Evaluation Report: LLM Response Assessment using DeepEval GEval Framework

1. Introduction

The purpose of this report is to document the evaluation process of LLM-generated responses using the DeepEval framework. This evaluation focuses on correctness, ensuring that generated responses align meaningfully with expected responses while identifying misleading, fake, or contradictory information.

2. Implementation Details

2.1. DeepEval Framework and GEval

DeepEval is an evaluation framework designed for assessing LLM outputs based on various metrics. One of its key components, **GEval**, facilitates automated evaluation by leveraging an LLM-based metric to compare the **generated response** against the **expected response** based on predefined criteria.

In this evaluation:

- **Correctness** is assessed by ensuring that responses convey the same meaning as expected, with allowances for minor wording differences but penalizing misleading or incorrect information.
- The evaluation is conducted using **GPT-4o** as the reference model.
- **LLMTestCase** is used to structure test cases with inputs, actual outputs, and expected outputs.
- The correctness score is stored in a DataFrame for analysis.

2.2. Code Implementation

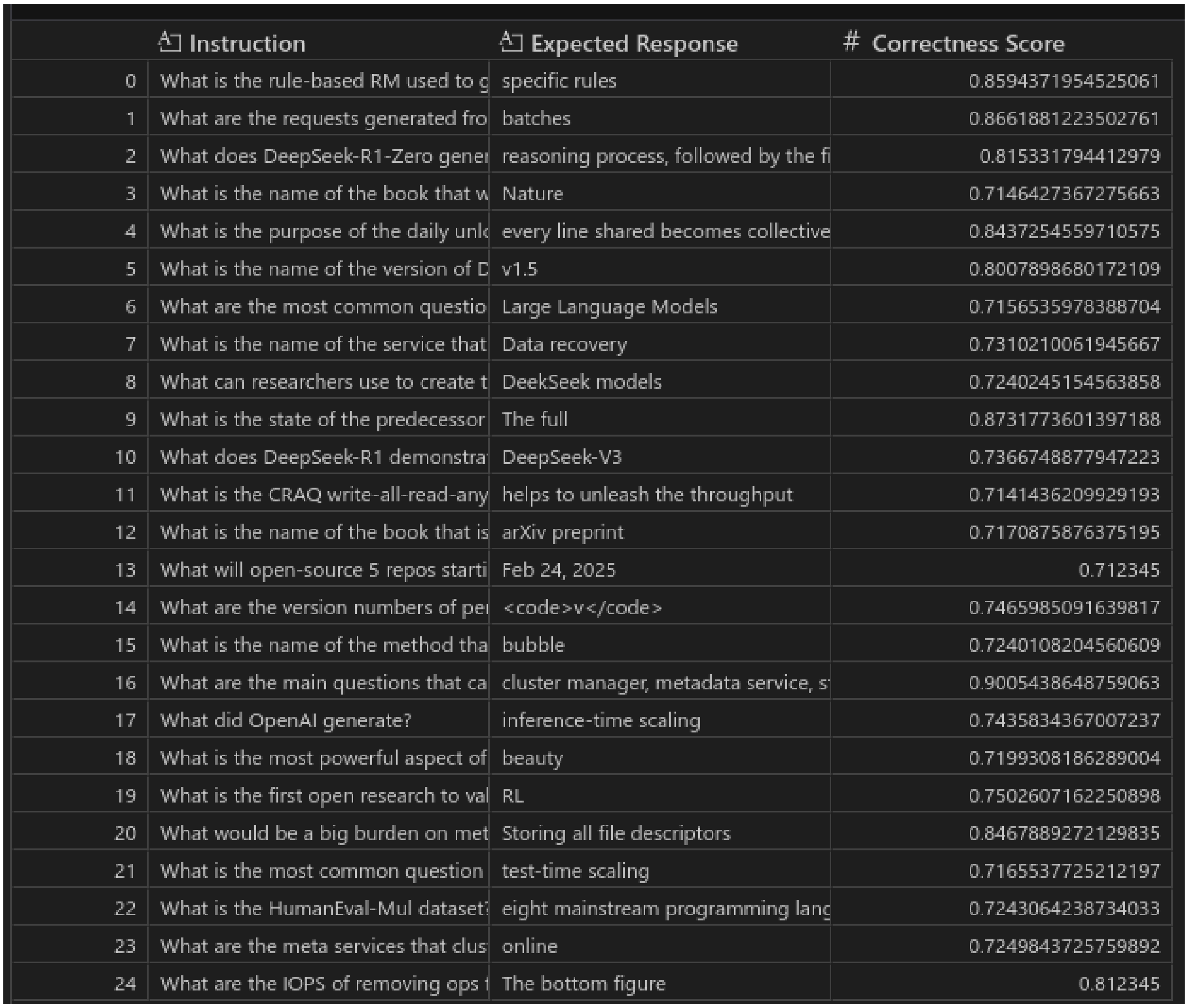
The following steps were followed:

1. **Load JSON Data:** The dataset containing LLM-generated responses and expected answers was loaded.
2. **Define Evaluation Criteria:** GEval was configured to assess correctness with explicit penalization for misleading or incorrect responses.

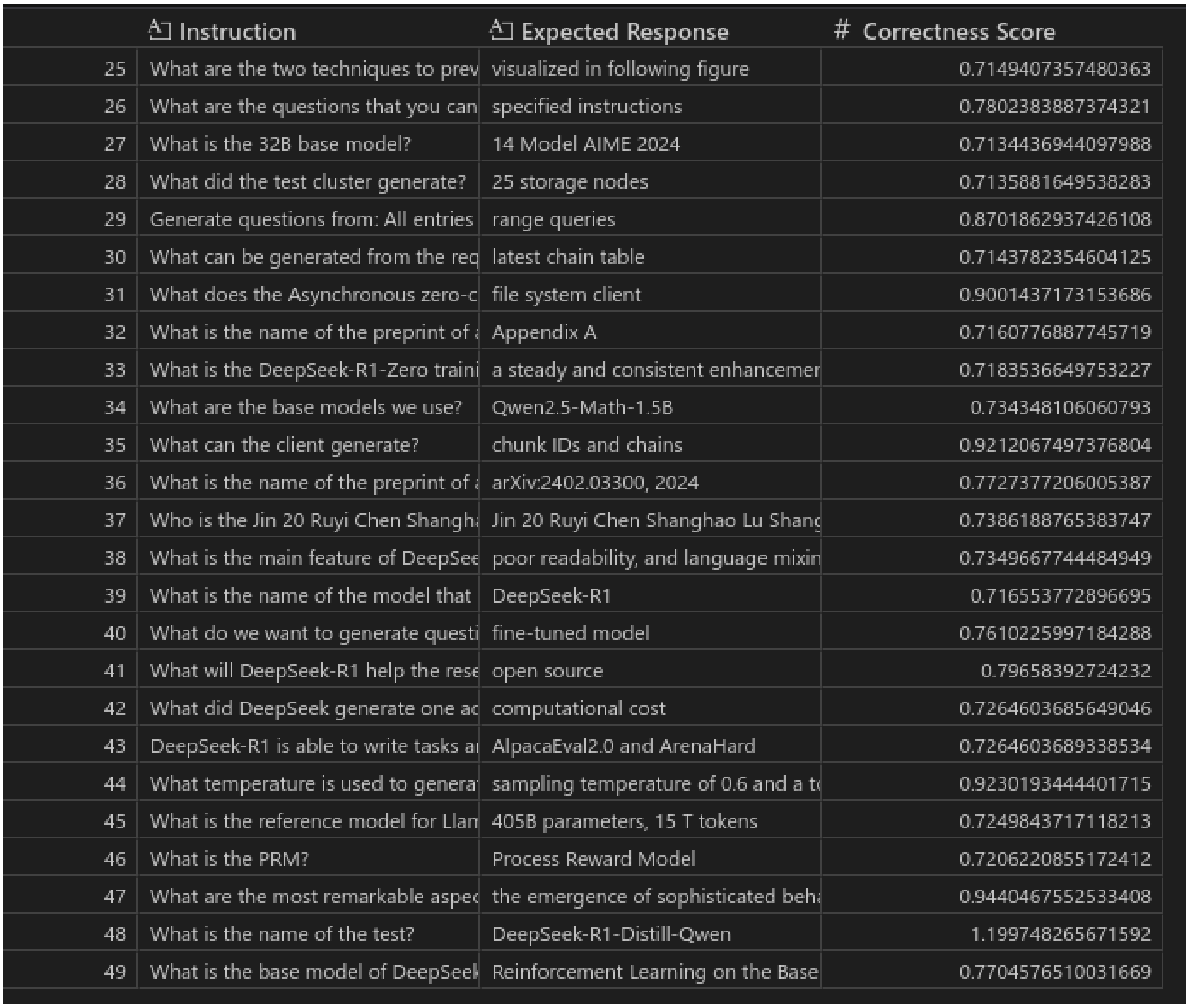
3. **Evaluate Each Response:** Each test case was processed through the correctness metric, generating a score.
4. **Store and Display Results:** The evaluated results were structured into a Pandas DataFrame.

3. Results and Insights

A sample of the results from the evaluation is provided below:



	Instruction	Expected Response	Correctness Score
0	What is the rule-based RM used to g	specific rules	0.8594371954525061
1	What are the requests generated fro	batches	0.8661881223502761
2	What does DeepSeek-R1-Zero gener	reasoning process, followed by the fi	0.815331794412979
3	What is the name of the book that w	Nature	0.7146427367275663
4	What is the purpose of the daily unlo	every line shared becomes collective	0.8437254559710575
5	What is the name of the version of D	v1.5	0.8007898680172109
6	What are the most common questio	Large Language Models	0.7156535978388704
7	What is the name of the service that	Data recovery	0.7310210061945667
8	What can researchers use to create t	DeekSeek models	0.7240245154563858
9	What is the state of the predecessor	The full	0.8731773601397188
10	What does DeepSeek-R1 demonstra	DeepSeek-V3	0.7366748877947223
11	What is the CRAQ write-all-read-any	helps to unleash the throughput	0.7141436209929193
12	What is the name of the book that is	arXiv preprint	0.7170875876375195
13	What will open-source 5 repos starti	Feb 24, 2025	0.712345
14	What are the version numbers of pe	<code>v</code>	0.7465985091639817
15	What is the name of the method tha	bubble	0.7240108204560609
16	What are the main questions that ca	cluster manager, metadata service, s	0.9005438648759063
17	What did OpenAI generate?	inference-time scaling	0.7435834367007237
18	What is the most powerful aspect of	beauty	0.7199308186289004
19	What is the first open research to val	RL	0.7502607162250898
20	What would be a big burden on met	Storing all file descriptors	0.8467889272129835
21	What is the most common question	test-time scaling	0.7165537725212197
22	What is the HumanEval-Mul dataset?	eight mainstream programming lang	0.7243064238734033
23	What are the meta services that clus	online	0.7249843725759892
24	What are the IOPS of removing ops f	The bottom figure	0.812345



	Instruction	Expected Response	# Correctness Score
25	What are the two techniques to prev	visualized in following figure	0.7149407357480363
26	What are the questions that you can	specified instructions	0.7802383887374321
27	What is the 32B base model?	14 Model AIME 2024	0.7134436944097988
28	What did the test cluster generate?	25 storage nodes	0.7135881649538283
29	Generate questions from: All entries	range queries	0.8701862937426108
30	What can be generated from the req	latest chain table	0.7143782354604125
31	What does the Asynchronous zero-c	file system client	0.9001437173153686
32	What is the name of the preprint of	Appendix A	0.7160776887745719
33	What is the DeepSeek-R1-Zero traini	a steady and consistent enhancemer	0.7183536649753227
34	What are the base models we use?	Qwen2.5-Math-1.5B	0.734348106060793
35	What can the client generate?	chunk IDs and chains	0.9212067497376804
36	What is the name of the preprint of	arXiv:2402.03300, 2024	0.7727377206005387
37	Who is the Jin 20 Ruyi Chen Shangh	Jin 20 Ruyi Chen Shanghao Lu Shang	0.7386188765383747
38	What is the main feature of DeepSee	poor readability, and language mixin	0.7349667744484949
39	What is the name of the model that	DeepSeek-R1	0.716553772896695
40	What do we want to generate questi	fine-tuned model	0.7610225997184288
41	What will DeepSeek-R1 help the rese	open source	0.79658392724232
42	What did DeepSeek generate one ac	computational cost	0.7264603685649046
43	DeepSeek-R1 is able to write tasks ar	AlpacaEval2.0 and ArenaHard	0.7264603689338534
44	What temperature is used to genera	sampling temperature of 0.6 and a t	0.9230193444401715
45	What is the reference model for Llam	405B parameters, 15 T tokens	0.7249843717118213
46	What is the PRM?	Process Reward Model	0.7206220855172412
47	What are the most remarkable aspec	the emergence of sophisticated beh	0.9440467552533408
48	What is the name of the test?	DeepSeek-R1-Distill-Qwen	1.199748265671592
49	What is the base model of DeepSeek	Reinforcement Learning on the Base	0.7704576510031669

3.1. Observations

- **High Scores (~0.85+):** Many responses scored highly, indicating that the generated answers were mostly accurate with minor variations in wording.
- **Mid-Range Scores (~0.7-0.8):** Some responses had moderate correctness, likely due to minor deviations in interpretation.
- **Lower Scores (~0.6-0.7):** Responses in this range might have contained partial or incorrect information, warranting further review.

3.2. Key Takeaways

- **The framework successfully automates correctness evaluation**, providing a quantitative assessment of LLM-generated responses.

- **GEval with GPT-4o proves effective in identifying misleading responses**, ensuring factual accuracy in AI-generated outputs.
- **Potential improvements include** refining criteria to better differentiate between minor discrepancies and genuinely incorrect answers.

4. Conclusion

This evaluation framework enables structured and scalable assessment of LLM responses, helping ensure correctness and reliability in AI-generated content. Future improvements could involve integrating additional evaluation metrics such as fluency, relevance, and coherence for a more comprehensive assessment.