15/05/2023

INTELLIGENCES ARTIFICIELLES CONVERSATIONNELLES & BIAIS COGNITIFS SYNTHÉTIQUES : CAS DU BIAIS DE DISPONIBILITÉ

Michael PICHAT, docteur en psychologie des processus cognitifs, maître de conférences des universités, fondateur de neocognition.ai

La psychologie de l'intelligence artificielle conversationnelle (PIAC) mobilise les principes de la psychologie scientifique pour analyser et adapter les mécanismes cognitifs et linguistiques propres aux agents conversationnels artificiels, afin de les rendre plus fonctionnels. Elle tire parti de divers domaines de la psychologie humaine en la transposant, de façon analogique et anthropomorphique, à ce qui peut être décrit comme une psychologie de l'intelligence artificielle.

La PIAC s'attache à la configuration cognitive des agents conversationnels artificiels, en paramétrant leurs modalités de traitement de l'information. Sur le plan technique, ce calibrage des processus cognitifs synthétiques s'effectue lors de plusieurs étapes clés de l'optimisation cognitive des intelligences artificielles conversationnelles.

Parmi ces étapes, la conception de prompts engineerings spécifiquement conçus pour atténuer les biais cognitifs synthétiques auxquels les IA conversationnelles sont sensibles en raison des méthodes mathématiques et statistiques utilisées pour traiter les données sur lesquelles elles sont entraînées. D'un point de vue opérationnel, la psychologie de l'IA conversationnelle offre alors des préconisations aux utilisateurs humains pour créer des instructions qui minimisent le potentiel d'activation, et donc l'impact, des biais cognitifs potentiellement impliqués.

CAS DU BIAIS DE DISPONIBILITÉ

Le biais cognitif de disponibilité (Kahneman & Tversky, 1973) est le fait, pour un modèle de langage, de produire une réponse relative à un objet (personne, objet physique, situation ou phénomène) mentionné dans un prompt, réponse qui est fortement basée sur des données (éventuellement stéréotypées ou simplistes) auxquelles le modèle a le plus accès. L'existence de ce biais synthétique provient du fait que ce modèle est fabriqué par optimisation de la fonction de perte, définie comme la mesure de l'écart entre les prévisions du modèle et les réponses effectives de l'ensemble de donnée sur lequel il est entrainé. Un type particulier de donnée sur-représenté dans un jeu de donnée d'entrainement sera ainsi plus appris et proposé en réponse par le modèle, réalisant ipso facto un biais de disponibilité.

A titre d'illustration, le biais synthétique de disponibilité provoque bien entendu des réponses de stéréotypisation (générationnels, genrés, ethniques, etc.) déjà largement documentées (exemple : « les infirmières, qui sont souvent des femmes, sont en charge des soins aux patients et de leur réconfort »). Il est également de nature à provoquer des outputs simplistes et manquant de nuance dans la mesure où les réponses unidimensionnelles sont souvent celles qui sont les plus fréquentes dans les datasets d'entrainement (exemple : « Les avocats passent la plupart de leur temps à plaider devant les tribunaux »).

En matière de prompt engineering, il conviendra dès lors à l'utilisateur humain de calibrer ses demandes de façon à minimiser la possibilité de manifestation du biais de disponibilité. Cela, en dotant ses prompts de balises antidotes spécifiques (i) demandant explicitement des réponses nuancées, diversifiées et non stéréotypées (« Quelles sont certaines activités variées et moins connues que les programmeurs peuvent effectuer » et non pas « Quelles sont les principales activités des programmeurs ») et (ii) fournissant des contextes ou des exemples clairs pour aider le modèle à comprendre la réponse attendue (« Quelles sont les compétences émotionnelles et interpersonnelles importantes pour un leader, en plus des compétences en gestion et prise de décision » et non pas « Quelles sont les compétences importantes pour être un bon leader »).

Une variété d'autres biais cognitifs synthétiques sont particulièrement de nature à impacter la qualité des réponses des IA conversationnelles et nourrissent allègrement leur passionnante comme nécessaire étude psychologique : l'amorçage, le cadrage, l'ancrage, la récence, la confirmation d'hypothèse pour n'en citer qu'un nombre limité.