

الجمهورية الجزائرية الديمقراطية الشعبية

People's Democratic Republic of Algeria

وزارة التعليم العالي والبحث العلمي

Ministry of Higher Education and Scientific Research

المدرسة العليا للإعلام الآلي 8 ماي 1945 - سيدي بلعباس

Higher School of Computer Science

8 Mai 1945 - Sidi Bel Abbès



Thesis

To obtain the diploma of Master's Degree

Field of Study: Computer Science

Specialization: Artificial Intelligence & Data Science

Theme

Multi-Objective Alignment of Large Language Models with Preference Vectors: A Survey of Methods and Approaches

Presented by

Dallaa Abdelsamad Taki-Eddine

Defended on: **July 4th, 2026**

In front of the jury composed of:

Mr. Belkacem Khaldi	President of the Jury
Mr. Souleyman Chaib	Supervisor
Mr. Zeyd Boukhers	Supervisor
Mr. Lingxiao Kong	Supervisor
Mrs. Rabab Bousmaha	Examiner

Academic Year: 2025/2026

Abstract

Multi-objective alignment of [Large Language Models \(LLMs\)](#) is the problem of training a single model to faithfully serve diverse, potentially conflicting user objectives, such as helpfulness, safety, verbosity, and correctness without committing to a fixed behavioural trade-off at training time. This report surveys the theoretical foundations and principal methods in this field, covering the post-training pipeline, preference optimisation objectives, parameter-efficient fine-tuning techniques, hypernetworks, and multi-objective optimisation theory. Three alignment approaches are analysed in depth: [Rewards-in-Context \(RiC\)](#), Rewarded Soups, and Panacea, alongside related methods including SteerLM, ArmoRM, and recent hypernetwork-based personalisation systems. A comparative analysis is conducted across dimensions of preference representation, training cost, inference flexibility, Pareto-front coverage, and prompt-awareness. The survey concludes by identifying prompt-agnosticism as a shared limitation of current weight-space methods, and positions context-conditioned hypernetwork adapters as a promising direction for future work.

Keywords: Large Language Models, Multi-Objective Alignment, Hypernetworks, Parameter-Efficient Fine-Tuning, Pareto Optimality, Preference Vectors, Singular Value Decomposition.

Résumé

L’alignement multi-objectif des modèles [LLM](#) est le problème de la formation d’un modèle unique pour servir fidèlement des objectifs d’utilisateur divers et potentiellement conflictuels, tels que l’utilité, la sécurité, la verbosité et l’exactitude, sans s’engager dans un compromis comportemental fixe au moment de la formation. Ce rapport examine les fondements théoriques et les principales méthodes dans ce domaine, couvrant le post-entraînement, les objectifs d’optimisation des préférences, les techniques de [Parameter-Efficient Fine-Tuning \(PEFT\)](#), les hyperréseaux et la théorie de l’optimisation multi-objectifs. Trois approches d’alignement sont analysées en profondeur : [Rewards-in-Context \(RiC\)](#), [Rewarded Soups](#) et [Panacea](#), ainsi que des méthodes connexes, notamment [SteerLM](#), [ArmoRM](#) et les récents systèmes de personnalisation basés sur les hyperréseaux. Une analyse comparative est menée sur les dimensions de la représentation des préférences, du coût de la formation, de la flexibilité de l’inférence, de la couverture Pareto-front et de la prise de conscience des invites. L’enquête se termine en identifiant l’agnosticisme rapide comme une limitation commune des méthodes actuelles d’espace de poids et positionne les adaptateurs d’hyperréseau conditionnés par le contexte comme une direction prometteuse pour les travaux futurs.

Mots-clés: LLMs, alignement multi-objectif, hyperréseaux, PEFT, optimalité de Pareto, vecteurs de préférence, décomposition en valeurs singulières.

ملخص

يمثل الـ Multi-objective Large Language Models (LLMs) مشكلة تدريب نموذج واحد لخدمة أهداف متنوعة ومتناقضة، مثل المساعدة والسلامة والإسهاب والصحة دون الحاجة إلى الإستغناء عن هدف للبلية آخر في وقت التدريب. يستعرض هذا التقرير الأسس النظرية والأساليب الرئيسية في هذا المجال، ويغطي عمليات ما بعد التدريب، وأهداف تحسين التفضيلات، وتقنيات الـ PEFT، والـ hypernetworks، ونظرية الـ Multi-objective alignment. يتم تحليل ثلاثة أساليب للمحاذاة بعمق: Rewards-in-Context (RiC)، و Rewarded، و Panacea و Soups، جنباً إلى جنب مع الأساليب ذات الصلة بما في ذلك، SteerLM و ArmoRM، وأنظمة التخصيص الحديثة القائمة على الشبكة الفائقة. يتم إجراء تحليل مقارنة عبر أبعاد تمثيل التفضيلات، وتكلفة التدريب، ومرونة الـ inference، وتغطية الـ Pareto، و Front، و الـ awareness. Prompt، ويختتم الاستطلاع بتحديد الـ prompt-agnosticism باعتبارها قيداً مشتركاً لطرق مساحة الوزن الحالية، ويضع الـ hypernetwork adapters المشروطة بالسياق كاتجاه واعد للعمل المستقبلي.

الكلمات الرئيسية: Parameter- Hypernetworks, Alignment, Multi-Objective Models, Language Large Decomposition. Value Singular Vectors, Preference Optimality, Pareto Fine-Tuning, Efficient