	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						clarity					
	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