

Policy Optimization and Language Models

Course title – Intitulé du cours	Policy Optimization and Language Models
Level / Semester – Niveau /semestre	M2, S2 (Spring 2027; exact dates TBD)
School – Composante	Ecole d'Economie de Toulouse
Teacher – Enseignant responsable	Gaurav Mahajan
Other teacher(s) – Autre(s) enseignant(s)	
Other teacher(s) – Autre(s) enseignant(s)	
Other teacher(s) – Autre(s) enseignant(s)	
Other teacher(s) – Autre(s) enseignant(s)	
Other teacher(s) – Autre(s) enseignant(s)	
Lecture Hours – Volume Horaire CM	10h
TA Hours – Volume horaire TD	
TP Hours – Volume horaire TP	
Course Language – Langue du cours	English
TA and/or TP Language – Langue des TD et/ou TP	

Teaching staff contacts – Coordonnées de l'équipe pédagogique :

Course Objectives – Objectifs du cours :

In this course, we study policy optimization, which provides a foundation for post-training language models. We study various policy-optimization algorithms, beginning with classical methods such as value iteration and policy iteration, and then moving to policy-gradient methods such as natural policy gradient, TRPO, PPO, DPO, and GRPO, which are used to post-train language models. We conclude by discussing modern post-training strategies used by recent open-source language models, including DeepSeek-R1 and Kimi K3.

Prerequisites – Pré requis :

Practical information about the sessions – Modalités pratiques de gestion du cours :

Lecture notes and material for reviewing the basics of reinforcement learning will be made available on the course website.

Grading system – Modalités d'évaluation : Final research project

Bibliography/references – Bibliographie/références :

Reinforcement Learning: Theory and Algorithms
https://rltheorybook.github.io/rltheorybook_ABJKS.pdf

Session planning – Planification des séances

- **Lecture 1:** Bellman optimality and fixed points
- **Lecture 2:** Value iteration and policy iteration
- **Lecture 3:** Policy gradient methods
- **Lecture 4:** Natural policy gradient
- **Lecture 5:** Trust-region and proximal policy optimization
- **Lecture 6:** Policy optimization for language models
- **Lecture 7:** Modern post-training strategies

Distance learning – Enseignement à distance :