Les performances des IA génératives sont-elles gonflées artificiellement ?

Benjamin Polge
JDN

Mis à jour le 11/09/24 10:17

Les résultats de la majorité des grands modèles de langage aux principaux benchmarks du marché sont biaisés. Récit d'une course aux résultats partiellement faussée.

MMLU, HumanEval, MATH… Les benchmarks censés mesurer les performances des LLM n'ont jamais été aussi scrutés que ces deux dernières années. Depuis la publication de ChatGPT en novembre 2022, la course aux performances ne cesse de s'accélérer. En tête le trio Google-OpenAI-Anthropic continue de sortir très régulièrement des modèles textuels et aujourd'hui multimodaux affichant des performances toujours plus hautes dans les benchmarks.

Cette saine concurrence, en apparence, n'en est pas moins artificielle. La communauté scientifique de l'IA s'accorde à dire que les performances affichées dans les benchmarks et dans la réalité diffèrent grandement. Quelles en sont les raisons exactes ? Voici de premiers éléments de réponse.

Des benchmarks aux données ouvertes

Le principal biais des benchmarks vient de leur mode de conception même. Leur point fort est également leur plus grand point faible : la transparence des données de test. A l'heure actuelle, la majorité des benchmarks du marché mettent en accès libre les questions et les problèmes soumis au modèle lors de la phase de test .Cette transparence accrue permet ainsi aux développeurs de modèle de comprendre assez finement la nature des questions et du raisonnement qui sera alors adressé au modèle pendant la phase de benchmarking. Partant de ce point, certains éditeurs de modèles n'hésiteraient pas à affiner leur IA avant le benchmarking en prenant pour référence un dataset proche de celui du benchmark.

"Il y a cette tendance à réaliser un alignement en fin de pré-entraînement. Sans l'appeler du fine-tuning, on cherche à exposer le modèle aux données des benchmarks. Bien sûr, cela se fait intelligemment : on n'utilise pas 100% du dataset. Par exemple, pour le MMLU avec ses 16 000 questions, si on en utilise 3 000 à 4 000, il y a de fortes chances que le modèle performe bien lors des tests, puisque les benchmarks sont ensuite tirés aléatoirement", rappelle Michel-Marie Maudet cofondateur de LINAGORA et faciliteur auprès d'OpenLLM France. Une pratique qui peut influer assez notablement sur le résultat final du benchmark. Le cas de MMLU (Massive Multitask Language Understanding) est assez parlant. Le benchmark est composé de plusieurs milliers de questions visant à évaluer la capacité des modèles de langage à comprendre et à raisonner à travers unelarge variété de domaines et de disciplines. Les résultats des modèles peuvent grandement varier grâce à certaines techniques.

Pour booster les performances, après avoir procédé à un entraînement du modèle assez long, en intégrant beaucoup de mathématiques, de contenu scientifique et de données synthétiques dans le jeu de pré-entraînement, les développeurs procèdent à une phase appelée "handling." Cette dernière repose "sur les derniers 1-2% de l'entraînement, où l'on expose le modèle à encore plus de contenu synthétique de très haute qualité, axé sur les mathématiques, les sciences et les bases de connaissances. Ces données sont très similaires à celles du MMLU pour booster les performances du modèle au dernier moment", explique avec détails Manuel Faysse, doctorant au MICS de CentraleSupélec, spécialiste des modèles de NLP.

Une technique qui permettrait, de gagner de précieux points lors de la notation finale du modèle. "Des études ont montré qu'en ajoutant simplement cette phase finale à un modèle ayant suivi le même pré-entraînement, on pouvait gagner 10-12% sur les scores MMLU. Cela démontre que le modèle n'est pas nécessairement meilleur de manière intrinsèque, mais qu'il a été optimisé pour ce benchmark spécifique", détaille encore le chercheur.

Pour exemple, le modèle Phi de Microsoft se classe toujours très haut sur MMLU mais dans les faits, lorsqu'on l'utilise en conditions réelles il n'est pas forcément meilleur. Le cas du GSM8K, un test de mathématiques réputé, illustre parfaitement cette problématique. Initialement considéré comme une référence, il a rapidement montré ses limites : de nombreux modèles y excellaient sans pour autant briller dans des applications réelles. Face à ce constat, une version plus courte et inédite, le GSM1K, a été développé. Cette nouvelle mouture a permis de mettre en lumière des disparités intéressantes : certains modèles maintiennent de bonnes performances sur les deux versions, tandis que d'autres, voient leurs résultats chuter sur le nouveau test. "Lorsque de nouvelles évaluations apparaissent, on se rend parfois compte que certains modèles ont été un peu trop overkill sur certains jeux de données spécifiques", s'amuse Manuel Faysse.

Les benchmarks "dynamiques" aussi concernés

Les benchmarks "dynamiques", dont les tests sont menés par des humains, ne sont pas non plus exempts de biais. Par exemple, la chatbot arena de LMSys met en avant les modèles dont la réponse est jugée la plus satisfaisante par un panel d'utilisateurs humains. Or "les évaluateurs ont tendance à préférer les réponses plus longues. Si on force un modèle à générer des réponses de 200 mots plutôt que de 50, elles seront en moyenne mieux notées, même si le contenu n'est pas nécessairement meilleur", avance le spécialiste.

L'autre biais principal de la majorité des benchmarks du marché, souvent mis en avant dans l'Hexagone, reste l'absence de véritable benchmark en langue française. "Bien qu'on puisse supposer qu'un modèle performant en anglais le sera aussi dans d'autres langues, cette hypothèse n'est pas toujours vérifiée ", souligne Michel-Marie Maudet.

Sur quelles données s'appuyer pour choisir son LLM ?

Partant du principe que les benchmarks ne donnent qu'une idée approximative des capacités réelles du modèle, sur quelles données faut-il s'appuyer pour choisir son LLM ?
Manuel Faysse recommande d'éviter de baser son choix uniquement sur les benchmarks et conseille d'opter pour le benchmarking privé. Après avoir identifié votre cas d'usage, il sera intéressant de tester les performances intrinsèques de chaque modèle. Le choix devrait alors logiquement s'orienter sur le modèle ayant les meilleures performances selon vos propres métriques, toujours adaptées au cas d'usage.

Enfin, pour des cas d'usage moins critiques ou plus polyvalents, le choix d'un modèle avec des performances élevées dans les benchmarks généraux peut être une approche pragmatique et efficace. Ces modèles bien classés offrent généralement une bonne polyvalence et des capacités solides dans divers domaines, ce qui peut convenir à de nombreuses applications courantes.