Zero-shot 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						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

Few-shots experimental stats

	coverage	coverage	coverage	coverage	coverage	coverage	completness	completness	completness	completness	completness	completness
	mean	median	std	min	max	count	mean	median	std	min	max	count
model												
deepseek-r1:1.5B	0.22	0.14	0.21	0.02	0.98	118	0.19	0.13	0.18	0.02	0.94	118
deepseek-r1:7B	0.51	0.56	0.31	0.03	0.99	87	0.44	0.41	0.29	0.03	0.99	87
gemma-2-9B	0.45	0.38	0.27	0.03	0.97	106	0.37	0.33	0.23	0.03	0.95	106
gemma-2-9B(IT)	0.65	0.73	0.29	0.1	0.99	119	0.58	0.62	0.28	0.09	0.99	119
llama-3.2-3B	0.31	0.2	0.29	0.03	0.98	119	0.27	0.2	0.25	0.03	0.97	119
llama-3.2-3B(IT)	0.59	0.64	0.28	0.04	0.99	119	0.5	0.49	0.25	0.04	0.99	119

	correctness	correctness	correctness	correctness	correctness	correctness	clarity	clarity	clarity	clarity	clarity	clarity
	mean	median	std	min	max	count	mean	median	std	min	max	count
model												
deepseek-r1:1.5B	0.46	0.34	0.29	0.09	0.99	118	0.85	0.96	0.25	0.05	0.99	118
deepseek-r1:7B	0.69	0.84	0.31	0.11	0.99	87	0.87	0.94	0.21	0.1	0.99	87
gemma-2-9B	0.75	0.86	0.26	0.08	0.99	106	0.84	0.88	0.16	0.03	0.98	106
gemma-2-9B(IT)	0.94	0.97	0.11	0.16	0.99	119	0.88	0.91	0.13	0.19	0.98	119
llama-3.2-3B	0.59	0.65	0.31	0.1	0.98	119	0.65	0.75	0.32	0.02	1.0	119
llama-3.2-3B(IT)	0.89	0.96	0.17	0.07	0.99	119	0.86	0.95	0.21	0.07	0.99	119

	hallucination	hallucination	hallucination	hallucination	hallucination	hallucination	grammar_score	grammar_score	grammar_score	grammar_score	grammar_score	grammar_score
	mean	median	std	min	max	count	mean	median	std	min	max	count
model												
deepseek-r1:1.5B	0.17	0.1	0.17	0.01	0.92	118	0.97	0.98	0.04	0.7	1.0	118
deepseek-r1:7B	0.29	0.19	0.27	0.01	0.97	87	0.93	0.95	0.06	0.57	1.0	87
gemma-2-9B	0.37	0.27	0.3	0.01	0.99	106	0.98	0.99	0.03	0.79	1.0	106
gemma-2-9B(IT)	0.33	0.26	0.24	0.01	0.94	119	0.95	0.95	0.03	0.86	1.0	119
llama-3.2-3B	0.42	0.42	0.26	0.03	0.99	119	0.95	0.98	0.11	0.03	1.0	119
llama-3.2-3B(IT)	0.37	0.34	0.26	0.02	0.93	119	0.9	0.94	0.12	0.32	1.0	119