

Table 1. Result of SentenceBert

Defense	No Defense	Reminder (Yi et al., 2023)	In-Context (Wei et al., 2023)	Isolation (Willison, 2024)	SysVec (Ours)
Model	Llama-2-7B-chat-hf				
Naive Attack (Liu et al., 2024)	0.17 $\pm$ 0.08	0.16 $\pm$ 0.07	0.16 $\pm$ 0.08	0.16 $\pm$ 0.07	<b>0.16</b> $\pm$ 0.06
Ignore Attack (Chen et al., 2024)	0.18 $\pm$ 0.09	0.16 $\pm$ 0.07	0.17 $\pm$ 0.08	0.17 $\pm$ 0.07	<b>0.15</b> $\pm$ 0.06
Completion Attack (Willison, 2024)	0.19 $\pm$ 0.08	0.17 $\pm$ 0.07	0.17 $\pm$ 0.09	0.18 $\pm$ 0.08	<b>0.16</b> $\pm$ 0.06
Ignore-Completion Attack	0.19 $\pm$ 0.08	0.18 $\pm$ 0.07	0.17 $\pm$ 0.07	0.19 $\pm$ 0.07	<b>0.16</b> $\pm$ 0.06
Remember-the-Start Attack	0.27 $\pm$ 0.12	0.26 $\pm$ 0.11	0.28 $\pm$ 0.13	0.27 $\pm$ 0.11	<b>0.16</b> $\pm$ 0.07
Ignore-Remember Attack	0.47 $\pm$ 0.16	0.41 $\pm$ 0.14	0.41 $\pm$ 0.15	0.37 $\pm$ 0.14	<b>0.33</b> $\pm$ 0.12
Completion-Remember Attack	0.30 $\pm$ 0.12	0.30 $\pm$ 0.11	0.27 $\pm$ 0.11	0.24 $\pm$ 0.11	<b>0.18</b> $\pm$ 0.07
Ignore-Completion-Remember Attack	0.49 $\pm$ 0.09	0.48 $\pm$ 0.10	0.48 $\pm$ 0.10	0.47 $\pm$ 0.11	<b>0.40</b> $\pm$ 0.12
PLeak (Hui et al., 2024)	0.49 $\pm$ 0.13	0.49 $\pm$ 0.13	0.49 $\pm$ 0.13	0.48 $\pm$ 0.15	<b>0.16</b> $\pm$ 0.06
Model	Llama-3-8B-Instruct				
Naive Attack (Liu et al., 2024)	0.17 $\pm$ 0.07	0.14 $\pm$ 0.07	0.14 $\pm$ 0.07	0.15 $\pm$ 0.07	<b>0.14</b> $\pm$ 0.08
Ignore Attack (Chen et al., 2024)	0.17 $\pm$ 0.07	0.16 $\pm$ 0.07	0.16 $\pm$ 0.07	0.17 $\pm$ 0.07	<b>0.15</b> $\pm$ 0.06
Completion Attack (Willison, 2024)	0.19 $\pm$ 0.07	0.15 $\pm$ 0.07	0.15 $\pm$ 0.07	0.17 $\pm$ 0.08	<b>0.14</b> $\pm$ 0.07
Ignore-Completion Attack	0.22 $\pm$ 0.06	0.22 $\pm$ 0.06	0.22 $\pm$ 0.06	0.22 $\pm$ 0.07	<b>0.15</b> $\pm$ 0.07
Remember-the-Start Attack	0.26 $\pm$ 0.08	0.26 $\pm$ 0.08	0.25 $\pm$ 0.08	0.26 $\pm$ 0.07	<b>0.15</b> $\pm$ 0.07
Ignore-Remember Attack	0.63 $\pm$ 0.30	0.63 $\pm$ 0.30	0.63 $\pm$ 0.28	0.32 $\pm$ 0.09	<b>0.30</b> $\pm$ 0.09
Completion-Remember Attack	0.21 $\pm$ 0.07	0.25 $\pm$ 0.05	0.25 $\pm$ 0.05	0.24 $\pm$ 0.05	<b>0.16</b> $\pm$ 0.07
Ignore-Completion-Remember Attack	0.98 $\pm$ 0.02	0.98 $\pm$ 0.02	0.98 $\pm$ 0.02	0.72 $\pm$ 0.32	<b>0.37</b> $\pm$ 0.10
PLeak (Hui et al., 2024)	0.69 $\pm$ 0.23	0.71 $\pm$ 0.27	0.71 $\pm$ 0.29	0.65 $\pm$ 0.19	<b>0.14</b> $\pm$ 0.06
Model	Mistral-7B-Instruct				
Naive Attack (Liu et al., 2024)	0.16 $\pm$ 0.07	0.16 $\pm$ 0.07	0.15 $\pm$ 0.06	0.16 $\pm$ 0.07	<b>0.14</b> $\pm$ 0.07
Ignore Attack (Chen et al., 2024)	0.16 $\pm$ 0.07	0.16 $\pm$ 0.07	<b>0.15</b> $\pm$ 0.06	0.16 $\pm$ 0.07	0.16 $\pm$ 0.07
Completion Attack (Willison, 2024)	0.17 $\pm$ 0.07	0.17 $\pm$ 0.07	0.15 $\pm$ 0.06	0.17 $\pm$ 0.07	<b>0.14</b> $\pm$ 0.07
Ignore-Completion Attack	0.18 $\pm$ 0.07	0.18 $\pm$ 0.07	0.16 $\pm$ 0.06	0.18 $\pm$ 0.07	<b>0.14</b> $\pm$ 0.07
Remember-the-Start Attack	0.35 $\pm$ 0.11	0.35 $\pm$ 0.10	0.28 $\pm$ 0.08	0.35 $\pm$ 0.12	<b>0.14</b> $\pm$ 0.07
Ignore-Remember Attack	0.37 $\pm$ 0.13	0.37 $\pm$ 0.13	<b>0.28</b> $\pm$ 0.08	0.39 $\pm$ 0.13	0.35 $\pm$ 0.16
Completion-Remember Attack	0.31 $\pm$ 0.06	0.31 $\pm$ 0.06	0.21 $\pm$ 0.07	0.35 $\pm$ 0.11	<b>0.17</b> $\pm$ 0.08
Ignore-Completion-Remember Attack	0.39 $\pm$ 0.09	0.39 $\pm$ 0.09	<b>0.28</b> $\pm$ 0.08	0.40 $\pm$ 0.11	0.35 $\pm$ 0.15
PLeak (Hui et al., 2024)	0.58 $\pm$ 0.17	0.58 $\pm$ 0.17	0.58 $\pm$ 0.17	0.54 $\pm$ 0.13	<b>0.16</b> $\pm$ 0.10

Table 2. Extra Defenses

Model	Llama-3-8B-Instruct				
Defense	No Defense	PPL	Paraphrasing	SFT	SysVec (Ours)
Naive Attack (Liu et al., 2024)	4.16 $\pm$ 1.73	4.07 $\pm$ 1.86	4.13 $\pm$ 1.68	<b>2.76</b> $\pm$ 1.30	2.83 $\pm$ 1.36
Ignore Attack (Chen et al., 2024)	4.83 $\pm$ 1.82	4.28 $\pm$ 2.61	4.91 $\pm$ 1.64	3.05 $\pm$ 1.53	<b>3.01</b> $\pm$ 1.48
Completion Attack (Willison, 2024)	4.12 $\pm$ 1.64	4.32 $\pm$ 1.97	4.30 $\pm$ 1.50	3.03 $\pm$ 1.34	<b>2.88</b> $\pm$ 1.47
Ignore-Completion Attack	4.29 $\pm$ 1.54	4.37 $\pm$ 1.61	4.16 $\pm$ 1.16	3.07 $\pm$ 1.23	<b>3.05</b> $\pm$ 1.51
Remember-the-Start Attack	7.21 $\pm$ 1.68	7.21 $\pm$ 1.19	7.01 $\pm$ 1.86	3.52 $\pm$ 1.87	<b>3.21</b> $\pm$ 1.61
Ignore-Remember Attack	8.43 $\pm$ 1.12	7.98 $\pm$ 1.60	7.95 $\pm$ 1.26	<b>3.41</b> $\pm$ 1.26	3.56 $\pm$ 1.69
Completion-Remember Attack	5.42 $\pm$ 1.82	5.49 $\pm$ 1.74	5.61 $\pm$ 1.82	<b>3.20</b> $\pm$ 1.57	3.30 $\pm$ 1.45
Ignore-Completion-Remember Attack	7.23 $\pm$ 1.21	7.05 $\pm$ 1.50	7.02 $\pm$ 1.85	3.78 $\pm$ 1.50	<b>3.45</b> $\pm$ 1.62
PLeak (Hui et al., 2024)	7.53 $\pm$ 1.84	4.67 $\pm$ 2.55	7.14 $\pm$ 1.67	<b>1.52</b> $\pm$ 0.96	2.02 $\pm$ 1.86

Table 3. Result of Updated Completion Attack and PLeak

Defense	No Defense	Reminder (Yi et al., 2023)	In-Context (Wei et al., 2023)	Isolation (Willison, 2024)	SysVec (Ours)
Llama-2-7B-chat-hf					
Completion Attack (Willison, 2024)	5.07 $\pm$ 1.67	5.05 $\pm$ 1.83	5.03 $\pm$ 1.69	4.75 $\pm$ 1.80	<b>3.15</b> $\pm$ 1.70
Ignore-Completion Attack	5.77 $\pm$ 1.71	5.46 $\pm$ 1.84	5.54 $\pm$ 1.77	5.48 $\pm$ 1.73	<b>3.44</b> $\pm$ 1.80
Completion-Remember Attack	3.62 $\pm$ 1.80	3.52 $\pm$ 1.83	3.49 $\pm$ 1.71	3.33 $\pm$ 1.93	<b>3.23</b> $\pm$ 1.58
Ignore-Completion-Remember Attack	5.33 $\pm$ 2.25	5.27 $\pm$ 2.22	5.39 $\pm$ 2.21	4.84 $\pm$ 2.40	<b>3.03</b> $\pm$ 1.66
PLeak (Hui et al., 2024)	6.14 $\pm$ 2.21	6.38 $\pm$ 2.08	6.30 $\pm$ 2.66	6.51 $\pm$ 2.82	<b>1.32</b> $\pm$ 0.93
Llama-3-8B-Instruct					
Completion Attack (Willison, 2024)	4.12 $\pm$ 1.64	4.04 $\pm$ 1.63	4.10 $\pm$ 1.66	3.82 $\pm$ 1.53	<b>2.88</b> $\pm$ 1.47
Ignore-Completion Attack	4.29 $\pm$ 1.54	4.30 $\pm$ 1.70	4.27 $\pm$ 1.70	4.16 $\pm$ 1.59	<b>3.05</b> $\pm$ 1.51
Completion-Remember Attack	5.42 $\pm$ 1.82	5.39 $\pm$ 1.82	5.49 $\pm$ 1.94	5.42 $\pm$ 1.75	<b>3.30</b> $\pm$ 1.45
Ignore-Completion-Remember Attack	7.23 $\pm$ 1.21	7.17 $\pm$ 1.18	7.20 $\pm$ 1.14	7.25 $\pm$ 1.36	<b>3.45</b> $\pm$ 1.62
PLeak (Hui et al., 2024)	7.53 $\pm$ 1.84	7.49 $\pm$ 1.68	7.51 $\pm$ 1.79	7.33 $\pm$ 1.86	<b>2.02</b> $\pm$ 1.86
Mistral-7B-Instruct					
Completion Attack (Willison, 2024)	5.85 $\pm$ 1.80	5.84 $\pm$ 1.75	5.83 $\pm$ 1.85	5.49 $\pm$ 1.86	<b>3.68</b> $\pm$ 1.40
Ignore-Completion Attack	5.08 $\pm$ 1.80	5.09 $\pm$ 1.82	4.93 $\pm$ 1.77	4.63 $\pm$ 1.83	<b>3.73</b> $\pm$ 1.44
Completion-Remember Attack	7.21 $\pm$ 1.19	7.13 $\pm$ 1.39	7.19 $\pm$ 1.30	6.44 $\pm$ 1.39	<b>3.42</b> $\pm$ 1.47
Ignore-Completion-Remember Attack	7.44 $\pm$ 0.98	7.50 $\pm$ 0.94	7.48 $\pm$ 0.96	7.0 $\pm$ 1.33	<b>3.68</b> $\pm$ 1.51
PLeak (Hui et al., 2024)	7.19 $\pm$ 1.94	7.03 $\pm$ 2.20	7.15 $\pm$ 2.12	6.79 $\pm$ 1.59	<b>1.40</b> $\pm$ 0.67

Table 4. Time Costs.

Application	D&D	Paimon	ML	Advisor	Stoic
Textual system prompt Avg. inference time (s/query), max_new_tokens=1	0.083	0.053	0.082	0.054	0.106
SysVec Avg. inference time (s/query), max_new_tokens=1	0.015	0.016	0.015	0.015	0.015
Textual system Prompt Avg. inference time (s/query), max_new_tokens=4096	1.730	1.946	2.890	4.573	5.166
SysVec Avg. inference time (s/query), max_new_tokens=4096	1.176	1.461	2.233	2.566	2.784
SysVec training time (s)	1540.2	1555.2	1655.2	1700.1	1705.1
Cost-benefit balance threshold, max_new_tokens=1, (worst case)	22650	40926	24705	43593	18738
Cost-benefit balance threshold, max_new_tokens=4096 (normal case)	2781	3207	2520	848	716