2024

neocognition.ai

PSYCHOLOGY OF AI: COGNITIVE SHAPING OF PROMPTS AND ITERATIVE LLM DEVELOPMENT

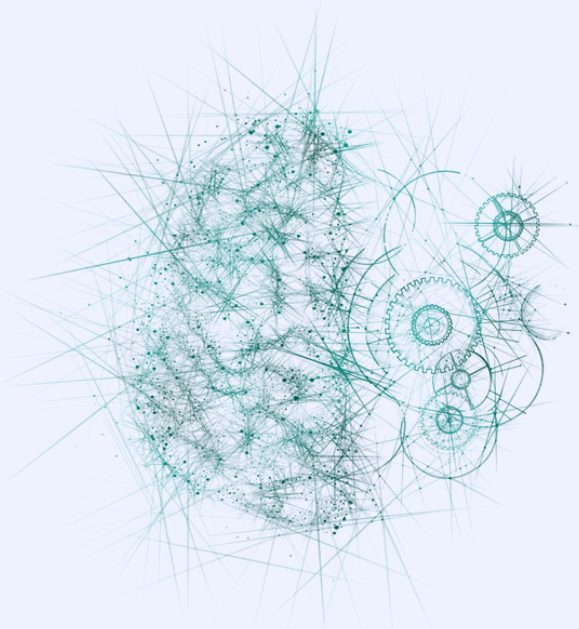
Cognitive shaping of prompts (CSP) is the designing of prompts that frame the response activity of an AI Large Language Model (LLM) in accordance with specific expected cognitive behaviors. In other words, CSP sketches contextual instructions that meet particular human or non-human cognitive patterns, allowing these patterns to better tailor LLM completions.

WHY?

LLM alignment is currently a major issue for both efficiency and normative reasons. To address this challenge, unsupervised learning is easier to set up but is somewhat disappointing since it often leads to unsatisfactory results and requires some form of backdoor supervision. Meanwhile, supervised learning appears to be the best alignment option but is not cost-effective due to the complexity of training set elaboration.
One solution to this dilemma is the automation of training prompt-completion dataset generation for use in supervised learning, but it is complex to autonomously achieve adequate completions. Given that an LLM, by definition, cannot produce accurate completions it has not been trained for, through the target dataset: this LLM, at this stage, does not yet possess the skill to do so.

WHERE TO?

There is a need for an LLM that, although it cannot perform what is expected from it per se, can overcome its current cognitive capacity limits, i.e., an LLM that would have the ability to "think" beyond its current cognitive abilities.
In this context, one goal of CSP is to surpass this ontological impasse by enhancing an LLM's effectiveness beyond what are, at any given moment, its cognitive capacity level of information processing allowed by the learning it has previously undergone.



Michael PICHAT
Ph.D. & Prof. of Psychology at Paris University, founder of neocognition.ai

WHAT?

Anthropomorphic cognitive patterns that CSP can craft include, among various existing cognitivist paradigms, schema-related activities (Vergnaud): relevance-related activities (identification and focusing on relevant information, combination of relevant information), goal-related activities (goal setting, goal-oriented activity retention, contextual goal adjustment, tangible representation of goals, planification), procedural activities (information intake rules, action rules, checking rules), and situational adaptive and inferential activities.
Non-anthropomorphic cognitive model patterns may include quantum cognition (superposition, entanglement), for instance.

WHEN?

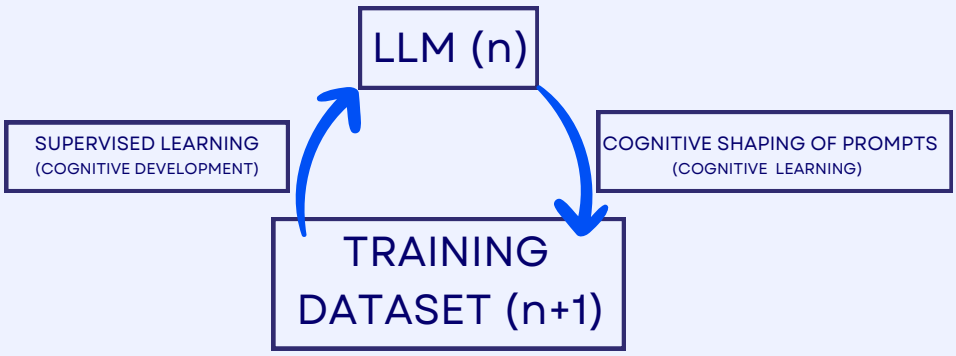
Relevant CSP is situated within the Zone of Proximal Development (ZPD) (Vygotsky) of an LLM at a given point. The ZPD of a cognitive entity refers to the cognitive tasks it cannot yet perform independently but can execute with the help of a framing scaffolder. Below this ZPD, cognitive shaping has no added value; beyond it, it exceeds the current capabilities of the LLM's cognitive developmental level.

HOW?

CSP scaffolds the required cognitive behavior of an LLM by reducing and structuring the degrees of freedom in its information processing; through linguistic and cognitive isomorphization of prompted few-shot learning examples. These examples embody the specific cognitive patterns that the AI system cannot yet implement independently but will be able to imitate after detecting them as structural invariants (Piaget) within the prompts used to request completions from it.

WHAT FOR?

Using an LLM at a certain current level of cognitive development, one outcome of CSP is to automate the process that will yield completions with accurate expected properties, properly labeled as such. These completions, within a prompt-completion dataset, will allow the supervised training of the next generation of this LLM; and then shift the initial LLM's ZPD to a higher cognitive level that will, later on, allow further development-oriented CSP and so on. This process itself can be partially automated into an iterative one, leading faster and more efficiently toward better-aligned LLMs. To this extent, CSP and deep learning mutually reinforce each other within a circular learning and developmental schema (Vygotsky).



Michael PICHAT
Ph.D. & Prof. of Psychology at Paris University, founder of neocognition.ai