

System and Method for Estimating Confidence and Implementing Metacognitive Abilities in
Artificial Intelligence Systems

Abstract

5 In a described embodiment, a system for information processing is provided including a data
acquisition module configured to receive feedback corresponding to one or more outputs
generated by a language model. The system further includes a cognitive reasoning module
configured to evaluate the reasoning process of the language model, emulate cognitive
functions including metacognitive processes, and generate an assessment based on an
analysis of the received feedback, wherein the assessment includes classifying the one or
10 more outputs into components, assigning quality scores for each component, and identifying
an improvement corresponding to the one or more outputs. Additionally, the system
includes a process adjustment module coupled to the cognitive reasoning module for
adjusting the reasoning process of the language model based on the assessment is
provided. A refinement module coupled to the process adjustment module is provided for
15 iteratively refining the reasoning process based on subsequent updates to the generated
assessment until a performance threshold is met.

FIG. 1