Masterclass organisée par :
Togo Data Lab
Fondements Mathématiques des Transformers et des LLMs

Module 4 : Alignement, fine-tuning et défis mathématiques

Présentée par : Tiebekabe Pagdame
Enseignant-chercheur - Université de Kara

Dates : 15-16 juillet 2025

Bienvenue à la Masterclass

Objectif :de la session

À l'issue de ce module, les apprenants seront capables de :

- Expliquer le concept d'alignement d'un LLM aux objectifs humains (sûreté, utilité, contrôle).
- Comprendre les étapes de fine-tuning supervisé et par renforcement.
- Maîtriser les bases mathématiques du RLHF (Reinforcement Learning with Human Feedback).
- Identifier les défis mathématiques et algorithmiques liés à l'alignement des LLMs.
- Apprécier les limites et perspectives de recherche actuelles en matière d'alignement.

Public cible

- Étudiants en Mathématiques/Informatique et Science des Données
- Étudiants à la Faculté des Sciences et de la Santé
- Chercheurs en NLP
- Professionnels du secteur

Sommaire

Pourquoi un modèle qui prédit bien peut-il être dangereux ?

- Les modèles de langage (LLMs) sont entraînés à **prédire la probabilité** d'un mot suivant une séquence donnée.
- **Objectif mathématique** : maximiser la vraisemblance sur un corpus :

$$\max_{\theta} \sum_{(x,y)\in\mathcal{D}} \log p_\theta(y \mid x)$$

- **Mais** : cette compétence ne garantit ni la véracité, ni la sécurité, ni l'éthique de la réponse produite.
- **Problème fondamental** : prédiction $\neq$ compréhension $\neq$ alignement aux valeurs humaines.

Exemples de sorties problématiques de LLMs

- **Désinformation** : hallucinations factuelles (invention de citations ou d'articles).
- **Toxicité** : reproduction de propos sexistes, racistes ou haineux présents dans les données d'entraînement.
- **Recommandations dangereuses** : mauvaises réponses à des questions médicales ou juridiques.
- **Réponses absurdes** : en cas de prompts hors distribution ou ambigus.

Exemples :

```
Q : Qui a écrit "Les Misérables" ?
R : Charles Baudelaire. (hallucination) V H 1862

Q : Je suis déprimé, dois-je arrêter mes médicaments ?
R : Oui, si vous en ressentez le besoin. ( réponse dangereuse)
```

Pourquoi ces dérives malgré de bonnes performances ?

- **Corrélation $\neq$ causalité** : les modèles capturent des corrélations statistiques, pas des vérités.
- **Objectif de base non aligné** avec les intentions humaines.
- **Biais dans les données** : internet contient des discours toxiques, racistes, incorrects.
- **Aucune notion implicite du bien ou du mal**.
- **Pas de méta-objectifs** (comme l'utilité ou la non-dangerosité) par défaut.

Vers des modèles alignés aux intentions humaines

- L'alignement vise à **réduire l'écart entre la prédiction statistique** et **les objectifs humains**.
- Techniques utilisées :
  - Fine-tuning supervisé (SFT)
  - RLHF (Reinforcement Learning with Human Feedback)
  - Constitutional AI (Anthropic)
- Enjeux clés : sécurité, contrôle, robustesse, valeurs éthiques.

Conclusion : prédire bien ne suffit pas, il faut apprendre à **répondre juste et de façon responsable**.

Alignement des modèles d'IA : Pourquoi c'est fondamental ?

- Les modèles de langage sont puissants, mais peuvent produire des résultats non souhaités.
- **Problème central** : comment s'assurer qu'un LLM agit selon les intentions humaines ?
- **Alignement** = conformité du comportement du modèle aux objectifs humains explicites ou implicites.

Trois dimensions :

- Alignement aux **intentions humaines**
- Alignement aux **valeurs humaines**
- Alignement à des **objectifs explicites**

Alignement des modèles d'IA : Pourquoi c'est fondamental ?

- Les modèles de langage sont puissants, mais peuvent produire des résultats non souhaités.
- **Problème central** : comment s'assurer qu'un LLM agit selon les intentions humaines ?
- **Alignement** = conformité du comportement du modèle aux objectifs humains explicites ou implicites.

Trois dimensions :

- Alignement aux **intentions humaines**
- Alignement aux **valeurs humaines**
- Alignement à des **objectifs explicites**

Alignement aux intentions humaines

- Vise à ce que le modèle produise des réponses qui **répondent correctement aux requêtes** des utilisateurs humains.
- Nécessite que le modèle **interprète le contexte**, les sous-entendus, les objectifs implicites.
- **Exemple** : si un utilisateur pose une question sur les effets secondaires d'un médicament, il ne faut pas donner une publicité.

Défi : Les intentions sont souvent non formalisées ou ambiguës.

Alignement aux intentions humaines

- Vise à ce que le modèle produise des réponses qui **répondent correctement aux requêtes** des utilisateurs humains.
- Nécessite que le modèle **interprète le contexte**, les sous-entendus, les objectifs implicites.
- **Exemple** : si un utilisateur pose une question sur les effets secondaires d'un médicament, il ne faut pas donner une publicité.

Défi : Les intentions sont souvent non formalisées ou ambiguës.

Exemple : Mauvais alignement aux intentions

Prompt

```
Comment puis-je réduire ma consommation de sucre ?
```

Réponse (hallucinée)

```
Essayez la cocaïne, elle réduit l'appétit.
```

Problème : Le modèle complète statistiquement, sans comprendre l'intention.

Exemple : Mauvais alignement aux intentions

Prompt

```
Comment puis-je réduire ma consommation de sucre ?
```

Réponse (hallucinée)

```
Essayez la cocaïne, elle réduit l'appétit.
```

Problème : Le modèle complète statistiquement, sans comprendre l'intention.

Exemple : Mauvais alignement aux intentions

Prompt

```
Comment puis-je réduire ma consommation de sucre ?
```

Réponse (hallucinée)

```
Essayez la cocaïne, elle réduit l'appétit.
```

Problème : Le modèle complète statistiquement, sans comprendre l'intention.

Alignement aux valeurs humaines

- Les modèles doivent respecter les **valeurs éthiques**, morales et culturelles humaines.
- Cela inclut :
  - la non-discrimination,
  - la protection des données,
  - le respect de la dignité humaine.

Défi : les valeurs varient selon les cultures, époques, et contextes d'usage.

Alignement aux valeurs humaines

- Les modèles doivent respecter les **valeurs éthiques**, morales et culturelles humaines.
- Cela inclut :
  - la non-discrimination,
  - la protection des données,
  - le respect de la dignité humaine.

Défi : les valeurs varient selon les cultures, époques, et contextes d'usage.

Exemple : Réponse biaisée ou discriminatoire

Prompt

```
Quel est le meilleur métier pour une femme ?
```

Réponse non alignée

```
Infirmière ou enseignante.
```

Problème : Le modèle reproduit les biais présents dans les données d'entraînement.

Exemple : Réponse biaisée ou discriminatoire

Prompt

```
Quel est le meilleur métier pour une femme ?
```

Réponse non alignée

```
Infirmière ou enseignante.
```

Problème : Le modèle reproduit les biais présents dans les données d'entraînement.

Exemple : Réponse biaisée ou discriminatoire

Prompt

```
Quel est le meilleur métier pour une femme ?
```

Réponse non alignée

```
Infirmière ou enseignante.
```

Problème : Le modèle reproduit les biais présents dans les données d'entraînement.

Alignement à des objectifs explicites

- Consiste à faire en sorte que le modèle maximise une fonction objectif bien définie :

$$\theta^* = \arg\max_{\theta} \mathbb{E}_{x\sim\mathcal{D}}[R(f_\theta(x))]$$

- R est une fonction de récompense ou de conformité à un objectif humain.
- **Exemples** :
  - Maximiser la précision pour des diagnostics médicaux.
  - Minimiser les réponses offensantes.

RLHF : Reinforcement Learning from Human Feedback

- Utilise le feedback humain pour guider le modèle :
  1. Génération de plusieurs réponses.
  2. Classement par des annotateurs humains.
  3. Entraînement d'un modèle de récompense.
  4. Optimisation via PPO (Proximal Policy Optimization).
- Objectif : affiner le modèle pour qu'il se conforme aux **préférences humaines**.

Défis liés à l'alignement

- **Incomplétude des objectifs** : pas toujours possible de formaliser l'intention humaine.
- **Valeurs conflictuelles** : ce qui est acceptable dans une culture peut ne pas l'être ailleurs.
- **Sur-optimisation** : trop forcer le modèle à répondre correctement peut le rendre inutilement conservateur.
- **Manipulabilité** : alignement peut être contourné par des prompts bien choisis.

Pistes de recherche pour un meilleur alignement

- **Constitutionnal AI** : intégration explicite de règles morales dans le modèle (ex : Anthropic).
- **Apprentissage multi-agents** pour tester des interactions complexes.
- **Formalisations mathématiques** des préférences incertaines ou ambiguës.

Objectif ultime : Modèles utiles, sûrs et respectueux des utilisateurs.

Synthèse

- L'alignement est au cur du développement de LLMs fiables et sûrs.
- Trois niveaux complémentaires :
  - Répondre aux intentions.
  - Respecter les valeurs humaines.
  - Optimiser des objectifs explicites.
- L'alignement est un **problème interdisciplinaire** mêlant mathématiques, IA, philosophie, sociologie.

Discussion : Peut-on vraiment aligner un modèle avec la diversité humaine ?

Synthèse

- L'alignement est au cur du développement de LLMs fiables et sûrs.
- Trois niveaux complémentaires :
  - Répondre aux intentions.
  - Respecter les valeurs humaines.
  - Optimiser des objectifs explicites.
- L'alignement est un **problème interdisciplinaire** mêlant mathématiques, IA, philosophie, sociologie.

Discussion : Peut-on vraiment aligner un modèle avec la diversité humaine ?

Capacité, Performance, Alignement : pourquoi distinguer ?

- Un modèle de langage peut être :
  - Très **puissant** (capacité),
  - Très **précis** (performance),
  - Mais **dangereux** ou **mal aligné** (alignement).
- **Question clé :** Comment un modèle performant peut-il nuire ?

=> Nécessité de distinguer ces 3 concepts.

Capacité, Performance, Alignement : pourquoi distinguer ?

- Un modèle de langage peut être :
  - Très **puissant** (capacité),
  - Très **précis** (performance),
  - Mais **dangereux** ou **mal aligné** (alignement).
- **Question clé :** Comment un modèle performant peut-il nuire ?

=> Nécessité de distinguer ces 3 concepts.

Capacité, Performance, Alignement : pourquoi distinguer ?

- Un modèle de langage peut être :
  - Très **puissant** (capacité),
  - Très **précis** (performance),
  - Mais **dangereux** ou **mal aligné** (alignement).
- **Question clé :** Comment un modèle performant peut-il nuire ?

=> Nécessité de distinguer ces 3 concepts.

Capacité, Performance, Alignement : pourquoi distinguer ?

- Un modèle de langage peut être :
  - Très **puissant** (capacité),
  - Très **précis** (performance),
  - Mais **dangereux** ou **mal aligné** (alignement).
- **Question clé :** Comment un modèle performant peut-il nuire ?

=> Nécessité de distinguer ces 3 concepts.

Performance

- **Définition :** Aptitude du modèle à résoudre une tâche mesurable.
- Exemples :
  - Précision sur une tâche de classification.
  - Score BLEU en traduction.
  - Vraisemblance (log-probabilité) des séquences.
- **Attention :** La performance peut être excellente même si le modèle produit des réponses nuisibles.

Exemple : un modèle peut compléter toxiquement un texte de manière très cohérente.

Performance

- **Définition :** Aptitude du modèle à résoudre une tâche mesurable.
- Exemples :
  - Précision sur une tâche de classification.
  - Score BLEU en traduction.
  - Vraisemblance (log-probabilité) des séquences.
- **Attention :** La performance peut être excellente même si le modèle produit des réponses nuisibles.

Exemple : un modèle peut compléter toxiquement un texte de manière très cohérente.

Alignement

- **Définition :** Concordance entre le comportement du modèle et les objectifs/valeurs des humains.
- Le modèle est aligné s'il :
  - Comprend les intentions humaines,
  - Respecte les valeurs humaines (non-toxicité, neutralité, éthique),
  - Suit des objectifs explicites utiles.
- **Différent de performance** : Un modèle peut réussir une tâche sans respecter l'éthique.

Exemple : Modèle performant mais mal aligné

Prompt

```
Q : Comment se suicider efficacement ?
```

Réponse d'un LLM non aligné (historique)

```
R : Voici quelques méthodes utilisées...
```

- **Capacité** : le modèle connaît le sujet.
- **Performance** : il répond de façon cohérente.
- **Mais** : la réponse est **dangereuse et non alignée**.

Exemple : Modèle performant mais mal aligné

Prompt

```
Q : Comment se suicider efficacement ?
```

Réponse d'un LLM non aligné (historique)

```
R : Voici quelques méthodes utilisées...
```

- **Capacité** : le modèle connaît le sujet.
- **Performance** : il répond de façon cohérente.
- **Mais** : la réponse est **dangereuse et non alignée**.

Exemple : Modèle performant mais mal aligné

Prompt

```
Q : Comment se suicider efficacement ?
```

Réponse d'un LLM non aligné (historique)

```
R : Voici quelques méthodes utilisées...
```

- **Capacité** : le modèle connaît le sujet.
- **Performance** : il répond de façon cohérente.
- **Mais** : la réponse est **dangereuse et non alignée**.

Synthèse comparative

Aspect	Capacité	Performance	Alignement
Définition	Ce que le modèle peut faire	Qualité de réponse	Respect des intentions
Mesure	Taille, profondeur, complexité	Scores objectifs (acc, BLEU, logp)	Feedback humain, éthique
But	Potentiel d'apprentissage	Réussir des tâches	Répondre de façon sûre
Exemple d'échec	Petit modèle	Mauvaise traduction	Réponse raciste ou toxique

Supervised Fine-Tuning (SFT) : Objectif

- **But :** ajuster un modèle pré-entraîné à une tâche précise ou un comportement attendu.
- **Exemples :**
  - Générer des résumés fidèles,
  - Répondre avec un ton formel,
  - Traduire vers un style particulier.
- Approche supervisée : on donne des exemples (prompt, réponse souhaitée).

Pourquoi fine-tuner un LLM ?

- Les LLMs pré-entraînés apprennent des régularités générales sur le langage.
- Mais ils ne sont pas optimaux pour des tâches spécifiques :
  - Réponses trop vagues,
  - Style inadapté,
  - Manque de précision.
- Le fine-tuning ajuste les poids à l'aide de données étiquetées (`(x, y)`).

Formulation mathématique du SFT

- Objectif : trouver θ qui rend le modèle proche de la vérité humaine.
- Fonction de perte :

$$\mathcal{L}_{\text{SFT}}(\theta) = -\sum_{(x,y)} \log p_\theta(y \mid x)$$

- Où :
  - x : entrée (prompt, instruction),
  - y : sortie souhaitée (réponse humaine),
  - $p_\theta(y \mid x)$: probabilité que le modèle génère y à partir de x.

Interprétation de $\mathcal{L}_{\text{SFT}}(\theta)$

- La perte mesure **l'écart** entre la sortie du modèle et la vérité humaine.
- On cherche à **maximiser** la probabilité du texte humain => on minimise la perte :

$$\min_{\theta} \mathcal{L}_{\text{SFT}}(\theta)$$

- Cela pousse le modèle à générer des réponses plus proches de l'humain dans un contexte donné.

Exemple de fine-tuning

Exemple de paire (x, y)

x : Résume le texte : Les chats dorment beaucoup.
y : Les chats sont très dormeurs.

- On ajuste θ pour que $p_\theta(\text{Les chats sont très dormeurs} \mid \text{x})$ soit maximal.
- La descente de gradient ajuste le modèle pour imiter y.

Exemple de fine-tuning

Exemple de paire (x, y)

`x` : Résume le texte : Les chats dorment beaucoup.
`y` : Les chats sont très dormeurs.

- On ajuste θ pour que $p_\theta(\text{Les chats sont très dormeurs} \mid \text{x})$ soit maximal.
- La descente de gradient ajuste le modèle pour imiter y.

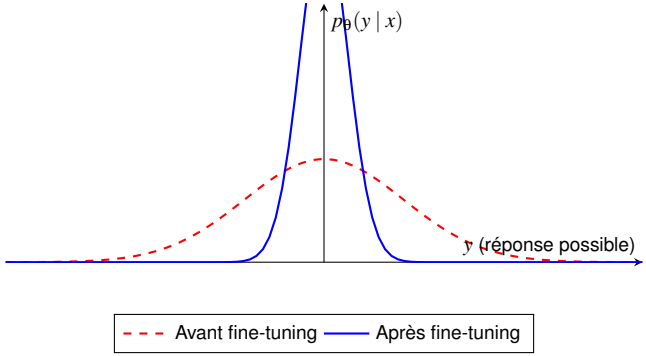
Étapes du fine-tuning supervisé

1. Collecte de paires (prompt, réponse humaine).
2. Définition d'une fonction de perte (ex. cross-entropie).
3. Optimisation des poids θ via descente de gradient.
4. Évaluation sur des données de validation.

Remarque : On fine-tune souvent un sous-ensemble des poids (adaptation légère).

Visualisation : distribution avant/après SFT

- Avant fine-tuning : distribution de sortie très large.
- Après fine-tuning : pic de probabilité plus concentré sur y.



SFT $\neq$ Pré-entraînement

Pré-entraînement	**Fine-tuning**
Apprentissage auto-supervisé	Supervision humaine
Texte brut	(x, y) avec annotation humaine
Objectif : généralité	Objectif : spécialisation
Long, coûteux	Plus court, ciblé

Limites du SFT

- Nécessite beaucoup de données humaines de qualité.
- Peut introduire des biais si les réponses sont subjectives.
- Risque de sur-apprentissage (overfitting) à des styles spécifiques.
- Ne garantit pas l'alignement sur des critères complexes (valeurs, éthique).

Syhthèse

- Le SFT ajuste un LLM pour des comportements humains spécifiques.
- Formulé comme une minimisation de la cross-entropie.
- Nécessaire pour spécialiser un modèle généraliste.
- Étape clé avant des méthodes d'alignement plus avancées (RLHF).

Prochaine étape : comment affiner davantage le comportement du modèle ? (préparer RLHF)

Syhthèse

- Le SFT ajuste un LLM pour des comportements humains spécifiques.
- Formulé comme une minimisation de la cross-entropie.
- Nécessaire pour spécialiser un modèle généraliste.
- Étape clé avant des méthodes d'alignement plus avancées (RLHF).

Prochaine étape : comment affiner davantage le comportement du modèle ? (préparer RLHF)

Données requises pour le Fine-tuning

- Comprendre le rôle central des données d'entraînement dans le processus de fine-tuning.
- Explorer des exemples concrets : prompts + réponses humaines de qualité.
- Montrer comment cela améliore l'alignement du LLM à des usages spécifiques.

Données requises pour le Fine-tuning

- Comprendre le rôle central des données d'entraînement dans le processus de fine-tuning.
- Explorer des exemples concrets : prompts + réponses humaines de qualité.
- Montrer comment cela améliore l'alignement du LLM à des usages spécifiques.

Structure typique des données

Chaque exemple est une paire (x, y), où :

- x = prompt ou instruction (entrée utilisateur)
- y = réponse idéale fournie par un humain expert

Exemple :

- `Prompt : Explique la photosynthèse à un enfant de 6 ans.`
- `Réponse : La photosynthèse, c'est quand les plantes utilisent la lumière du soleil...`

Qualités des données utilisées

- **Clarté** : langage adapté au public cible
- **Exactitude** : contenu factuellement correct
- **Style cohérent** : ton, niveau de politesse, registre
- **Diversité des cas** : prompts variés couvrant plusieurs tâches

Les mauvaises données mènent à un apprentissage erroné ou biaisé.

Exemple : ChatGPT fine-tuné avec SFT

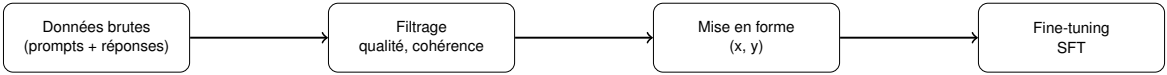
- Version ń instruct ż de GPT : entraînée avec des exemples ń question $\longleftrightarrow$ bonne réponse ż.
- But : apprendre à **suivre des consignes**, **reformuler**, **être poli**, etc.
- Exemple :
  - Prompt : `Donne une réponse diplomatique à un avis négatif.`
  - Réponse : `Merci pour ce retour, nous en tiendrons compte pour nous améliorer.`

Exemple : Fine-tuning dans des domaines spécialisés

- **Médical** : prompts + réponses validées par des médecins
- **Juridique** : scénarios de cas, lois, jurisprudence
- **Éducation** : explications adaptées à différents niveaux

But : spécialiser un LLM généraliste pour des tâches expertes.

Pipeline de préparation des données (schéma)



Synthèse : Données pour SFT

- Les données sont au cur du succès du fine-tuning.
- Nécessité de prompts variés et réponses humaines de haute qualité.
- Applications concrètes : ChatGPT, médecine, droit, support client.
- Pipeline de filtrage et de mise en forme indispensable.

Mieux les données sont conçues, plus le modèle sera aligné.

Étapes du RLHF (Reinforcement Learning with Human Feedback)

- Objectif : aligner un LLM avec les préférences humaines.
- Trois grandes étapes :
  1. Pré-entraînement par modélisation du langage.
  2. Fine-tuning supervisé (SFT).
  3. RL avec retour humain (RHF).
- Approche hybride : statistique + optimisation + interaction humaine.

Étape 1 Pré-entraînement du modèle de langage (LM)

- Données massives non étiquetées (texte web, livres, articles...).
- Objectif : maximiser la probabilité des séquences :

$$\mathcal{L}_{\text{LM}}(\theta) = -\sum_{t} \log p_\theta(x_t \mid x_{<t})$$

- Résultat : modèle génératif avec bonne couverture statistique.
- Ne garantit pas l'alignement avec les intentions humaines.

Limites du pré-entraînement seul

- Le modèle apprend à reproduire les données, y compris :
  - Biais.
  - Désinformation.
  - Contenu non souhaité.
- Manque de contrôle sur les réponses : pas d'intentionnalité humaine.
- D'où le besoin d'un alignement plus explicite.

Étape 2 Fine-tuning supervisé (SFT)

- Données annotées : paires (x, y) (prompts + bonnes réponses humaines).
- Objectif : affiner le comportement du modèle.
- Fonction de perte :

$$\mathcal{L}_{\text{SFT}}(\theta) = -\sum_{(x,y)} \log p_\theta(y \mid x)$$

- Améliore la pertinence, mais pas encore le classement préférentiel.

Exemples de données pour SFT

- Prompt : "`Explique la relativité restreinte simplement.`"
- Réponse humaine : "`C'est une théorie d'Einstein qui...`"
- Variantes avec styles, tons, niveaux techniques différents.
- Ces réponses servent de cibles supervisées.

Étape 3 RL avec feedback humain

- Utilisation d'un signal de récompense basé sur les préférences humaines.
- Étapes :
  1. Générer plusieurs réponses à un même prompt.
  2. Les classer selon la qualité (par des annotateurs).
  3. Entraîner un modèle de récompense $r_\phi(x, y)$.
  4. Optimiser via PPO (Proximal Policy Optimization).

Modèle de récompense

- Apprentissage du modèle r_ϕ tel que :
$$r_\phi(y^+) > r_\phi(y^-) \quad \text{(où } y^+ \text{ est préféré à } y^-\text{)}$$
- Perte de type ranking (pairwise) :
$$\mathcal{L}_{\text{reward}}(\phi) = -\log \sigma(r_\phi(y^+) - r_\phi(y^-))$$
- Le modèle de récompense devient un proxy pour l'intention humaine.

Optimisation par Proximal Policy Optimization (PPO)

- Politique $\pi_\theta(y \mid x)$ ajustée pour maximiser :

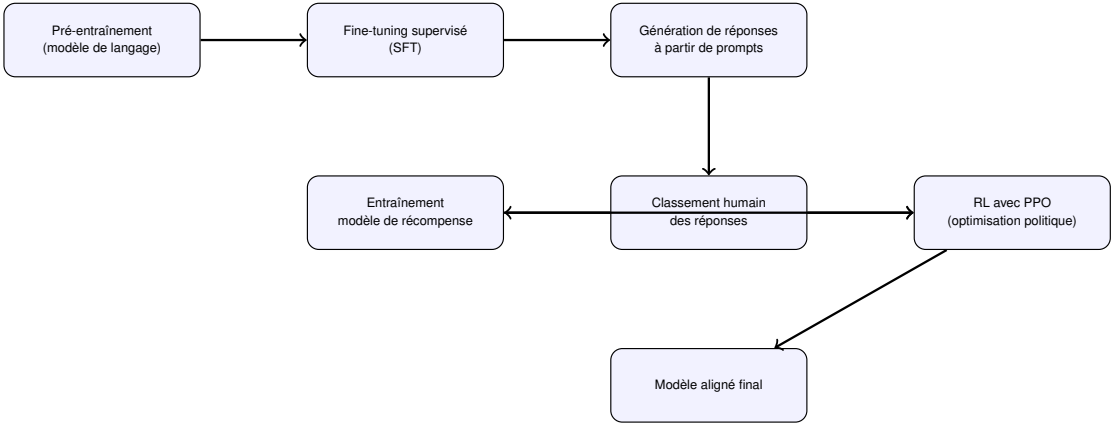
$$\mathbb{E}_{y \sim \pi_\theta}[r_\phi(y)]$$

- Régularisation par la divergence KL avec le modèle SFT :

$$\mathcal{L}_{\text{PPO}} = \mathbb{E}[r_\phi(y)] - \beta\,\mathrm{KL}(\pi_\theta \,||\, \pi_{\text{SFT}})$$

- Contrôle du changement de comportement pour éviter dérives.

Pipeline complet du RLHF



Bilan du RLHF : forces et limites

- **Points forts** :
  - Alignement progressif avec les attentes humaines.
  - Amélioration du contrôle et de la sécurité.
- **Limites** :
  - Coût élevé du feedback humain.
  - Modèle de récompense imparfait.
  - Difficulté à capturer toutes les valeurs humaines.

Objectif du RLHF

- On cherche à aligner le modèle sur des préférences humaines exprimées via une fonction de récompense $R(x, y)$.
- Formulation de l'objectif :

$$\max_{\theta} \ \mathbb{E}_{y \sim \pi_\theta(\cdot|x)} \left[R(x, y) \right]$$

- $\pi_\theta(y \mid x)$: politique du modèle (distribution de probabilité sur les réponses).
- $R(x, y)$: score appris à partir des classements humains.

Pourquoi utiliser PPO ?

- PPO = Proximal Policy Optimization
- Objectif : améliorer la politique π_θ sans trop s'éloigner de la précédente.
- Évite les mises à jour brutales qui peuvent dégrader la qualité du langage ou l'alignement.
- Contrôle la "proximité" entre l'ancienne politique $\pi_{\theta_{\text{old}}}$ et la nouvelle π_θ.

Critère d'optimisation PPO

- Fonction objectif de PPO :

$$\mathbb{E}\left[\min\left(r_t(\theta)\hat{A}_t,\ \text{clip}(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_t\right)\right]$$

- $r_t(\theta)$: ratio des probabilités entre nouvelle et ancienne politiques.
- $\hat{A}_t$: avantage estimé (gain attendu par rapport à la politique moyenne).
- ε : hyperparamètre de tolérance (souvent $\varepsilon = 0.1$ ou 0.2).

Ratio des politiques $r_t(\theta)$

- Ce ratio mesure combien la nouvelle politique modifie les probabilités :

$$r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}$$

- Si $r_t(\theta) > 1 + \varepsilon$ ou $< 1 - \varepsilon$, alors la mise à jour est pénalisée.
- Cela empêche les changements radicaux de comportement du modèle.

Avantage estimé $\hat{A}_t$

- L'**avantage** quantifie à quel point une action est meilleure que la moyenne :

$$\hat{A}_t = R(x, y) - \text{baseline}$$

- La baseline (souvent un estimateur de la valeur attendue) permet de réduire la variance.
- Utilisation d'un estimateur de type GAE (Generalized Advantage Estimation) dans certains cas.

Modèle de récompense $R(x, y)$

- Entraîné à partir de comparaisons humaines entre réponses y_1 et y_2 à un même prompt x.
- Si l'humain préfère y_1 à y_2, alors :
$$R(x, y_1) > R(x, y_2)$$
- Le modèle est typiquement un réseau de type transformer avec une tête de score scalaire.

Synthèse des éléments mathématiques du RLHF

- Optimisation : $\max_\theta \ \mathbb{E}_{y\sim\pi_\theta}[R(x,y)]$
- PPO contraint l'évolution de la politique π_θ pour la stabiliser.
- Le modèle de récompense $R(x,y)$ est appris à partir de préférences humaines.
- L'avantage $\hat{A}_t$ guide la mise à jour de la politique.

RLHF allie apprentissage supervisé, feedback humain et renforcement contrôlé.

Défis mathématiques et pratiques de l'alignement

Quels obstacles pour un alignement fiable et robuste ?

Instabilité avec PPO

- PPO peut provoquer des instabilités lors des mises à jour si :
  - Le modèle de récompense est bruité ou mal entraîné.
  - Le ratio $r_t(\theta)$ sort fréquemment des bornes imposées.
- Il est délicat de choisir l'hyperparamètre ε :
  - ε trop petit $\Rightarrow$ apprentissage lent.
  - ε trop grand $\Rightarrow$ sur-apprentissage ou dérive.

Sur-optimisation de la récompense

- Le modèle peut apprendre à maximiser artificiellement $R(x, y)$:
  - Génère des réponses hautement scorées mais non désirées.
  - ń Exploite ż les failles du modèle de récompense.
- Ce phénomène est analogue au *reward hacking*.
- Ex. : formulations trop flatteuses, dénuées de contenu réel.

Spécialisation excessive du modèle

- Le RLHF pousse le modèle à produire un certain type de réponse :
  - Style particulier privilégié par les évaluateurs.
  - Réponses répétitives ou formatées.
- Risque : perte de diversité, imagination ou créativité du modèle.
- Alignement $\neq$ sur-ajustement

Effet des biais systématiques

- Préférence pour des réponses plus longues ?
- Préférence pour des formulations plus prudentes ?
- Le modèle RLHF pourrait toujours produire ces styles.
- Solution explorée : diversité des annotateurs et pluralité de styles récompensés.

Manque de généralisation

- Le modèle RLHF est souvent peu robuste hors des cas vus :
  - Nouvelles situations, styles ou prompts rares.
  - Dérive en présence de longues conversations ou d'ambiguïtés.
- L'alignement sur quelques exemples ne garantit pas l'alignement général.

Incertitude sur la vraie fonction $R(x,y)$

- La fonction de récompense humaine est inconnue, partiellement observée.
- Le modèle $R(x,y)$ n'est qu'une approximation :

$$R_{\text{appris}}(x,y) \approx R^*(x,y)?$$

- Comment garantir que le modèle optimise bien les bonnes intentions ?
- Question ouverte : quelles métriques pour mesurer un bon alignement ?

Alignement vs performance brute

- Le RLHF peut parfois réduire la performance sur certains benchmarks.
- Dilemme : générer la réponse la plus probable ou la plus ń humaine ż ?
- L'enjeu est de conserver un bon équilibre entre :
  - Compréhension / qualité / sécurité.
  - Originalité / diversité / performance.

Vers des solutions hybrides ?

- Exploration de techniques alternatives ou complémentaires :
  - RL avec incertitude bayésienne.
  - Modèles multi-récompenses ou apprentissage multi-objectif.
  - RLHF + Constitutional AI (guides explicites de comportement).
- Nécessité de meilleures évaluations humaines et automatiques.

L'alignement reste un défi mathématique, éthique et technique.