

Standard Error of Experimental Results

Figure 1: Standard Error of BLEU scores in algorithm design and oracle construction.

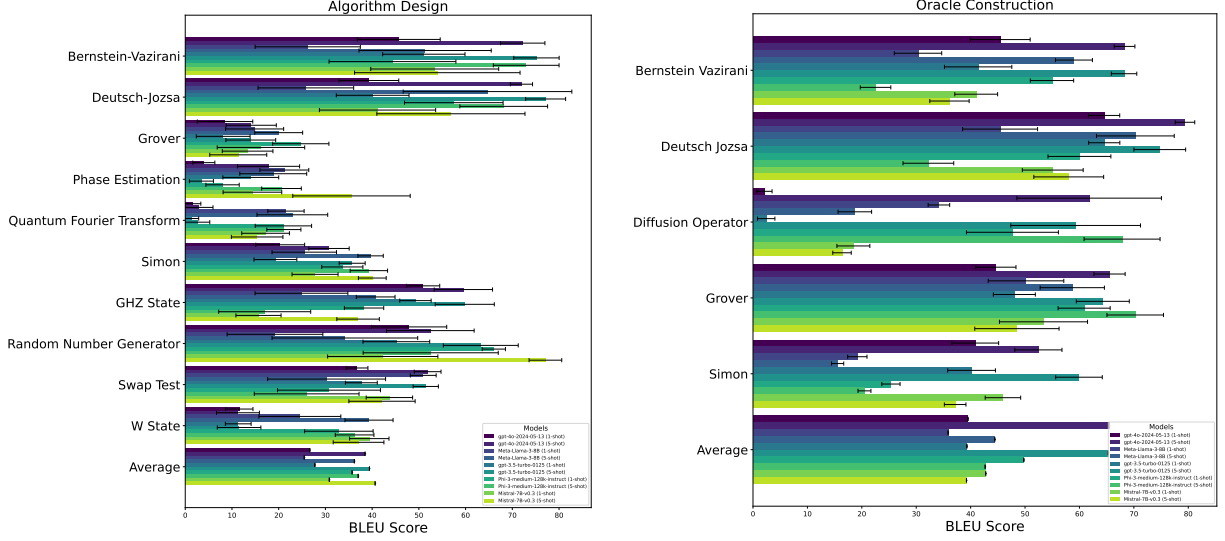


Table 1: Standard Error of verification function scores in algorithm design.

Model	Shot	Bernstein-Vazirani	Deutsch-Jozsa	Grover	Phase Estimation	Quantum Fourier Transform	Simon	GHZ State	Random Number Generator	Swap Test	W State	Average
gpt-4o-2024-05-13	1	-0.8462 (± 0.1042)	-0.5538 (± 0.1986)	-0.7089 (± 0.1879)	-1.0000 (± 0.0000)	-1.0000 (± 0.0000)	-0.6692 (± 0.1447)	-0.8462 (± 0.1538)	-1.0000 (± 0.0000)	-1.0000 (± 0.0000)	-1.0000 (± 0.0000)	-0.8624
gpt-4o-2024-05-13	5	-0.3054 (± 0.2086)	0.0135 (± 0.2070)	-0.2071 (± 0.2089)	-0.5357 (± 0.3107)	-0.6154 (± 0.1804)	-0.3692 (± 0.1443)	-0.1538 (± 0.2738)	-0.4967 (± 0.2210)	-0.8700 (± 0.1300)	-0.9231 (± 0.0769)	-0.4463
Meta-Llama-3-8B	1	-0.2308 (± 0.2571)	-0.7692 (± 0.1216)	-0.7143 (± 0.1844)	-0.8571 (± 0.1429)	-0.9231 (± 0.0769)	-1.0000 (± 0.0000)	-0.6154 (± 0.1804)	-0.9285 (± 0.0715)	-1.0000 (± 0.0000)	-0.3846 (± 0.1404)	-0.7423
Meta-Llama-3-8B	5	0.0709 (± 0.2107)	-0.2308 (± 0.1662)	-0.5393 (± 0.2185)	-1.0000 (± 0.0000)	-0.7692 (± 0.1662)	-0.8462 (± 0.1042)	-0.3846 (± 0.1404)	-0.7276 (± 0.1468)	-1.0000 (± 0.0000)	-0.1538 (± 0.1042)	-0.5575
gpt-3.5-turbo-0125	1	-0.8462 (± 0.1042)	-0.7154 (± 0.1503)	-0.5679 (± 0.2037)	-1.0000 (± 0.0000)	-1.0000 (± 0.0000)	-0.6231 (± 0.1680)	-0.8462 (± 0.1538)	-1.0000 (± 0.0000)	-1.0000 (± 0.0000)	-1.0000 (± 0.0000)	-0.8599
gpt-3.5-turbo-0125	5	-0.6154 (± 0.1404)	-0.0571 (± 0.1934)	-0.0500 (± 0.1687)	-1.0000 (± 0.0000)	-0.6538 (± 0.1538)	-0.1646 (± 0.1395)	-0.2308 (± 0.2809)	-0.4513 (± 0.2410)	-0.8778 (± 0.1222)	-0.8462 (± 0.1042)	-0.4947
Phi-3-medium-128k-instruct	1	-0.8462 (± 0.1538)	-0.7750 (± 0.1527)	-1.0000 (± 0.0000)	-1.0000 (± 0.0000)	-1.0000 (± 0.0000)	-1.0000 (± 0.0000)	-0.3846 (± 0.2130)	-1.0000 (± 0.0000)	-0.8878 (± 0.1122)	-0.8462 (± 0.1042)	-0.8740
Phi-3-medium-128k-instruct	5	-0.6577 (± 0.1891)	-0.3821 (± 0.2342)	-0.8286 (± 0.1714)	-1.0000 (± 0.0000)	-1.0000 (± 0.0000)	-0.6100 (± 0.1425)	-0.9231 (± 0.0769)	-0.3569 (± 0.2402)	-0.8333 (± 0.1667)	-0.8462 (± 0.1042)	-0.7438
Mistral-7B-v0.3	1	-0.8462 (± 0.1042)	-0.8590 (± 0.1410)	-0.7107 (± 0.1868)	-1.0000 (± 0.0000)	-1.0000 (± 0.0000)	-0.9192 (± 0.0808)	-0.7692 (± 0.1662)	-1.0000 (± 0.0000)	-1.0000 (± 0.0000)	-0.6923 (± 0.1332)	-0.8797
Mistral-7B-v0.3	5	-0.6246 (± 0.1664)	-0.6667 (± 0.1820)	-0.4071 (± 0.2106)	-0.8571 (± 0.1429)	-0.9231 (± 0.0769)	-0.9115 (± 0.0885)	-0.6923 (± 0.1332)	-0.8820 (± 0.1180)	-1.0000 (± 0.0000)	-0.5385 (± 0.1439)	-0.7563

Table 2: Standard Error of verification function scores in oracle construction.

Model	Shot	Bernstein-Vazirani	Deutsch-Jozsa	Diffusion-Operator	Grover	Simon	Average
gpt-4o-2024-05-13	1	-0.3200 (± 0.0530)	-0.0100 (± 0.0438)	-0.8462 (± 0.1538)	-0.9885 (± 0.0115)	-0.4674 (± 0.0545)	-0.5264
gpt-4o-2024-05-13	5	-0.1100 (± 0.0399)	0.0800 (± 0.0506)	-0.3077 (± 0.2083)	-0.9540 (± 0.0279)	-0.0870 (± 0.0295)	-0.2757
Meta-Llama-3-8B	1	-0.7300 (± 0.0468)	-0.5000 (± 0.0704)	-0.3846 (± 0.1404)	-1.0000 (± 0.0000)	-0.6848 (± 0.0487)	-0.6599
Meta-Llama-3-8B	5	-0.0500 (± 0.0359)	0.1700 (± 0.0551)	-0.8462 (± 0.1042)	-1.0000 (± 0.0000)	-0.6413 (± 0.0503)	-0.4735
gpt-3.5-turbo-0125	1	-0.3500 (± 0.0539)	-0.0400 (± 0.0470)	-0.8462 (± 0.1538)	-1.0000 (± 0.0000)	-0.3696 (± 0.0529)	-0.5211
gpt-3.5-turbo-0125	5	-0.1100 (± 0.0373)	0.0200 (± 0.0531)	-0.3077 (± 0.2083)	-0.9770 (± 0.0162)	-0.1087 (± 0.0326)	-0.2967
Phi-3-medium-128k-instruct	1	-0.6800 (± 0.0510)	-0.6100 (± 0.0584)	-0.9231 (± 0.0769)	-1.0000 (± 0.0000)	-0.7500 (± 0.0454)	-0.7926
Phi-3-medium-128k-instruct	5	-0.5400 (± 0.0521)	-0.4300 (± 0.0685)	-1.0000 (± 0.0000)	-1.0000 (± 0.0000)	-0.8370 (± 0.0387)	-0.7614
Mistral-7B-v0.3	1	-0.4000 (± 0.0512)	-0.4300 (± 0.0640)	-0.9231 (± 0.0769)	-0.9540 (± 0.0279)	-0.6087 (± 0.0512)	-0.6632
Mistral-7B-v0.3	5	-0.3700 (± 0.0506)	-0.1300 (± 0.0734)	-1.0000 (± 0.0000)	-0.9195 (± 0.0373)	-0.2391 (± 0.0447)	-0.5317