How to make LLM-generated results align more closely with standards has long lacked effective validation, as there is insufficient experimental data to support such verification.

Our method is to input each chaotic input into the LLM, using different prompts for each. Then, we obtain a claim and compare it with the standardized normalized claim to compute a score. Finally, we repeat this process for all inputs and calculate the average score. for 4 seconds
Our approach is to input each chaotic input into the LLM using different prompts to generate a claim. This claim is then compared with the standardized normalized claim to compute a score. Finally, all inputs are processed and their scores are averaged.

The average scores for each method are as follows:
key_drive: 0.2616
code_like: 0.2642
example_base: 0.2710
negative_positive: 0.2731
repeition_emphasie: 0.2878

Based on the experimental data, we can have those conclusions:
The "repeition_emphasie" strategy achieved the highest score, and the "code_like" and "key_drive" methods yielded lower scores. Different prompts can indeed affect the accuracy of LLM-generated results, but the score differences between them are not significant. I used 300 tokens, which may be due to the insufficient token quantity causing this effect.