Évaluation des capacités de réponse de larges modèles de langage (LLM) pour des questions d'historiens

Mathieu Chartier*, Nabil Dakkoune**,
Guillaume Bourgeois* et Stéphane Jean**

*CRIHAM, université de Poitiers
**LIAS, ISAE-ENSMA et université de Poitiers
prenom.nom@univ-poitiers.fr

Résumé. Les larges modèles de langage (LLM) tels que ChatGPT ou Bard ont bouleversé la recherche d'informations et conquis le public par leur facilité à générer des réponses sur mesure en un temps record, qu'importe le sujet. Dans cet article, nous analysons les capacités de différents LLM à produire des réponses sur des faits historiques en français avec fiabilité, exhaustivité et suffisamment de pertinence pour être directement exploitables ou extractibles. Pour cela, nous avons élaboré un banc d'essai constitué de multiples questions d'histoire. Ces dernières sont de différents types, thématiques et de niveaux de difficulté variables. Notre évaluation des réponses fournies par dix LLM, que nous avons jugés pertinents, montre de nombreuses limites sur le fond comme dans la forme. Au-delà d'un taux de précision globalement insuffisant, nous mettons en évidence le traitement inégal du français ou encore des problèmes de loquacité et d'inconstance des réponses fournies par les LLM.

1 Introduction

Le succès et l'engouement rencontrés par ChatGPT d'OpenAI depuis son lancement le 30 novembre 2022 mettent en évidence les progrès des larges modèles de langage (LLM) auprès du grand public, atteignant un million d'utilisateurs actifs seulement cinq jours après l'annonce officielle et même cent millions deux mois plus tard. Les LLM démontrent des capacités en termes de production et complétion de textes, de traduction automatique mais aussi de génération de réponses à des questions de tout ordre. Ces outils se révèlent notamment performants dans ces différentes tâches grâce à un entraînement non supervisé ou semi-supervisé sur d'immenses corpus textuels, totalisant des milliards de mots parmi des sources diverses et variées.

Avec de telles masses d'informations ingurgitées, nous pouvons estimer que ces modèles sont entraînés sur des volumes de connaissances paraissant dépasser les capacités mémorielles humaines (Bowman (2023)), laissant penser à tort qu'ils auraient réponse à tout. En effet, bien que les LLM s'appuient sur des millions de sources diversifiées, ils sont capables de produire avec une certaine assurance des réponses biaisées voire totalement fausses ou farfelues (Zheng et al. (2023)), mais aussi développer un certain niveau d'hallucination (Ji et al. (2023)). Par conséquent, il semble nécessaire d'analyser leur aptitude à générer des réponses approfondies