Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Un formalisme fondé sur des approximations du problème du sac à dos pour modéliser l'alignement

Ayoub Hammal[1] Pierre Zweigenbaum[1] Caio Corro[2]
(1) Université Paris-Saclay, CNRS, LISN
(2) INSA Rennes, IRISA, CNRS, Université de Rennes
{ayoub.hammal,pz}@lisn.fr. caio.corro@irisa.f

RÉSUMÉ

Des travaux antérieurs ont conclu que la majeure partie des capacités de génération des grands modèles de langue (*large language models*, LLM) est apprise lors du pré-entraînement. Cependant, les LLM nécessitent une étape d'alignement supplémentaire afin de satisfaire aux exigences des tâches cibles ainsi qu'aux préférences stylistiques, entre autres. Avec la croissance du nombre de paramètres des LLM, le coût computationnel des procédures d'alignement devient de plus en plus prohibitif. Dans ce travail, nous proposons une nouvelle approche permettant d'éviter ces coûts grâce à un alignement implicite du LLM lors de la génération. Notre approche se fonde sur l'utilisation d'un petit LLM auxiliaire correctement aligné à coût bien moindre et sur la construction d'un mélange des distributions de sortie des deux LLM. Le calcul des paramètres de la loi de mélange est réduit à un problème de sac à dos binaire. Grâce à ce formalisme, nous dérivons des approximations primales et duales de la loi de mélange optimale. Nous montrons expérimentalement les bénéfices de notre méthode, tant en termes de performance sur les tâches cibles que de vitesse de génération en utilisant un décodage spéculatif.

ABSTRACT

KAD : A Framework for Proxy-based Test-time Alignment with Knapsack Approximation Deferral.

Several previous works concluded that the largest part of the generation capabilities of large language models (LLM) is learned (early) during pre-training. However, LLMs still require further alignment to adhere to downstream task requirements and stylistic preferences, among other desired properties. As LLMs continue to scale in size, the computational cost of alignment procedures becomes prohibitively high. In this work, we propose a novel approach to circumvent these costs via proxy-based test-time alignment, *i.e.* using guidance from a small aligned model. Our approach can be described as a token-specific cascading method, where the token-specific deferral rule is reduced to 0-1 knapsack problem. In this setting, we derive primal and dual approximations of the optimal deferral decision. We experimentally demonstrate the benefits of our method across both task performance and speculative decoding speed.

MOTS-CLÉS : Grands modèles de langue, problème du sac à dos, alignement de modèle de langue.

KEYWORDS: Large language models, knapsack problem, language model alignment.

ARTICLE ACCEPTÉ À : The 19th Conference of the European Chapter of the Association for Computational Linguistics.
URL : https://arxiv.org/pdf/2510.27017

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 2 : articles déjà publiés, pages 52–52.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.