Toward Autonomy: Metacognitive Learning for Enhanced AI Performance

Brendan Conway-Smith, Robert L. West
Carleton University
brendan.conwaysmith@carleton.ca, robert.west@carleton.ca

Abstract

Metacognition is an array of cognitive processes that guide ordinary cognition in order to improve its functioning. For example, a student can recognize they learn better when they study in the morning instead of the evening. In the case of LLMs, prompt engineering can be considered a type of metacognition that is provided by humans. Prompts are explicit instructions that are intended to direct computational processes during the completion of some task. However, LLMs are not very adept of generating their own task-specific prompts, which impedes their performance and autonomy. Generally, we think of metacognition as conscious, deliberate efforts to control and enhance cognitive processes, however, the repeated practice of a metacognitive strategy can result in it becoming automatic. An effective way to model this is through the compilation mechanism in ACT-R, which takes explicit strategies stored in declarative memory and compiles them into implicit productions in procedural memory, making them faster, automatic, and unconscious. ACT-R must be supplied with task-specific strategies from declarative memory, which is similar to humans receiving and remembering instructions, however, strategic instructions are not al- ways available for humans. Here it is important to distinguish between learning to do a particular task better and learning how to learn a task better. Learning how to learn is metacognitive learning. Metacognitive learning produces knowledge about different types of learning strategies, and where they are best applied. Although humans routinely do this, they are not always skilled at it. One example of humans exhibiting effective metacognitive learning is the activity of accomplished research scientists. In terms of AI, this ability for metacognitive learning can be understood in terms of three is- sues: (1) choosing the learning mechanism, (2) choosing how to structure the information, and (3) choosing how to adjust the parameters of the learning mechanism. In terms of LLMs, creating prompts corresponds to issue 2 and adjusting noise corresponds to issue 3. However, issue 1 is missing because learning is frozen (until the next update), and issues 2 and 3 are inputted by humans. We argue that metacognitive learn- ing in both humans and LLMs can be modelled through a dedicated learning mechanism that associates learning choices and parameters with outcomes across different classes of problems. This could potentially inform the development of more autonomous and versatile learning mechanisms in AI, as well as improved problem-solving capabilities and performance across diverse tasks.