length of Description	ChatGPT-4o			Gemini			Llama3		
	EM	SSI	ED	EM	SSI	ED	EM	SSI	ED
0-5	72.98%	80.51%	2.54	72.84%	84.18%	2.53	70.15%	77.67%	2.62
6-9	73.61%	81.47%	2.60	71.77%	79.12%	1.82	78.12%	78.22%	2.15
10-13	73.40%	83.02%	3.37	71.26%	80.62%	3.45	74.62%	75.21%	3.88
14-17	75.58%	84.96%	3.41	72.13%	83.80%	3.42	79.83%	80.85%	2.73
18-21	74.03%	84.64%	3.31	70.88%	78.63%	3.69	81.35%	82.58%	2.82
22-25	72.25%	83.04%	3.19	70.42%	79.69%	3.36	77.44%	75.93%	3.29
26-29	73.02%	83.73%	3.29	69.20%	78.87%	3.64	75.63%	75.76%	3.31
30-33	74.59%	82.68%	3.51	70.14%	79.71%	3.57	80.46%	77.86%	3.68
34-37	73.89%	83.67%	3.55	71.65%	78.18%	3.46	75.00%	77.32%	3.52
40+	71.30%	82.60%	3.42	68.78%	80.25%	3.71	78.58%	80.73%	3.32