15/05/2023

CONVERSATIONAL AI & SYNTHETIC COGNITIVE BIASES: THE CASE OF AVAILABILITY BIAS

Michael PICHAT, Ph.D. in psychology of cognitive processes, university lecturer & researcher, founder of neocognition

Conversational artificial intelligence psychology (CAIP) applies principles of scientific psychology to analyze and adapt the cognitive and linguistic mechanisms specific to conversational AI agents, making them more functional. CAIP draws from various fields of human psychology, using analogous and anthropomorphic transpositions of it, to create what can be described as artificial intelligence psychology.

CAIP focuses on the cognitive configuration of conversational AI agents, tailoring their information processing modalities. Technically, this calibration of synthetic cognitive processes occurs during several key stages of conversational AI's cognitive optimization.

Among these stages is the design of prompt engineering specifically crafted to mitigate synthetic cognitive biases that conversational AI is susceptible to, due to the mathematical and statistical methods used to process the data on which they are trained. From an operational standpoint, conversational AI psychology provides recommendations to human users for creating instructions that minimize the potential of activation and impact of cognitive biases that are involved.

A variety of synthetic cognitive biases significantly impact the quality of conversational AI responses and contribute to their intriguing and necessary psychological study, such as priming, framing, anchoring, recency, and hypothesis confirmation, to name just a few.

THE AVAILABILITY BIAS CASE

Availability bias (Kahneman & Tversky, 1973) occurs when a language model produces a response related to an object (person, physical object, situation, or phenomenon) mentioned in a prompt, this response being heavily based on the most accessible data by the model, which may be stereotypical or simplistic. The existence of this synthetic bias stems from the fact that the model is built through loss function optimization, defined as the measure of the difference between the model's predictions and actual responses from the training dataset. A particular type of overrepresented data in a training dataset will thus be learned more by the model and more proposed as a response, resulting in availability bias.

For example, synthetic availability bias can lead to stereotyping responses (generational, gender-based, ethnic, etc.) that have been well-documented (e.g., "Nurses, who are often women, are responsible for patient care and comfort"). It can also produce simplistic and unnuanced outputs, as one-dimensional responses are frequently found in training datasets (e.g., "Lawyers spend most of their time arguing cases in court").

In prompt engineering, human users should calibrate their requests to minimize the likelihood of availability bias manifestation. This shall be achieved by adding specific antidote markers to their prompts (i) explicitly asking for nuanced, diverse, and non-stereotypical responses (e.g., "What are some lesser-known and varied activities that programmers can perform?" instead of "What are the main activities of programmers?") and (ii) providing clear contexts or examples to help the model understand the expected response (e.g., "What are important emotional and interpersonal skills for a leader, in addition to management and decision-making abilities?" instead of "What are the important skills to be a good leader?").