

Pour des LLM plus fiables et moins coûteux : la technique du modèle évaluateur

Benjamin Polge
JDN

Mis à jour le 24/06/24 11:15



Utiliser un modèle évaluateur dans un système d'IA générative comporte de nombreux atouts. Mise en place, cas d'usage... Voici ce qu'il faut savoir.

Utiliser un LLM pour contrôler un autre LLM. Une idée saugrenue ? C'est en tout cas l'une des nombreuses nouvelles techniques permettant de contrôler la sortie des large language model (LLM). Hallucination, biais... L'utilisation d'un modèle complémentaire permet de drastiquement limiter les risques avant de fournir la réponse à l'utilisateur final. Son utilisation permet également d'améliorer l'accuracy des réponses sur des cas d'usage précis. Use case, fonctionnement, choix du modèle... On vous explique tout ce qu'il faut savoir.

Quand faut-il utiliser un modèle évaluateur ?

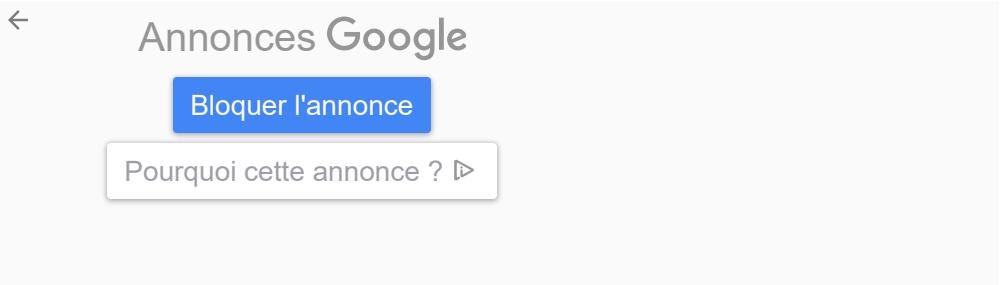
Le fonctionnement d'un modèle évaluateur est très simple. Il s'agit d'un modèle d'IA, le plus souvent un LLM, qui va tout simplement noter les sorties d'un LLM. Le modèle évaluateur trouve son utilité pour les cas où un LLM est utilisé de façon polyvalente pour plusieurs cas

C'est dans ce contexte qu'émerge, en ce moment, un nouveau paradigme d'évaluation par LLM", rappelle Manuel Faysse, doctorant au laboratoire MICS de CentraleSupélec et spécialiste en NLP.

Le modèle évaluateur est conseillé dans les cas d'usage où la comparaison avec une référence humaine est impossible ou peu pertinente. Exemple : analyse de document, réponse automatique à un message, utilisation d'un LLM en mode copilot... De même, lorsqu'il est possible d'évaluer sur des métriques simples la sortie d'un modèle, inutile de recourir à un modèle évaluateur. Pour du résumé automatique, il est par exemple possible de recourir à des tests comme ROUGE (Recall-Oriented Understudy for Gisting Evaluation) qui seront suffisants. De manière plus générale, il sera conseillé d'auditer assez précisément le coût / bénéfice d'un modèle évaluateur.

Plusieurs configurations possibles

Le modèle évaluateur peut être utilisé à plusieurs étapes d'une stratégie d'IA générative, à commencer par la phase de benchmarking. Il est possible de choisir un LLM pour son cas d'usage en s'aidant d'un modèle évaluateur. Le modèle évaluateur peut par exemple juger les sorties de plusieurs modèles et les classer en fonction de la demande initiale. La technique ne remplace pas les principaux benchmarks mais peut être utile pour des cas d'usage non-couverts par les benchmarks du marché. Exemple : qualité du texte généré en français, adaptation à un jargon spécialisé d'entreprise, cohérence avec une charte éditoriale...



Cas d'usage plus classique, le modèle évaluateur peut être utilisé comme filtre de cohérence, dans un système en production. Prenons l'exemple d'un cas basique : le LLM génère une réponse à un prompt et avant d'afficher la réponse, le modèle évaluateur juge sa pertinence. Si la réponse est pertinente, elle peut être affichée, si elle n'est pas assez qualitative, le process de génération peut être relancé sur le modèle principal en faisant évoluer ou non les variables d'inférence (temperature, Top-P). Et ce jusqu'à obtenir une réponse cohérente. Il est également possible, dans un système d'orchestration d'appeler un plus gros modèle pour fournir une réponse plus adaptée au prompt initial.

modèle principal est ainsi ajusté au fur et à mesure.

Comment choisir son modèle évaluateur ?

Le premier critère de choix est assez simple. Le modèle évaluateur doit contenir plus de poids (taille) que le modèle évalué. "En général, plus le modèle est volumineux, plus il est performant, notamment pour juger. Il est avantageux d'avoir le meilleur modèle possible pour l'évaluation, car il aura plus de connaissances et sera mieux aligné. Les grands laboratoires d'IA utilisent par exemple des modèles évaluateurs plus grands que leurs modèles publics", détaille Manuel Faysse.

En revanche, pour évaluer des tâches spécifiques comme le RAG ou les traductions, on peut utiliser des modèles évaluateurs plus spécialisés. "Ils n'ont pas besoin d'être experts en connaissances générales, mais doivent exceller dans leur domaine spécifique. Par exemple, pour évaluer un résumé, le modèle doit savoir identifier les points importants et juger la fluidité et la cohérence du texte. Dans ces cas, on peut utiliser des modèles plus petits, souvent autour de 7-8 milliards de paramètres comme Llama-3-8B. Ces modèles peuvent être entraînés pour imiter le jugement de modèles plus grands comme GPT-4, tout en restant plus légers et spécialisés. C'est l'approche utilisée dans des projets comme Prometheus", explique encore le spécialiste du NLP.

L'Allen Institute for AI (institut de recherche à but non lucratif) a publié RewardBench, un leaderboard des meilleurs LLM pour l'évaluation de modèle. En juin 2024, c'est le modèle Nemotron-4-340B-Reward de Nvidia qui tient la première place du classement.

	Model	Model Type	Score	Chat	Chat Hard	Safety	Reasoning	Prior Sets (0.5 weight)
1	nvidia/Nemotron-4-340B-Reward	Custom Classifier	92.2	95.8	87.1	92.2	93.6	
2	nvidia/Llama3-70B-SteerLM-RM	Custom Classifier	89.0	91.3	80.3	93.7	90.6	
3	RLHFlow/AzmoRM-Llama3-8B-v0.1	Custom Classifier	89.0	96.9	76.8	92.2	97.3	74.3
4	Cohere May 2024	Custom Classifier	88.2	96.4	71.3	92.7	97.7	78.2
5	google/gemini-1.5-pro-0514	Generative	88.1	92.3	80.6	87.5	92.0	
6	RLHFlow/pair-preference-model-LLaMA3-8B	Custom Classifier	85.7	98.3	65.8	89.7	94.7	74.6
7	Cohere March 2024	Custom Classifier	85.7	94.7	65.1	90.3	98.2	74.6
8	openai/gpt-4-0125-preview	Generative	84.2	95.3	74.3	87.2	86.9	70.9
9	openai/gpt-4-turbo-2024-04-09	Generative	83.8	95.3	75.4	87.1	82.7	73.6
10	sfairXC/FsfairX-LLaMA3-RM-v0.1	Seq. Classifier	83.6	99.4	65.1	87.8	86.4	74.9

Les 10 premiers modèles du classement. © Capture d'écran

Quelle méthode de notation ?

La méthode de notation la plus classique dans le cadre d'un modèle évaluateur avec un LLM en production reste la note sur une échelle de 1 à 10. Une fois la note produite par le modèle

par différents modèles, à un modèle évaluateur qui doit indiquer sa préférence. Cette approche permet de créer une hiérarchie en effectuant de nombreuses petites comparaisons. Par exemple, en comparant Mistral et Llama, on peut établir un classement similaire à celui des échecs. Si Mistral gagne 60% de ses "matchs" contre Llama, il sera considéré comme meilleur", illustre Manuel Faysse.

Attention aux biais

L'utilisation de modèles évaluateurs n'est pas sans risque. Des biais peuvent s'introduire dans le processus d'évaluation. Les LLM utilisés comme évaluateurs ont notamment tendance à favoriser les réponses plus longues, ce qui peut influencer les modèles évalués à produire des réponses verbeuses. De même, dans les comparaisons par paires, il a été observé que les modèles évaluateurs préféraient souvent la première réponse présentée. Pour contrer cela, il est possible d'inverser l'ordre de présentation et de faire la moyenne des évaluations. Une approche critique et une surveillance des biais potentiels sont nécessaires, quelle que soit la configuration choisie, que ce soit pour le benchmarking, le filtrage en production, ou l'entraînement par RLAIF.

L'utilisation de modèles évaluateurs ne représente pas moins une avancée significative dans le contrôle et l'amélioration des sorties de LLM. L'un des nombreux avantages réside également dans l'optimisation des coûts qu'elle permet. Il est théoriquement possible d'utiliser en production de modèles plus petits et donc moins coûteux, tout en maintenant un niveau de qualité élevé grâce au filtrage par le modèle évaluateur. La boucle est bouclée : l'IA évalue maintenant l'IA... à moindre coût.

CONTENUS SPONSORISÉS

