					Zero-sl	not experimental	stats					
	coverage						completness					
	mean	median	std	min	max	count	mean	median	std	min	max	count
model												
deepseek-r1:1.5B	0.28	0.18	0.24	0.04	0.97	121	0.24	0.17	0.22	0.04	0.88	121
deepseek-r1:7B	0.49	0.47	0.32	0.02	0.99	124	0.42	0.35	0.3	0.02	0.99	124
gemma-2-9B	0.53	0.58	0.32	0.04	0.99	120	0.45	0.4	0.29	0.04	0.99	120
gemma-2-9B(IT)	0.46	0.42	0.29	0.05	0.99	120	0.37	0.32	0.24	0.05	0.99	120
llama-3.2-3B	0.68	0.75	0.27	0.06	0.99	120	0.58	0.62	0.27	0.06	0.99	120
llama-3.2-3B(IT)	0.61	0.68	0.28	0.04	0.98	121	0.52	0.53	0.27	0.04	0.98	121
	correctness								cla	rity		
	mean	median	std	min	max	count	mean	median	std	min	max	count
model												
deepseek-r1:1.5B	0.56	0.51	0.32	0.04	0.99	121	0.88	0.96	0.19	0.1	0.99	121
deepseek-r1:7B	0.86	0.97	0.24	0.05	0.99	124	0.86	0.93	0.2	0.03	0.99	124
gemma-2-9B	0.89	0.96	0.16	0.17	0.99	120	0.84	0.92	0.22	0.09	0.99	120
gemma-2-9B(IT)	0.92	0.96	0.11	0.28	0.99	120	0.92	0.94	0.07	0.58	0.99	120
llama-3.2-3B	0.87	0.94	0.18	0.16	0.99	120	0.75	0.79	0.17	0.04	0.98	120
llama-3.2-3B(IT)	0.84	0.96	0.24	0.13	0.99	121	0.94	0.98	0.12	0.24	0.99	121
	hallucination						grammar_score					
	mean	median	std	min	max	count	mean	median	std	min	max	count
model												
deepseek-r1:1.5B	0.18	0.07	0.22	0.0	0.96	121	0.98	0.98	0.02	0.92	1.0	121
deepseek-r1:7B	0.26	0.13	0.28	0.01	0.96	124	0.95	0.96	0.04	0.78	1.0	124
gemma-2-9B	0.35	0.27	0.3	0.0	0.98	120	0.97	0.98	0.03	0.82	1.0	120
gemma-2-9B(IT)	0.38	0.26	0.3	0.01	0.97	120	0.95	0.95	0.03	0.82	0.99	120
llama-3.2-3B	0.46	0.46	0.3	0.01	0.97	120	0.9	0.94	0.11	0.28	1.0	120
llama-3.2-3B(IT)	0.2	0.11	0.22	0.01	0.89	121	0.94	0.97	0.11	0.28	1.0	121
						ots experimental	stats					
	coverage	coverage									•	•
			coverage	coverage	coverage	coverage	completness	completness	completness	completness	completness	
	mean	median	std	min	coverage	coverage count	completness mean	completness median	completness std	completness	completness max	completness
model		median	std	min	max	count	mean	median	std	min	max	count
deepseek-r1:1.5B	0.22	median 0.14	std 0.21	min 0.02	max 0.98	count 118	mean 0.19	median 0.13	std 0.18	min 0.02	max 0.94	118
deepseek-r1:1.5B deepseek-r1:7B	0.22 0.51	0.14 0.56	0.21 0.31	0.02 0.03	0.98 0.99	118 87	0.19 0.44	0.13 0.41	0.18 0.29	0.02 0.03	0.94 0.99	count 118 87
deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B	0.22 0.51 0.45	0.14 0.56 0.38	0.21 0.31 0.27	0.02 0.03 0.03	0.98 0.99 0.97	118 87 106	0.19 0.44 0.37	0.13 0.41 0.33	0.18 0.29 0.23	0.02 0.03 0.03	0.94 0.99 0.95	118 87 106
deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B gemma-2-9B(IT)	0.22 0.51 0.45 0.65	0.14 0.56 0.38 0.73	0.21 0.31 0.27 0.29	0.02 0.03 0.03 0.1	0.98 0.99 0.97 0.99	118 87 106 119	0.19 0.44 0.37 0.58	0.13 0.41 0.33 0.62	0.18 0.29 0.23 0.28	0.02 0.03 0.03 0.09	0.94 0.99 0.95 0.99	118 87 106 119
deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B gemma-2-9B(IT) llama-3.2-3B	0.22 0.51 0.45 0.65 0.31	0.14 0.56 0.38 0.73	0.21 0.31 0.27 0.29	0.02 0.03 0.03 0.03 0.1 0.03	0.98 0.99 0.97 0.99 0.98	count 118 87 106 119 119	0.19 0.44 0.37 0.58 0.27	0.13 0.41 0.33 0.62 0.2	0.18 0.29 0.23 0.28 0.25	0.02 0.03 0.03 0.09 0.03	0.94 0.99 0.95 0.99 0.97	118 87 106 119
deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B gemma-2-9B(IT)	0.22 0.51 0.45 0.65	0.14 0.56 0.38 0.73	0.21 0.31 0.27 0.29	0.02 0.03 0.03 0.1	0.98 0.99 0.97 0.99	118 87 106 119	0.19 0.44 0.37 0.58	0.13 0.41 0.33 0.62	0.18 0.29 0.23 0.28	0.02 0.03 0.03 0.09	0.94 0.99 0.95 0.99	118 87 106 119
deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B gemma-2-9B(IT) llama-3.2-3B	0.22 0.51 0.45 0.65 0.31 0.59	0.14 0.56 0.38 0.73 0.2 0.64	0.21 0.31 0.27 0.29 0.29 0.28	0.02 0.03 0.03 0.1 0.03 0.04	0.98 0.99 0.97 0.99 0.98 0.99	count 118 87 106 119 119 119	0.19 0.44 0.37 0.58 0.27 0.5	0.13 0.41 0.33 0.62 0.2 0.49	0.18 0.29 0.23 0.28 0.25 0.25	0.02 0.03 0.03 0.09 0.03 0.04	0.94 0.99 0.95 0.99 0.97	118 87 106 119 119 119
deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B gemma-2-9B(IT) llama-3.2-3B	0.22 0.51 0.45 0.65 0.31 0.59	median 0.14 0.56 0.38 0.73 0.2 0.64 correctness	0.21 0.31 0.27 0.29 0.29 0.28 correctness	min 0.02 0.03 0.03 0.1 0.03 0.04 correctness	max 0.98 0.99 0.97 0.99 0.98 0.99 correctness	count 118 87 106 119 119 119 correctness	0.19 0.44 0.37 0.58 0.27 0.5	median 0.13 0.41 0.33 0.62 0.2 0.49 clarity	0.18 0.29 0.23 0.28 0.25 0.25	min 0.02 0.03 0.03 0.09 0.03 0.04 clarity	max 0.94 0.99 0.95 0.99 0.97 0.99	118 87 106 119 119 119
deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B gemma-2-9B(IT) llama-3.2-3B llama-3.2-3B(IT)	0.22 0.51 0.45 0.65 0.31 0.59	0.14 0.56 0.38 0.73 0.2 0.64	0.21 0.31 0.27 0.29 0.29 0.28	0.02 0.03 0.03 0.1 0.03 0.04	0.98 0.99 0.97 0.99 0.98 0.99	count 118 87 106 119 119 119	0.19 0.44 0.37 0.58 0.27 0.5	0.13 0.41 0.33 0.62 0.2 0.49	0.18 0.29 0.23 0.28 0.25 0.25	0.02 0.03 0.03 0.09 0.03 0.04	0.94 0.99 0.95 0.99 0.97	118 87 106 119 119 119
deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B gemma-2-9B(IT) llama-3.2-3B llama-3.2-3B(IT)	0.22 0.51 0.45 0.65 0.31 0.59 correctness mean	0.14 0.56 0.38 0.73 0.2 0.64 correctness median	0.21 0.31 0.27 0.29 0.29 0.28 correctness std	min 0.02 0.03 0.03 0.1 0.03 0.04 correctness min	0.98 0.99 0.97 0.99 0.98 0.99	118 87 106 119 119 119 correctness count	0.19 0.44 0.37 0.58 0.27 0.5 clarity mean	0.13 0.41 0.33 0.62 0.2 0.49 clarity median	0.18 0.29 0.23 0.28 0.25 0.25 clarity std	min 0.02 0.03 0.03 0.09 0.03 0.04 clarity min	0.94 0.99 0.95 0.99 0.97 0.99	118 87 106 119 119 119 clarity
deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B gemma-2-9B(IT) llama-3.2-3B llama-3.2-3B(IT) model deepseek-r1:1.5B	0.22 0.51 0.45 0.65 0.31 0.59 correctness mean	0.14 0.56 0.38 0.73 0.2 0.64 correctness median	std 0.21 0.31 0.27 0.29 0.29 0.28 correctness std	min 0.02 0.03 0.03 0.1 0.03 0.04 correctness min 0.09	0.98 0.99 0.97 0.99 0.98 0.99 correctness max	count 118 87 106 119 119 119 119 119 119 119 119 118	0.19 0.44 0.37 0.58 0.27 0.5 Clarity mean	0.13 0.41 0.33 0.62 0.2 0.49 Clarity median 0.96	0.18 0.29 0.23 0.28 0.25 0.25 0.25 0.25	min 0.02 0.03 0.03 0.09 0.03 0.04 clarity min	0.94 0.99 0.95 0.95 0.97 0.97 0.99	count 118 87 106 119 119 119 clarity count
deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B gemma-2-9B(IT) llama-3.2-3B llama-3.2-3B(IT) model deepseek-r1:1.5B deepseek-r1:7B	0.22 0.51 0.45 0.65 0.31 0.59 correctness mean 0.46 0.69	0.14 0.56 0.38 0.73 0.2 0.64 correctness median 0.34 0.84	0.21 0.31 0.27 0.29 0.29 0.28 correctness std 0.29 0.31	min 0.02 0.03 0.03 0.1 0.03 0.04 correctness min 0.09	0.98 0.99 0.97 0.99 0.98 0.99 0.99 0.99 0.99	count 118 87 106 119 119 119 119 119 119 119 119 119 11	mean 0.19 0.44 0.37 0.58 0.27 0.5 clarity mean 0.85 0.87	0.13 0.41 0.33 0.62 0.2 0.49 clarity median 0.96 0.94	0.18 0.29 0.23 0.28 0.25 0.25 clarity std 0.25 0.25	min 0.02 0.03 0.03 0.09 0.03 0.04 clarity min 0.05 0.1	0.94 0.99 0.95 0.99 0.97 0.99 clarity max 0.99	count 118 87 106 119 119 119 119 119 118 87
deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B gemma-2-9B(IT) llama-3.2-3B llama-3.2-3B(IT) model deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B	0.22 0.51 0.45 0.65 0.31 0.59 correctness mean 0.46 0.69 0.75	0.14 0.56 0.38 0.73 0.2 0.64 correctness median 0.34 0.84 0.86	0.21 0.31 0.27 0.29 0.29 0.28 correctness std 0.29 0.31 0.26	min 0.02 0.03 0.03 0.1 0.03 0.04 correctness min 0.09 0.11 0.08	0.98 0.99 0.97 0.99 0.98 0.99 0.99 0.99	count 118 87 106 119 119 119 119 119 119 118 87 106	0.19 0.44 0.37 0.58 0.27 0.5 clarity mean 0.85 0.87	0.13 0.41 0.33 0.62 0.2 0.49 clarity median 0.96 0.94 0.88	0.18 0.29 0.23 0.28 0.25 0.25 clarity std 0.25 0.21 0.16	min 0.02 0.03 0.03 0.09 0.03 0.04 clarity min 0.05 0.1	0.94 0.99 0.95 0.99 0.97 0.99 0.97 0.99 clarity max 0.99 0.99	count 118 87 106 119 119 119 119 119 118 87 106
deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B gemma-2-9B(IT) llama-3.2-3B llama-3.2-3B(IT) model deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B gemma-2-9B(IT)	0.22 0.51 0.45 0.65 0.31 0.59 correctness mean 0.46 0.69 0.75	0.14 0.56 0.38 0.73 0.2 0.64 correctness median 0.34 0.84 0.86 0.97	0.21 0.31 0.27 0.29 0.29 0.28 correctness std 0.29 0.31 0.26	min 0.02 0.03 0.03 0.1 0.03 0.04 correctness min 0.09 0.11 0.08 0.16	0.98 0.99 0.97 0.99 0.98 0.99 0.99 0.99 0.99	count 118 87 106 119 119 119 119 119 119 1106 1118	0.19 0.44 0.37 0.58 0.27 0.5 clarity mean 0.85 0.87 0.84 0.88	0.13 0.41 0.33 0.62 0.2 0.49 Clarity median 0.96 0.94 0.88 0.91	0.18 0.29 0.23 0.28 0.25 0.25 clarity std 0.25 0.21 0.16 0.13	min 0.02 0.03 0.09 0.03 0.04 clarity min 0.05 0.1 0.03 0.19	0.94 0.99 0.95 0.99 0.97 0.99 0.97 0.99 clarity max 0.99 0.98	count 118 87 106 119 119 119 119 119 119 110 clarity count 118 87 106 119
deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B gemma-2-9B(IT) llama-3.2-3B llama-3.2-3B(IT) model deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B gemma-2-9B(IT) llama-3.2-3B	0.22 0.51 0.45 0.65 0.31 0.59 correctness mean 0.46 0.69 0.75 0.94	median 0.14 0.56 0.38 0.73 0.2 0.64 correctness median 0.34 0.84 0.86 0.97	std 0.21 0.31 0.27 0.29 0.29 0.28 correctness std 0.29 0.31 0.26 0.11 0.31	min 0.02 0.03 0.03 0.1 0.03 0.04 correctness min 0.09 0.11 0.08 0.16 0.1	max 0.98 0.99 0.97 0.99 0.98 0.99 correctness max 0.99 0.99 0.99 0.99	count 118 87 106 119 119 119 119 correctness count 118 87 106 119 119	mean 0.19 0.44 0.37 0.58 0.27 0.5 Clarity mean 0.85 0.87 0.84 0.88	0.13 0.41 0.33 0.62 0.2 0.49 Clarity median 0.96 0.94 0.88 0.91 0.75	0.18 0.29 0.23 0.28 0.25 0.25 0.25 0.21 0.16 0.13 0.32	min 0.02 0.03 0.03 0.09 0.03 0.04 clarity min 0.05 0.1 0.03 0.19 0.02	0.94 0.99 0.95 0.99 0.97 0.99 0.97 0.99 0.99 0.99 0.99	count 118 87 106 119 119 119 119 clarity count 118 87 106 119 119
deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B gemma-2-9B(IT) llama-3.2-3B llama-3.2-3B(IT) model deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B	0.22 0.51 0.45 0.65 0.31 0.59 correctness mean 0.46 0.69 0.75	0.14 0.56 0.38 0.73 0.2 0.64 correctness median 0.34 0.84 0.86 0.97	0.21 0.31 0.27 0.29 0.29 0.28 correctness std 0.29 0.31 0.26	min 0.02 0.03 0.03 0.1 0.03 0.04 correctness min 0.09 0.11 0.08 0.16	0.98 0.99 0.97 0.99 0.98 0.99 0.99 0.99 0.99	count 118 87 106 119 119 119 119 119 119 1106 1118	0.19 0.44 0.37 0.58 0.27 0.5 clarity mean 0.85 0.87 0.84 0.88	0.13 0.41 0.33 0.62 0.2 0.49 Clarity median 0.96 0.94 0.88 0.91	0.18 0.29 0.23 0.28 0.25 0.25 clarity std 0.25 0.21 0.16 0.13	min 0.02 0.03 0.09 0.03 0.04 clarity min 0.05 0.1 0.03 0.19	0.94 0.99 0.95 0.99 0.97 0.99 0.97 0.99 clarity max 0.99 0.98	count 118 87 106 119 119 119 119 119 119 110 clarity count 118 87 106 119
deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B gemma-2-9B(IT) llama-3.2-3B llama-3.2-3B(IT) model deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B gemma-2-9B(IT) llama-3.2-3B	0.22 0.51 0.45 0.65 0.31 0.59 correctness mean 0.46 0.69 0.75 0.94 0.59	median 0.14 0.56 0.38 0.73 0.2 0.64 correctness median 0.34 0.84 0.86 0.97 0.65 0.96	std 0.21 0.31 0.27 0.29 0.29 0.28 correctness std 0.29 0.31 0.26 0.11 0.31 0.17	min 0.02 0.03 0.03 0.1 0.03 0.04 correctness min 0.09 0.11 0.08 0.16 0.1 0.07	max 0.98 0.99 0.97 0.99 0.98 0.99 correctness max 0.99 0.99 0.99 0.99 0.99	count 118 87 106 119 119 119 119 119 119 118 87 106 119 119 119	mean 0.19 0.44 0.37 0.58 0.27 0.5 clarity mean 0.85 0.87 0.84 0.88 0.65 0.86	median 0.13 0.41 0.33 0.62 0.2 0.49 clarity median 0.96 0.94 0.88 0.91 0.75 0.95	std 0.18 0.29 0.23 0.28 0.25 0.25 clarity std 0.25 0.21 0.16 0.13 0.32 0.21	min 0.02 0.03 0.03 0.09 0.03 0.04 clarity min 0.05 0.1 0.03 0.19 0.02 0.07	0.94 0.99 0.95 0.99 0.97 0.99 clarity max 0.99 0.98 0.98 1.0 0.99	count 118 87 106 119 119 119 119 119 119 118 87 106 119 119 119 119
deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B gemma-2-9B(IT) llama-3.2-3B llama-3.2-3B(IT) model deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B gemma-2-9B(IT) llama-3.2-3B	0.22 0.51 0.45 0.65 0.31 0.59 correctness mean 0.46 0.69 0.75 0.94 0.59 0.89	median 0.14 0.56 0.38 0.73 0.2 0.64 correctness median 0.34 0.86 0.97 0.65 0.96	std 0.21 0.31 0.27 0.29 0.29 0.28 correctness std 0.29 0.31 0.26 0.11 0.31 0.17	min 0.02 0.03 0.03 0.1 0.03 0.04 correctness min 0.09 0.11 0.08 0.16 0.1 0.07	max 0.98 0.99 0.97 0.99 0.98 0.99 correctness max 0.99 0.99 0.99 0.99 0.99	count 118 87 106 119 119 119 119 119 119 118 87 106 119 119 119 119	mean 0.19 0.44 0.37 0.58 0.27 0.5 clarity mean 0.85 0.87 0.84 0.88 0.65 0.86 grammar_score	median 0.13 0.41 0.33 0.62 0.2 0.49 clarity median 0.96 0.94 0.88 0.91 0.75 0.95 grammar_score	std 0.18 0.29 0.23 0.28 0.25 0.25 clarity std 0.25 0.21 0.16 0.13 0.32 0.21 grammar_score	min 0.02 0.03 0.03 0.09 0.03 0.04 clarity min 0.05 0.1 0.03 0.19 0.02 0.07	0.94 0.99 0.95 0.99 0.97 0.99 0.97 0.99 clarity max 0.99 0.98 0.98 1.0 0.99 grammar_score	count 118 87 106 119 119 119 119 119 118 87 106 119 119 119 119 grammar_sco
deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B gemma-2-9B(IT) llama-3.2-3B llama-3.2-3B(IT) model deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B gemma-2-9B llama-3.2-3B llama-3.2-3B	0.22 0.51 0.45 0.65 0.31 0.59 correctness mean 0.46 0.69 0.75 0.94 0.59	median 0.14 0.56 0.38 0.73 0.2 0.64 correctness median 0.34 0.84 0.86 0.97 0.65 0.96	std 0.21 0.31 0.27 0.29 0.29 0.28 correctness std 0.29 0.31 0.26 0.11 0.31 0.17	min 0.02 0.03 0.03 0.1 0.03 0.04 correctness min 0.09 0.11 0.08 0.16 0.1 0.07	max 0.98 0.99 0.97 0.99 0.98 0.99 correctness max 0.99 0.99 0.99 0.99 0.99	count 118 87 106 119 119 119 119 119 119 118 87 106 119 119 119	mean 0.19 0.44 0.37 0.58 0.27 0.5 clarity mean 0.85 0.87 0.84 0.88 0.65 0.86	median 0.13 0.41 0.33 0.62 0.2 0.49 clarity median 0.96 0.94 0.88 0.91 0.75 0.95	std 0.18 0.29 0.23 0.28 0.25 0.25 clarity std 0.25 0.21 0.16 0.13 0.32 0.21	min 0.02 0.03 0.03 0.09 0.03 0.04 clarity min 0.05 0.1 0.03 0.19 0.02 0.07	0.94 0.99 0.95 0.99 0.97 0.99 clarity max 0.99 0.98 0.98 1.0 0.99	count 118 87 106 119 119 119 119 119 119 118 87 106 119 119 119 119
deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B(IT) llama-3.2-3B llama-3.2-3B(IT) model deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B gemma-2-9B(IT) llama-3.2-3B llama-3.2-3B(IT)	0.22 0.51 0.45 0.65 0.31 0.59 correctness mean 0.46 0.69 0.75 0.94 0.59 0.89	median 0.14 0.56 0.38 0.73 0.2 0.64 correctness median 0.34 0.84 0.86 0.97 0.65 0.96	std 0.21 0.31 0.27 0.29 0.29 0.28 correctness std 0.29 0.31 0.26 0.11 0.31 0.17	min 0.02 0.03 0.03 0.1 0.03 0.04 correctness min 0.09 0.11 0.08 0.16 0.1 0.07	max 0.98 0.99 0.97 0.99 0.98 0.99 0.99 0.99 0.99 0.99 0.99	count 118 87 106 119 119 119 119 119 118 87 106 119 119 119 119 119 119 hallucination count	mean 0.19 0.44 0.37 0.58 0.27 0.5 clarity mean 0.85 0.87 0.84 0.88 0.65 0.86 grammar_score mean	median 0.13 0.41 0.33 0.62 0.2 0.49 clarity median 0.96 0.94 0.88 0.91 0.75 0.95 grammar_score median	Std	min 0.02 0.03 0.03 0.09 0.03 0.04 clarity min 0.05 0.1 0.03 0.19 0.02 0.07	max 0.94 0.99 0.95 0.99 0.97 0.99 clarity max 0.99 0.99 0.98 1.0 0.99 grammar_score max	count 118 87 106 119 119 119 clarity count 118 87 106 119 119 119 119 119 119 grammar_scc count
deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B gemma-2-9B(IT) llama-3.2-3B llama-3.2-3B(IT) model deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B gemma-2-9B(IT) llama-3.2-3B llama-3.2-3B(IT)	0.22 0.51 0.45 0.65 0.31 0.59 correctness mean 0.46 0.69 0.75 0.94 0.59 0.89 hallucination mean	median 0.14 0.56 0.38 0.73 0.2 0.64 correctness median 0.34 0.84 0.86 0.97 0.65 0.96 hallucination median	Std	min 0.02 0.03 0.03 0.1 0.03 0.04 correctness min 0.09 0.11 0.08 0.16 0.1 0.07 hallucination min	max 0.98 0.99 0.97 0.99 0.98 0.99 0.99 0.99 0.99 0.99 0.99	count 118 87 106 119 119 119 119 119 119 correctness count 118 87 106 119 119 119 119 119 hallucination count	mean 0.19 0.44 0.37 0.58 0.27 0.5 clarity mean 0.85 0.87 0.84 0.88 0.65 0.86 grammar_score mean 0.97	median 0.13 0.41 0.33 0.62 0.2 0.49 clarity median 0.96 0.94 0.88 0.91 0.75 0.95 grammar_score median	Std	min 0.02 0.03 0.03 0.09 0.03 0.04 clarity min 0.05 0.1 0.03 0.19 0.02 0.07 grammar_score min	max 0.94 0.99 0.95 0.99 0.97 0.99 clarity max 0.99 0.98 0.98 1.0 0.99 grammar_score max	count 118 87 106 119 119 119 119 clarity count 118 87 106 119 119 119 119 119 119 119 119 119
deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B gemma-2-9B(IT) llama-3.2-3B llama-3.2-3B(IT) model deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B(IT) llama-3.2-3B llama-3.2-3B(IT)	0.22 0.51 0.45 0.65 0.31 0.59 correctness mean 0.46 0.69 0.75 0.94 0.59 0.89 hallucination mean	median 0.14 0.56 0.38 0.73 0.2 0.64 correctness median 0.34 0.84 0.86 0.97 0.65 0.96 hallucination median 0.1 0.19	std 0.21 0.31 0.27 0.29 0.29 0.28 correctness std 0.29 0.31 0.17 hallucination std	min 0.02 0.03 0.03 0.04 correctness min 0.09 0.11 0.08 0.16 0.1 0.07 hallucination min	max 0.98 0.99 0.97 0.99 0.98 0.99 0.99 correctness max 0.99 0.99 0.99 0.99 0.99 0.99 0.99 0.	count 118 87 106 119 119 119 119 119 119 110 correctness count 118 87 106 119 119 119 119 119 hallucination count	mean 0.19 0.44 0.37 0.58 0.27 0.5 clarity mean 0.85 0.87 0.84 0.88 0.65 0.86 grammar_score mean 0.97	median	Std O.18 O.29 O.23 O.28 O.25 O.25 O.25 O.25 O.21 O.16 O.13 O.21 O	min 0.02 0.03 0.03 0.09 0.03 0.04 clarity min 0.05 0.1 0.03 0.19 0.02 0.07 grammar_score min 0.7	0.94 0.99 0.95 0.99 0.97 0.99 0.97 0.99 clarity max 0.99 0.98 0.98 1.0 0.99 grammar_score max 1.0	count 118 87 106 119 119 119 119 clarity count 118 87 106 119 119 119 119 grammar_scc count 118 87
deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B gemma-2-9B(IT) llama-3.2-3B llama-3.2-3B(IT) model deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B gemma-2-9B(IT) llama-3.2-3B llama-3.2-3B(IT) model deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B	0.22 0.51 0.45 0.65 0.31 0.59 correctness mean 0.46 0.69 0.75 0.94 0.59 0.89 hallucination mean 0.17 0.29 0.37	median 0.14 0.56 0.38 0.73 0.2 0.64 correctness median 0.34 0.84 0.86 0.97 0.65 0.96 hallucination median 0.1 0.19 0.27	std 0.21 0.31 0.27 0.29 0.29 0.28 correctness std 0.29 0.31 0.26 0.11 0.31 0.17 hallucination std	min 0.02 0.03 0.03 0.1 0.03 0.04 correctness min 0.09 0.11 0.08 0.16 0.1 0.07 hallucination min 0.01 0.01	max 0.98 0.99 0.97 0.99 0.98 0.99 0.99 correctness max 0.99 0.99 0.99 0.99 0.99 0.99 0.99 0.	count 118 87 106 119 119 119 119 correctness count 118 87 106 119 119 119 119 119 hallucination count	mean 0.19 0.44 0.37 0.58 0.27 0.5 clarity mean 0.85 0.87 0.84 0.88 0.65 0.86 grammar_score mean 0.97 0.93 0.98	median 0.13 0.41 0.33 0.62 0.2 0.49 Clarity median 0.96 0.94 0.88 0.91 0.75 0.95 grammar_score median 0.98 0.98	Std	min 0.02 0.03 0.03 0.09 0.03 0.04 clarity min 0.05 0.1 0.03 0.19 0.02 0.07 grammar_score min 0.7 0.57 0.79	max 0.94 0.99 0.95 0.99 0.97 0.99 0.97 0.99 clarity max 0.99 0.99 0.98 1.0 0.99 grammar_score max 1.0 1.0	count 118 87 106 119 119 119 119 clarity count 118 87 106 119 119 119 119 119 119 119 119 110 grammar_sco count
deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B gemma-2-9B(IT) llama-3.2-3B llama-3.2-3B(IT) model deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B llama-3.2-3B(IT) model deepseek-r1:7B deepseek-r1:7B deepseek-r1:7B gemma-2-9B gemma-2-9B gemma-2-9B gemma-2-9B gemma-2-9B	0.22 0.51 0.45 0.65 0.31 0.59 correctness mean 0.46 0.69 0.75 0.94 0.59 0.89 hallucination mean 0.17 0.29 0.37 0.33	median 0.14 0.56 0.38 0.73 0.2 0.64 correctness median 0.34 0.84 0.86 0.97 0.65 0.96 hallucination median 0.1 0.1 0.19 0.27 0.26	std 0.21 0.31 0.27 0.29 0.29 0.28 correctness std 0.29 0.31 0.26 0.11 0.31 0.17 hallucination std 0.17 0.27 0.3 0.24	min 0.02 0.03 0.03 0.1 0.03 0.04 correctness min 0.09 0.11 0.08 0.16 0.1 0.07 hallucination min 0.01 0.01 0.01	max 0.98 0.99 0.97 0.99 0.98 0.99 0.99 0.99 0.99 0.99 0.99	count 118 87 106 119 119 119 119 correctness count 118 87 106 119 119 119 119 119 119 119 119 119 11	mean 0.19 0.44 0.37 0.58 0.27 0.5 clarity mean 0.85 0.87 0.84 0.88 0.65 0.86 grammar_score mean 0.97 0.93 0.98 0.95	median 0.13 0.41 0.33 0.62 0.2 0.49 clarity median 0.96 0.94 0.88 0.91 0.75 0.95 grammar_score median 0.98 0.98 0.99	Std	min 0.02 0.03 0.03 0.09 0.03 0.04 clarity min 0.05 0.1 0.03 0.19 0.02 0.07 grammar_score min 0.7 0.57 0.79 0.86	max 0.94 0.99 0.95 0.99 0.97 0.99 clarity max 0.99 0.99 0.98 1.0 0.99 grammar_score max 1.0 1.0	count 118 87 106 119 119 119 clarity count 118 87 106 119 119 119 119 119 119 119 119 119 11
deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B gemma-2-9B(IT) llama-3.2-3B llama-3.2-3B(IT) model deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B gemma-2-9B(IT) llama-3.2-3B llama-3.2-3B(IT) model deepseek-r1:1.5B deepseek-r1:7B gemma-2-9B	0.22 0.51 0.45 0.65 0.31 0.59 correctness mean 0.46 0.69 0.75 0.94 0.59 0.89 hallucination mean 0.17 0.29 0.37	median 0.14 0.56 0.38 0.73 0.2 0.64 correctness median 0.34 0.84 0.86 0.97 0.65 0.96 hallucination median 0.1 0.19 0.27	std 0.21 0.31 0.27 0.29 0.29 0.28 correctness std 0.29 0.31 0.26 0.11 0.31 0.17 hallucination std	min 0.02 0.03 0.03 0.1 0.03 0.04 correctness min 0.09 0.11 0.08 0.16 0.1 0.07 hallucination min 0.01 0.01	max 0.98 0.99 0.97 0.99 0.98 0.99 0.99 correctness max 0.99 0.99 0.99 0.99 0.99 0.99 0.99 0.	count 118 87 106 119 119 119 119 correctness count 118 87 106 119 119 119 119 119 hallucination count	mean 0.19 0.44 0.37 0.58 0.27 0.5 clarity mean 0.85 0.87 0.84 0.88 0.65 0.86 grammar_score mean 0.97 0.93 0.98	median 0.13 0.41 0.33 0.62 0.2 0.49 Clarity median 0.96 0.94 0.88 0.91 0.75 0.95 grammar_score median 0.98 0.98	Std	min 0.02 0.03 0.03 0.09 0.03 0.04 clarity min 0.05 0.1 0.03 0.19 0.02 0.07 grammar_score min 0.7 0.57 0.79	max 0.94 0.99 0.95 0.99 0.97 0.99 0.97 0.99 clarity max 0.99 0.99 0.98 1.0 0.99 grammar_score max 1.0 1.0	count 118 87 106 119 119 119 119 clarity count 118 87 106 119 119 119 119 119 119 grammar_sco count