

System and Method for Reinforcing Artificial Intelligence Model Safety

ABSTRACT

In a described embodiment a method for safety reinforcement in a large language model (LLM) is provided. The method comprises receiving or identifying one or more malicious prompts that produce an unsafe output from an LLM, capturing one or more activation tensors corresponding to the one or more malicious prompts, generating a difference tensor by performing differential analysis between the one or more activation tensors and baseline activation data of the LLM, localising a set of anomalous neurons associated with the unsafe output by analysing the difference tensor to identify neurons with statistically significant activation deviations, generating one or more corrective prompt tokens to modify one or more activation patterns of the anomalous neurons, and converting the one or more corrective prompt tokens to text tokens to thereby form one or more corrective prompts.

15 Figure 1