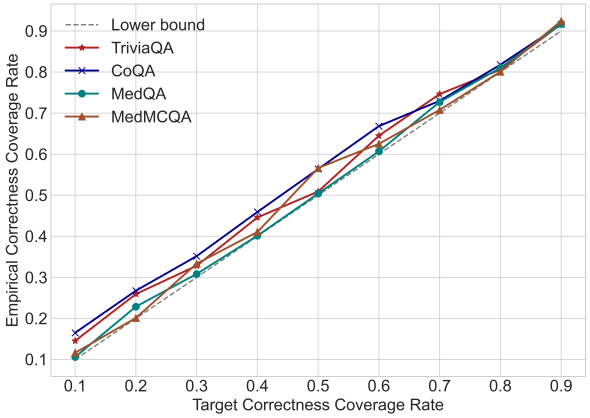

Figure 7: Target vs. empirical correctness coverage rate. We test the 4 datasets utilizing the LLaMA-2-13B-Chat model as the generator.

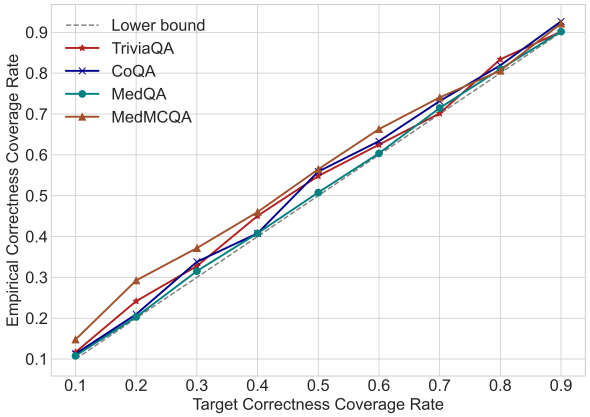

Figure 8: Target vs. empirical correctness coverage rate. We test the 4 datasets utilizing the Vicuna-13B-v1.5 model as the generator.

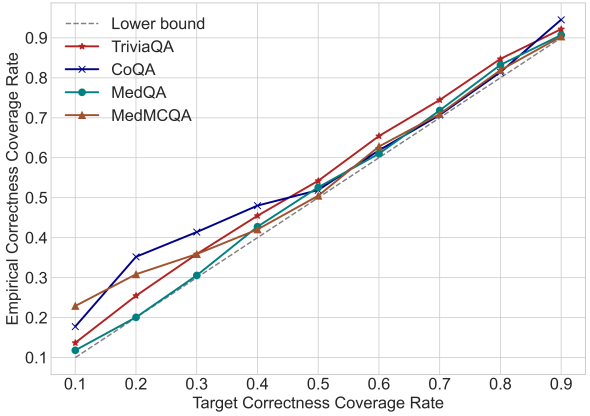

Figure 9: Target vs. empirical correctness coverage rate. We test the 4 datasets utilizing the LLaMA-3-70B-Instruct model as the generator.

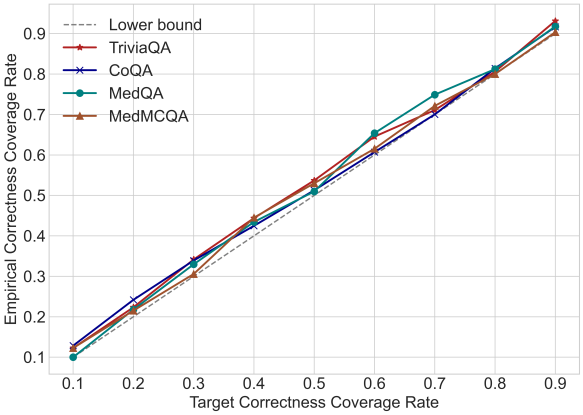

Figure 10: Target vs. empirical correctness coverage rate. We test the 4 datasets utilizing the GPT-3.5-turbo model as the generator.

Table 5: The results of correctness coverage rate (%) on 7 LLMs across 4 open-ended NLG datasets. The user-accepted error rate α is strictly set to 0.05.

LLMs	TriviaQA	CoQA	MedQA	MedMCQA
LLaMA-2-7B-Chat	95.26	96.45	100.00	95.99
Mistral-7B-Instruct-v0.3	95.01	95.72	95.79	95.12
LLaMA-3-8B-Instruct	98.17	95.23	95.78	98.38
LLaMA-2-13B-Chat	95.04	96.96	95.15	96.59
Vicuna-13B-v1.5	97.28	95.33	95.51	97.29
LLaMA-3-70B-Instruct	95.38	95.33	95.51	97.29
GPT-3.5-turbo	97.02	97.60	95.62	95.19

Table 6: The results of correctness coverage rate (%) on 7 LLMs across 4 open-ended NLG datasets. The user-accepted error rate α is strictly set to 0.01.

LLMs	TriviaQA	CoQA	MedQA	MedMCQA
LLaMA-2-7B-Chat	99.93	99.83	100.00	99.14
Mistral-7B-Instruct-v0.3	99.38	99.27	99.15	99.81
LLaMA-3-8B-Instruct	99.79	99.53	100.00	99.76
LLaMA-2-13B-Chat	99.06	99.13	99.51	99.48
Vicuna-13B-v1.5	99.52	100.00	99.94	100.00
LLaMA-3-70B-Instruct	99.84	99.75	99.15	99.82
GPT-3.5-turbo	99.17	99.82	99.51	99.95