

System and Method for Diagnosis and Intervention of an Artificial Intelligence Model

ABSTRACT

In a described embodiment a computer-implemented method for diagnosing security status of an algorithm engine is provided. The method comprises capturing activation patterns of the algorithm engine when processing an input prompt; analysing the activation patterns for deviation against a baseline representation of activation patterns that are in response to a benign input prompt in compliance with safety protocols, the baseline representation being constructed based on output from each neural network layer of the algorithm engine when processing the benign input prompt; classifying the activation patterns based on the deviation; and generating a risk assessment signal based on the activation patterns classification.

Figure 1