

	Prompt					Prompt				
	A	B	C	D	E	A	B	C	D	E
Paraphrase Error (%)										
GPT-3.5	11.80061038	9.359104781	10.78331638	12.41086678	9.15564582	88.19538562	90.64089522	89.21668362	87.56501322	90.84435402
GPT-4	8.138351984	6.103763888	6.612410987	6.612410987	3.357070193	91.86164802	93.89623601	93.38758901	93.38758901	96.64252981
GPT-4o	3.967446592	3.763687792	3.967446592	3.662258393	2.543234995	96.03285341	96.23601221	96.03285341	96.33774161	97.45676501
Paraphrase Accuracy (%)										
GPT-3.5	36.92777213	36.52058453	37.63987792	37.63987792	31.33265514	63.07222787	63.47914547	62.36012208	62.36012208	68.66734486
GPT-4	22.17700916	23.39776195	24.82197355	23.80467955	19.83723296	77.82299084	76.60223805	75.17802645	76.19532045	80.16276704
GPT-4o	16.48016277	24.00813835	21.26144456	16.78535097	18.10783316	83.51983723	79.99186165	78.73855544	83.21464903	81.89216884
Paraphrase F1 Score										
GPT-3.5	0.385	0.398	0.381	0.374	0.446	0.385	0.398	0.384	0.378	0.45
GPT-4	0.427	0.431	0.428	0.429	0.53	0.427	0.431	0.428	0.429	0.53
GPT-4o	0.456	0.447	0.448	0.446	0.522	0.456	0.447	0.448	0.446	0.522
Paraphrase SNE										
GPT-3.5	0.129	0.134	0.125	0.121	0.213	0.129	0.134	0.125	0.121	0.213
GPT-4	0.138	0.138	0.137	0.137	0.283	0.138	0.138	0.137	0.137	0.283
GPT-4o	0.167	0.156	0.164	0.157	0.271	0.167	0.156	0.164	0.157	0.271
Paraphrase SME										
GPT-3.5	0.362	0.395	0.381	0.374	0.446	0.362	0.395	0.381	0.374	0.446
GPT-4	0.425	0.428	0.425	0.427	0.526	0.425	0.428	0.425	0.427	0.526
GPT-4o	0.452	0.443	0.445	0.443	0.517	0.452	0.443	0.445	0.443	0.517
Paraphrase Inference Time (s)										
GPT-3.5	0.8416138889	1.743335	1.143433889	2.212084722	4.971445	0.8416138889	1.743335	1.143433889	2.212084722	4.971445
GPT-4	1.455085	4.158043611	4.345294444	5.164868889	6.631765278	1.455085	4.158043611	4.345294444	5.164868889	6.631765278
GPT-4o	0.7398538889	1.689626111	1.731704722	2.470251667	5.404273889	0.7398538889	1.689626111	1.731704722	2.470251667	5.404273889
Paraphrase Inference Cost (\$)										
GPT-3.5	1.933	2.96	3.039	3.466	5.686	1.933	2.96	3.039	3.466	5.686
GPT-4	20.232	37.416	38.161	45.144	66.74	20.232	37.416	38.161	45.144	66.74
GPT-4o	3.44	7.523	7.653	9.262	12.842	3.44	7.523	7.653	9.262	12.842

#Records 983

		Weaker	Equal	Stronger
A	GPT-3	5	27	39
	GPT-4	3	27	103
	GPT-4o	11	43	85
B	GPT-3	5	24	52
	GPT-4	4	24	113
	GPT-4o	5	35	99
C	GPT-3	7	23	45
	GPT-4	4	24	108
	GPT-4o	8	44	94
D	GPT-3	4	17	56
	GPT-4	4	27	112
	GPT-4o	8	41	79
E	GPT-3	8	103	58
	GPT-4	12	144	97
	GPT-4o	24	137	63

