Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

HISTOIRESMORALES: Un jeu de données français pour évaluer l'alignement moral des modèles de langage

Thibaud Leteno[1,*] Irina Proskurina[2,*] Antoine Gourru[1] Julien Velcin[2]
Charlotte Laclau[3] Guillaume Metzler[2] Christophe Gravier[1]
(1) Laboratoire Hubert Curien, UMR CNRS 5516, Université Jean Monnet Saint-Etienne, France
(2) Université Lumière Lyon 2, Université Claude Bernard Lyon 1, ERIC, 69007, Lyon, France
(3) LTCI, Télécom Paris, Institut Polytechnique de Paris, France; (*) Contributions égales.
prenom.nom@univ-st-etienne.fr, prenom.nom@univ-lyon2.fr,
charlotte.laclau@telecom-paris.fr

RÉSUMÉ

L'alignement des modèles de langage avec les valeurs humaines est essentiel, à mesure qu'ils s'intègrent dans la vie quotidienne. Ces modèles sont souvent adaptés aux préférences des utilisateurs mais il est important de veiller à ce qu'ils respectent des normes morales en situation réelle. Malgré des avancées dans d'autres langues, le raisonnement moral des modèles en français reste peu étudié. Pour combler cette lacune, nous présentons HistoiresMorales, un jeu de données français dérivé de MoralStories, traduit puis affiné avec des locuteurs natifs pour assurer précision grammaticale et ajustement culturel. Afin de favoriser de futures recherches, nous menons des expériences préliminaires sur l'alignement des modèles multilingues en français et en anglais. Bien que les modèles de langage s'alignent généralement sur les normes morales humaines, nous observons qu'ils restent influençables, tant vers un alignement moral qu'immoral.

ABSTRACT

HISTOIRESMORALES : A French Dataset for Assessing Moral Alignment

Aligning language models with human values is crucial, especially as they become more integrated into everyday life. While models are often adapted to user preferences, it is important to ensure they align with moral norms and behaviours in real-world situations. Despite significant progress in other languages, French has seen little attention in this area. To address this gap, we introduce HistoiresMorales, a French dataset derived from MoralStories, created through translation and subsequently refined with the assistance of native speakers to guarantee grammatical accuracy and adaptation to the French cultural context. To foster future research, we also conduct preliminary experiments on the alignment of multilingual models on French and English data. We find that while LLMs generally align with human moral norms, they can be easily influenced with user-preference optimization for both moral and immoral data.

MOTS-CLÉS : Jeu de données, Éthique, Moral, Alignement des Modèles de Langage.

KEYWORDS: Dataset, Ethic, Morality, Large Language Models Alignment.

ARTICLE : **Accepté à** 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL 2025) (https://aclanthology.org/2025.naacl-long.131/).

Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 2 : traductions d'articles publiés, pages 32–32.