tesing on 9 testcases with anser relevancy and tool correctedness metric on parallel

 You're running DeepEval's latest Answer Relevancy (GEval) Metric! (using llama3.2, strict=False, async_mode=True)... You're running DeepEval's latest Tool Correctness (GEval) Metric! (using llama3.2, strict=False, async_mode=True)...

==========

Metrics Summary

- Answer Relevancy (GEval) (score: 0.9, threshold: 0.5, strict: False, evaluation model: llama3.2, reason: The input is clear and concise, answering a specific question. The actual output matches the input's intent by providing a direct answer to the question about France's capital. Both the input and actual output are free from typos and grammatical errors. However, the actual output could be more detailed or comprehensive in its response., error: None)
- Tool Correctness (GEval) (score: 0.9, threshold: 0.5, strict: False, evaluation model: llama3.2, reason: The agent used the correct tool (Google), the input matches the actual output, and the answer to the question is correct. However, there might be a slight variation in formatting or wording between the input and actual output., error: None)

For test case:

- input: What is the capital of France?
- actual output: The capital of France is Paris.
- expected output: Paris
- context: None
- retrieval context: None

==========

Overall Metric Pass Rates

Answer Relevancy (GEval): 88.89% pass rate Tool Correctness (GEval): 100.00% pass rate

==========

==========

Metrics Summary

- Answer Relevancy (GEval) (score: 0.9, threshold: 0.5, strict: False, evaluation model: llama3.2, reason: The answer directly addresses the input question, providing a clear and relevant response. However, it could be improved by including more specific details about the planet's size in comparison to other planets or its unique features., error: None)
- Tool Correctness (GEval) (score: 0.9, threshold: 0.5, strict: False, evaluation model: llama3.2, reason: The response correctly identifies the largest planet in our solar system as Jupiter, includes its diameter value, and provides a comparative statement with Earth's diameter. The format of the answer is also clear and easy to understand., error: None)

For test case:

- input: What is the largest planet in our solar system?
- actual output: The largest planet in our solar system is Jupiter. It has a diameter of about 142,984 kilometers (88,846 miles), which is more than 11 times the diameter of the Earth!
- expected output: Jupiter
- context: None
- retrieval context: None

==========

Overall Metric Pass Rates

Answer Relevancy (GEval): 88.89% pass rate Tool Correctness (GEval): 100.00% pass rate

==========

==========

Metrics Summary

- Answer Relevancy (GEval) (score: 0.9, threshold: 0.5, strict: False, evaluation model: llama3.2, reason: The response fully answers the question, but it is more detailed than expected. The main information should be provided in a shorter manner., error: None)
- Tool Correctness (GEval) (score: 0.9, threshold: 0.5, strict: False, evaluation model: llama3.2, reason: The response correctly identifies the artist, era, and notable features of the Mona Lisa. The output is detailed and well-structured, making it clear that the agent has used a tool to provide accurate information., error: None)

For test case:

- input: Who painted the Mona Lisa?
- actual output: The Mona Lisa was painted by the Italian artist Leonardo da Vinci (1452-1519). Specifically, it is believed to have been painted in 1503-1506, during the Italian Renaissance. The portrait is widely considered one of the greatest paintings of all time, and it is known for its enigmatic smile, intricate details, and masterful use of sfumato, a painting technique that creates a soft, hazy effect.

Da Vinci's Mona Lisa is housed at the Louvre Museum in Paris, France, where it has been on display since 1797. It is considered one of the most famous paintings in the world, and it continues to be celebrated for its beauty, mystery, and historical significance.

- expected output: Leonardo da Vinci
- context: None
- retrieval context: None

==========

Overall Metric Pass Rates

Answer Relevancy (GEval): 88.89% pass rate Tool Correctness (GEval): 100.00% pass rate

==========

==========

Metrics Summary

- ✅ Answer Relevancy (GEval) (score: 0.9, threshold: 0.5, strict: False, evaluation model: llama3.2, reason: The question is clear and concise. The actual output directly relates to the input. However, it does not address other parts of the question, such as asking for the boiling point at standard atmospheric pressure., error: None)
- ✅ Tool Correctness (GEval) (score: 0.8, threshold: 0.5, strict: False, evaluation model: llama3.2, reason: The agent provided the correct boiling point of water in Celsius (100°C) at standard atmospheric pressure, demonstrating a strong understanding of the physical property. However, the overly formal tone and unnecessary mention of 'standard atmospheric pressure' detract from clarity., error: None)

For test case:

- input: What is the boiling point of water in Celsius?
- actual output: The boiling point of water is 100 degrees Celsius (°C) at standard atmospheric pressure.
- expected output: 100
- context: None
- retrieval context: None

==========

Overall Metric Pass Rates

Answer Relevancy (GEval): 88.89% pass rate Tool Correctness (GEval): 100.00% pass rate

==========

==========

Metrics Summary

- ✅ Answer Relevancy (GEval) (score: 0.2, threshold: 0.5, strict: False, evaluation model: llama3.2, reason: The answer partially matches the input question, but includes additional information about the novel's influence, publication date, and themes. It does not directly address who wrote 'To Kill a Mockingbird'., error: None)
- ✅ Tool Correctness (GEval) (score: 0.8, threshold: 0.5, strict: False, evaluation model: llama3.2, reason: The response accurately matches the input, but it goes beyond providing a simple answer. It includes extraneous information about the novel's influence and publication details, which is not requested in the input., error: None)

For test case:

- input: Who wrote 'To Kill a Mockingbird'?
- actual output: The Pulitzer Prize-winning novel "To Kill a Mockingbird" was written by Harper Lee. It's one of the most famous and influential novels of the 20th century, published in 1960 and exploring issues of racial injustice, tolerance, and the loss of innocence through the eyes of Scout Finch, a young girl growing up in the Deep South during the 1930s.
- expected output: Harper Lee
- context: None
- retrieval context: None

==========

Overall Metric Pass Rates

Answer Relevancy (GEval): 88.89% pass rate Tool Correctness (GEval): 100.00% pass rate

==========

==========

Metrics Summary

- ✅ Answer Relevancy (GEval) (score: 0.9, threshold: 0.5, strict: False, evaluation model: llama3.2, reason: The input is clearly stated. The actual output directly relates to the input, providing both the symbol and its origin in a concise manner. However, there seems to be an inconsistency between the expected response format and the provided one; while the answer itself is correct, the inclusion of extraneous information may not align with the question's requirements., error: None)
- ✅ Tool Correctness (GEval) (score: 0.9, threshold: 0.5, strict: False, evaluation model: llama3.2, reason: The tool used is correct, the input matches the requirements, and the output aligns with the expected outcome. However, it does not explicitly state the Latin origin of the chemical symbol 'Au', which might be considered a minor omission., error: None)

For test case:

- input: What is the chemical symbol for gold?
- actual output: The chemical symbol for gold is Au. This comes from the Latin word for gold, "aurum".
- expected output: Au
- context: None
- retrieval context: None

==========

Overall Metric Pass Rates

Answer Relevancy (GEval): 88.89% pass rate Tool Correctness (GEval): 100.00% pass rate

==========

==========

Metrics Summary

- ✅ Answer Relevancy (GEval) (score: 0.9, threshold: 0.5, strict: False, evaluation model: llama3.2, reason: The input asks a specific question about the main ingredient in guacamole. The actual output provides a detailed description of the primary ingredient, while also including some secondary ingredients. However, it does not explicitly answer the question posed in the input, as it starts by stating that avocados are the 'main event' and then elaborates on additional ingredients. This minor deviation from directly answering the question prevents a perfect score., error: None)
- ✅ Tool Correctness (GEval) (score: 0.9, threshold: 0.5, strict: False, evaluation model: llama3.2, reason: The response correctly identifies the main ingredient in

guacamole as avocados. However, it could be more concise and focused on a single key point, such as 'avocados' or 'the primary ingredient.', error: None)

For test case:

- input: What is the main ingredient in guacamole?
- actual output: The main ingredient in traditional guacamole is avocados! Specifically, ripe and mashed or diced avocados are the star of this popular Mexican dip. Along with avocados, other common ingredients include onions, tomatoes, cilantro, lime juice, salt, and sometimes garlic or jalapeños for added flavor. But the base, the foundation, and the main event is indeed the creamy avocado!
- expected output: Avocado
- context: None
- retrieval context: None

==========

Overall Metric Pass Rates

Answer Relevancy (GEval): 88.89% pass rate Tool Correctness (GEval): 100.00% pass rate

==========

==========

Metrics Summary

- ✅ Answer Relevancy (GEval) (score: 1.0, threshold: 0.5, strict: False, evaluation model: llama3.2, reason: The response directly answers the question, provides a clear and detailed explanation of the discovery of penicillin, and matches the expected output. The actual output is relevant to the question being asked and includes all necessary details about Alexander Fleming's discovery., error: None)
- ✅ Tool Correctness (GEval) (score: 1.0, threshold: 0.5, strict: False, evaluation model: llama3.2, reason: The tool used matches the problem description, output is verified against an expected result, input and actual output are compared for correctness, and any discrepancies or errors are reported., error: None)

For test case:

- input: Who discovered penicillin?
- actual output: A question about one of the most significant discoveries in medical history!

Penicillin was discovered by Scottish scientist Alexander Fleming in 1928. Fleming, a biologist and pharmacologist at St Mary's Hospital in London, was studying the bacteria Staphylococcus aureus when he made his groundbreaking discovery.

Here's what happened:

Fleming had returned to his laboratory after a two-week vacation during which he had left some bacterial cultures on his workbench. One of these cultures, which contained Staphylococcus aureus bacteria, was contaminated with a mold. The mold turned out to be a type of fungus called Penicillium notatum.

To Fleming's surprise, the mold had killed off the surrounding bacteria! He observed that the mold produced a substance that seemed to inhibit the growth of the bacteria.

Fleming isolated the active ingredient from the mold and named it penicillin. He realized its potential as an antibacterial agent and began experimenting with it.

However, it was not Fleming alone who developed penicillin into a usable medicine. He collaborated with two other scientists, Howard Florey and Ernst Boris Chain, to purify and mass-produce penicillin, which led to the development of penicillin as a treatment for bacterial infections.

The three scientists were jointly awarded the Nobel Prize in Physiology or Medicine in 1945 for their work on penicillin.

- expected output: Alexander Fleming
- context: None
- retrieval context: None

==========

Overall Metric Pass Rates

Answer Relevancy (GEval): 88.89% pass rate Tool Correctness (GEval): 100.00% pass rate

==========

==========

Metrics Summary

- ✅ Answer Relevancy (GEval) (score: 0.8, threshold: 0.5, strict: False, evaluation model: llama3.2, reason: The input is clear and concise, but the actual output does not directly address the question. Instead, it provides an answer and additional information about Jane Austen and the novel, which may be seen as addressing multiple questions simultaneously., error: None)
- ✅ Tool Correctness (GEval) (score: 0.9, threshold: 0.5, strict: False, evaluation model: llama3.2, reason: Correct tool used for input, but the response exceeds expected result length with unnecessary details., error: None)

For test case:

- input: Who wrote 'Pride and Prejudice'?
- actual output: A classic question!

The author of the beloved novel "Pride and Prejudice" is Jane Austen (1775-1817), an English novelist known for her witty, insightful, and engaging portrayals of life in 19th-century England. Published in 1813, "Pride and Prejudice" remains one of Austen's most popular works and a timeless favorite among readers worldwide.

Would you like to know more about Jane Austen or the novel itself?

- expected output: Jane Austen
- context: None
- retrieval context: None

==========

Overall Metric Pass Rates

Answer Relevancy (GEval): 88.89% pass rate Tool Correctness (GEval): 100.00% pass rate

==========

✓ Tests finished 🎉! Run 'deepeval view' to analyze, debug, and save evaluation results on Confident AI.

{'Answer Relevancy (GEval)': 0.8}