Apprentissage par renforcement pour l'alignement des agents LLMs avec des environnements interactifs : quantification et réduction du surapprentissage aux prompts

Mohamed Salim Aissi[1], **Clement Romac**[2,3], **Thomas Carta**[3], **Sylvain Lamprier**[4], **Pierre-Yves Oudeyer**[3], **Olivier Sigaud**[1], **Laure Soulier** [1], **Nicolas Thome** [1]
[1]Sorbonne Université, CNRS, ISIR, F-75005 Paris, France [2]Hugging Face
[3]Inria (Flowers), University of Bordeaux, France [4]Univ Angers, LERIA, Angers, France
{salim.aissi, laure.soulier, olivier.sigaud, nicolas.thome}@isir.upmc.fr
{clement.romac, thomas.carta, pierre-yves.oudeyer}@inria.fr
{sylvain.lamprier}@univ-angers.fr

RÉSUMÉ

L'apprentissage par renforcement constitue une approche prometteuse pour aligner les connaissances des Grands Modèles de Langue (LLMs) avec des tâches de prise de décision séquentielle. Cependant, peu d'études ont analysé en profondeur l'impact de l'ajustement des LLMs par apprentissage par renforcement dans un environnement spécifique. Dans cet article, nous proposons un nouveau cadre d'analyse pour évaluer la sensibilité des LLMs aux formulations de prompt après un entraînement par renforcement dans un environnement textuel. Nos résultats montrent que la performance des LLMs se dégrade lorsqu'ils sont confrontés à des formulations de prompt différentes de celles utilisées durant la phase d'entraînement par renforcement. Par ailleurs, nous analysons l'origine de cette sensibilité en examinant les représentations internes du modèle ainsi que les tokens saillants. Enfin, nous proposons l'utilisation d'une fonction de coût contrastive afin d'atténuer cette sensibilité et d'améliorer la robustesse et les capacités de généralisation des LLMs.

ABSTRACT

Reinforcement Learning for Aligning Large Language Models Agents with Interactive Environments : Quantifying and Mitigating Prompt Overfitting.

Reinforcement learning (RL) is a promising approach for aligning large language models (LLMs) knowledge with sequential decision-making tasks. However, few studies have thoroughly investigated the impact on LLM agents capabilities of fine-tuning them with RL in a specific environment. In this paper, we propose a novel framework to analyze the sensitivity of LLMs to prompt formulations following RL training in a textual environment. Our findings reveal that the performance of LLMs degrades when faced with prompt formulations different from those used during the RL training phase. Besides, we analyze the source of this sensitivity by examining the model's internal representations and salient tokens. Finally, we propose to use a contrastive loss to mitigate this sensitivity and improve the robustness and generalization capabilities of LLMs.

MOTS-CLÉS : LLM, Apprentissage par renforcement, Prise de décision séquentielle.

KEYWORDS: LLM, Reinforcement Learning, Sequential Decision Making.

ARTICLE : **Accepté à** NAACL2025 Findings
lien : https://aclanthology.org/2025.findings-naacl.390/.