EVALUATION AUTOMATIQUE DES REPONSES EN LANGAGE NATUREL

Proposé par Redha Moulla
redha_moulla@yahoo.fr

Description du projet
Le projet consiste à utiliser des LLMs (Large Language Models) pour l'évaluation automatique des réponses des étudiants rédigées en langage naturel. Le modèle doit être capable à la fois de donner des notes pertinentes, aussi proches que possible de celles données par l'enseignant, et de proposer un feedback pour justifier sa note ; il doit ainsi mettre en évidence les séquences pertinentes comme les éventuelles séquences fausses dans la réponse de l'étudiant. Pour ce faire, le modèle doit se référer à la réponse donnée par l'enseignant ; celle-ci est désignée par la suite par « réponse référence ».
Le choix du LLM (ChatGPT, Gemini, Mistral, etc.) est laissé à la discrétion des étudiants. Cependant, un benchmarking entre différents modèles sera fortement apprécié.

Données
Les données seront mises à la disposition des étudiants. Elles consistent en un dataset de plusieurs questions de machine learning, auxquels 38 étudiants ont répondu. Ainsi, à chaque question sont associées 38 réponses potentielles, déjà évaluées par un enseignant, et une réponse référence.

Méthodologie
L'approche pour aborder cette problématique peut être décomposée en trois étapes graduelles.

Étape 1 : il s'agit de l'approche la plus immédiate et la plus élémentaire. Elle consiste simplement à prompter le LLM avec la question et la réponse référence pour lui demander d'évaluer la réponse de l'étudiant et de proposer un feedback. Le choix du prompt peut néanmoins être déterminant dans la performance du LLM.

Étape 2 : dans un deuxième temps, des prompts plus sophistiqués de type *Chain of Thoughts* (CoT) peuvent être envisagés pour améliorer les performances du LLM. Il s'agit, plus concrètement, d'apprendre au LLM le raisonnement sous-jacent à l'évaluation de l'enseignant ; un ensemble de 5 exemples peut être fourni au LLM, qui doit extrapoler ensuite sur les réponses restantes. Autrement dit, il faut inclure dans le prompt, outre la réponse référence, 5 exemples de corrections avec les notes et les feedbacks correspondants

Étape 3 : celle-ci consiste à fine-tuner un LLM open source, de type Mistral, sur un dataset contenant à la fois les questions, les réponses, les notes et les feedbacks souhaités. La construction du dataset est une étape importante qui peut impliquer la recherche de nouvelles données, en utilisant par exemple un modèle comme GPT-4 pour l'annotation automatique (génération de feedbacks).

Livrable

Le livrable est un notebook qui inclut à la fois le code Python et les commentaires nécessaires. A chaque étape, la méthodologie doit être décrite et documentée. Les résultats obtenus doivent être mis en évidence, aussi bien les notes que les feedbacks.

Évaluation

L'évaluation des rendus sera basée sur les performances obtenus en considérant à la fois les notes et les feedbacks données par le modèle. S'agissant des notes, dans la mesure où le barème utilisé dans le dataset fourni comprend seulement trois valeurs (0, 1, 2), les métriques de classification seront privilégiées, en particulier l'accuracy et le F1-score. Quant à l'évaluation des feedbacks, elle sera basée sur un échantillon de 5 réponses tirées aléatoirement pour 4 questions tirées également aléatoirement. Ces réponses ne doivent évidemment pas avoir été utilisées dans les exemples fournis dans les CoT ou le fine-tuning.

Bibliographie

Sultan, M. A. (2016, June). Fast and easy short answer grading with high accuracy. *In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* , pp. 1070-1075.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824-24837.