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

Large Language Models (LLMs) lack robust metacognitive learning abilities and depend on human-provided algorithms and prompts for learning and output generation. Metacognition involves processes that monitor and enhance cognition. Learning how to learn - metacognitive learning - is crucial for adapting and optimizing learning strategies over time. Although LLMs possess limited metacognitive abilities, they cannot autonomously refine or optimize these strategies. Humans possess innate mechanisms for metacognitive learning that enable at least two unique abilities: discerning which metacognitive strategies are best and automatizing learning strategies. These processes have been effectively modeled in the ACT-R cognitive architecture, providing insights on a path toward greater learning autonomy in AI. Incorporating human-like metacognitive learning abilities into AI could potentially lead to the development of more autonomous and versatile learning mechanisms, as well as improved problem-solving capabilities and performance across diverse tasks.

Introduction

Currently, Large Language Models (LLMs) do not possess a robust set of self-directing learning abilities, and rely on human-designed algorithms, training data, and prompts to learn and generate outputs. For an LLM to become adept at metacognitive learning would require significant advancements in AI, including the development of AI systems capable of self-modification, self-assessment, and autonomous strategy development.

Metacognition is an array of cognitive processes that monitor and guide ordinary cognition in order to improve its functioning (Flavell 1979). For example, a student can recognize they learn better when they study in the morning instead of the evening. Generally, we think of metacognition as conscious, deliberate efforts to control and enhance cognitive processes, however, the practice of a metacognitive strategy can result in it becoming automatic (Conway-Smith, West, and Mylopoulos 2023). For example, a student's choice to work in the morning can become an automatic habit.

An effective way to model metacognitive automization is through the proceduralization mechanism in ACT-R, which takes explicit strategies stored in declarative memory and compiles them into automatic productions in procedural memory, making them faster and unconscious (Anderson 2013). Importantly, once compiled, the automatic productions (that do not consult declarative memory) compete with the original productions (that do consult declarative memory). This competition is significant as the productions must prove their utility across time and can be unlearned (via time-delayed learning algorithms). The proceduralization mechanism tested in experiments focused on learning strategies and shows similar speed up curves to humans (Anderson et al. 2019). However, we are unaware of any broader testing of this mechanism in situations where multiple strategies compete and their effectiveness varies over time or conditions. In theory, this mechanism should prevent automization except in cases where it is reliably effective.

The process of proceduralization in ACT-R requires that strategies have been stored in declarative memory. This maps most directly to humans who have decided that a strategy is useful, memorized it, and then practiced to make it automatic, which is an example of metacognition (i.e., cognition designed to influence, control, or improve cognition).

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 (Van Velzen 2015). While humans can ordinarily perform metacognitive learning, it takes practice to become skilled. An example is of a research scientist with expert knowledge of strategies for learning about different topics in their field.

Metacognitive Learning in LLMs

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 computa-

tional processes during the completion of some task. In addition to providing the initial prompt, a human prompt engineer monitors the LLM's responses and adjust the prompts on an ongoing basis to improve the output. This is very much like a human engaged in metacognitive learning. They begin with a strategy and then monitor and adjust it as they move forward. In the example of the research scientist, when exploring a novel problem they may begin with one research method but alter it when monitoring identifies it as not the best choice. By generalizing what is learned and how, the scientist understands which research strategies elicit useful results for a specific class of problem. Similarly, a prompt engineer learns which prompting techniques elicit the most useful results for a given class of interaction with the LLM.

We argue that LLMs such as Chat GPT, which have been trained on extensive databases, have some limited metacognitive abilities. Specifically, if an LLM has absorbed enough of the right content it is able to output metacognitive suggestions as part of its answers. For example, when Chat GPT is asked for advice on writing a scientific abstract it responded (among other things):

> "Maintain a clear focus on your objective and what you want to communicate. This helps to organize your thoughts and present your findings coherently."

> "Approach your writing with an organized plan. Decide on the structure of your abstract beforehand and allocate your focus accordingly."

In this sense, LLMs motivate a question that has been ignored in the ACT-R approach i.e., where do metacognitive strategies come from in the first place? In this example, one could argue that Chat GPT simply reproduced strategies that were in its learning set. Alternatively, one could argue that Chat GPT completely understands that metacognitive suggestions should be part of the answer. Either way, it is able to supply some forms of metacognitive strategies. This is not entirely different from humans, who mainly receive metacognitive strategies from other sources (e.g., teachers, books) and store them in declarative memory for later use.

However, beyond this, LLMs are limited in their ability to employ metacognition effectively. While a metacognitive strategy may be included in a prompt, this inclusion is not the same as actually applying the strategy. The prompt would allow the strategy to moderately influence the process but not to guide the process. Furthermore, LLMs cannot perform the two types of metacognitive learning that we described above — automatization and learning which metacognitive strategies are best. Both of these are important and related to each other. Metacognitive automatization compares and seeks the best metacognitive strategy using reinforcement learning. This method could also be used to find the best metacognitive learning strategy, but it would require more overhead. Also, we should consider whether or not the automatization learning algorithm is open for modification through metacognitive practice. A simple example of this would be learning to occasionally check if an already automatized procedure is still the best, something that humans are capable of but often struggle with. This is common in Cognitive Behavioural Therapy, where therapists will encourage clients to periodically re-evaluate their coping strategies to determine their current effectiveness (Beck 2020).

An important component of almost all cognitive architectures (Laird, Lebiere, and Rosenbloom 2017) is the separation of associative learning in declarative memory and reinforcement learning in procedural memory. The most direct way to implement this type of system would be to treat the LLM as declarative memory and implement prompt engineering in procedural memory. Procedural memory (in most architectures) represents the task using graph structures. Hence, some method to convert language outputs to graph-based outputs would be needed. To some degree, LLMs are already capable of this, and are able to interpret graph-based code as well. However, unlike LLMs which operate from a bottom-up approach, cognitive architectures, like humans, can exert strong top-down controls, effectively mimicking an expert system. This is also reflected in learning as procedural memory rewards are largely based on achieving task goals.

We argue that equipping LLMs with human-like metacognitive capabilities would require a distinct procedural module embodying the characteristics we've outlined. Such a module would allow for the ongoing refinement of internal strategies for prompt optimization, promoting autonomous learning and the creation of more effective prompts. This could be used to augment human-provided prompts or enable self-prompting, allowing the LLM to better act as an independent agent. Metacognition is particularly important for these agents, as it facilitates the self-monitoring and self-correction necessary for addressing safety and ethical issues.

References

Anderson, J. R. 2013. *Cognitive skills and their acquisition.* Psychology Press.

Anderson, J.; Betts, S.; Bothell, D.; Hope, R.; and Lebiere, C. 2019. Learning rapid and precise skills. *Psychological review,* 126(5), 727.

Beck, J. S. 2020. *Cognitive behavior therapy: Basics and beyond.* Guilford Publications.

Conway-Smith, B.; West, R. L; and Mylopoulos, M. 2023. Metacognitive skill: how it is acquired. In *Proceedings of the 45th Annual Conference of the Cognitive Science Society*.

Flavell, J. 1979. Metacognition and cognitive monitoring: A new area of cognitive inquiry. *American psychologist*, 34(10), 906.

Laird, J. E.; Lebiere, C.; and Rosenbloom, P. S. 2017. A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. *Ai Magazine,* 38, no. 4 (2017): 13-26.

Van Velzen, J. 2015. *Metacognitive learning.* New York, NY: Springer International Publishing.