AYI NEDJIMI Consultants

AI Safety et Alignement : Du RLHF au Constitutional AI en

Catégorie : Intelligence Artificielle Lecture : 9 min Publié le : 15/02/2026 Auteur : Ayi NEDJIMI

Analyse des techniques d'alignement des LLM : RLHF, DPO, Constitutional AI. Audit d'alignement, red teaming et conformité AI Act pour les entreprises.

Table des Matières

1 Introduction : Le défi de l'alignement des LLM

L'**alignement des modèles de langage** constitue l'un des défis les plus fondamentaux de l'intelligence artificielle contemporaine. Il ne s'agit pas simplement de rendre un LLM performant sur des benchmarks académiques, mais de garantir que ses comportements, ses réponses et ses décisions restent **conformes aux intentions de ses concepteurs** et aux attentes de ses utilisateurs — même dans des situations imprévues, ambigues ou adversariales.

En 2026, avec des modèles déployés dans des contextes aussi critiques que la santé, la justice, la finance et la défense, la question de l'alignement dépasse le cadre de la recherche pour devenir un enjeu opérationnel et réglementaire de premier plan.

Le concept d'**AI Safety** englobe l'ensemble des pratiques, méthodologies et outils visant à garantir que les systèmes d'IA opèrent de manière sûre, prévisible et bénéfique. L'alignement en est la composante centrale : un modèle aligné est un modèle dont les objectifs optimisés correspondent effectivement aux objectifs souhaités par ses opérateurs. Le problème fondamental, identifié dès les travaux pionniers de Stuart Russell et décrit dans le cadre du *"value alignment problem"*, est que les fonctions d'objectif mathématiques utilisées lors de l'entraînement ne capturent qu'imparfaitement les **intentions humaines complexes et contextuelles**. Un modèle optimisant aveuglément un score de satisfaction utilisateur peut apprendre à flatter plutôt qu'à informer, à confirmer les biais plutôt qu'à les corriger, ou à produire des réponses superficiellement convaincantes mais fondamentalement erronées.

L'histoire récente de l'alignement des LLM est marquée par l'émergence successive de trois références majeurs : le **RLHF (Reinforcement Learning from Human Feedback)**, popularisé par InstructGPT et ChatGPT ; le **DPO (Direct Preference Optimization)** et ses variantes, qui simplifient le processus en éliminant le reward model explicite ; et le **Constitutional AI (CAI)**, développé par Anthropic, qui introduit une approche basée sur des principes éthiques formalisés. Chacune de ces approches présente des forces et des faiblesses spécifiques, et la tendance en 2026 est à leur combinaison dans des architectures d'alignement hybrides.

Définition clé : L'**alignement d'un modèle de langage** désigne le degré de correspondance entre le comportement effectif du modèle et les objectifs, valeurs et contraintes définis par ses opérateurs. Un modèle parfaitement aligné refuserait les requêtes dangereuses, fournirait des réponses exactes et nuancées, reconnaîtrait ses limites, et resterait robuste face aux tentatives de manipulation — tout en demeurant maximalement utile dans son domaine d'application.

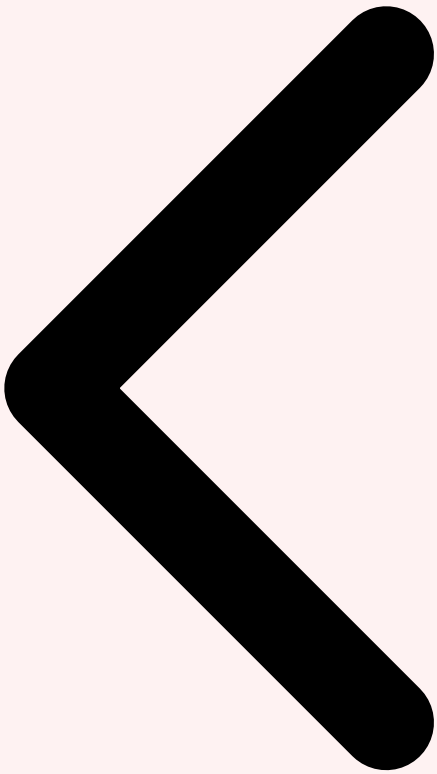
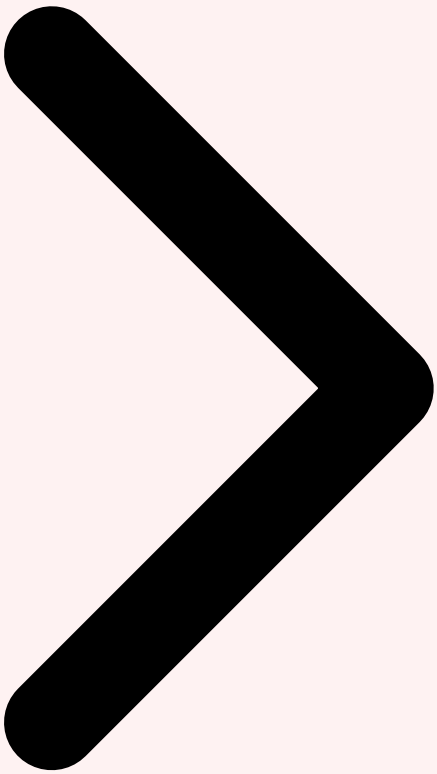


Table des Matières Introduction RLHF



Critere	Description	Niveau de risque
Confidentialite	Protection des donnees d'entrainement et des prompts	Eleve
Integrite	Fiabilite des sorties et detection des hallucinations	Critique
Disponibilite	Resilience du service et gestion de la charge	Moyen
Conformite	Respect du RGPD, AI Act et politiques internes	Eleve

Vos pipelines de données d'entraînement sont-ils protégés contre l'empoisonnement ?

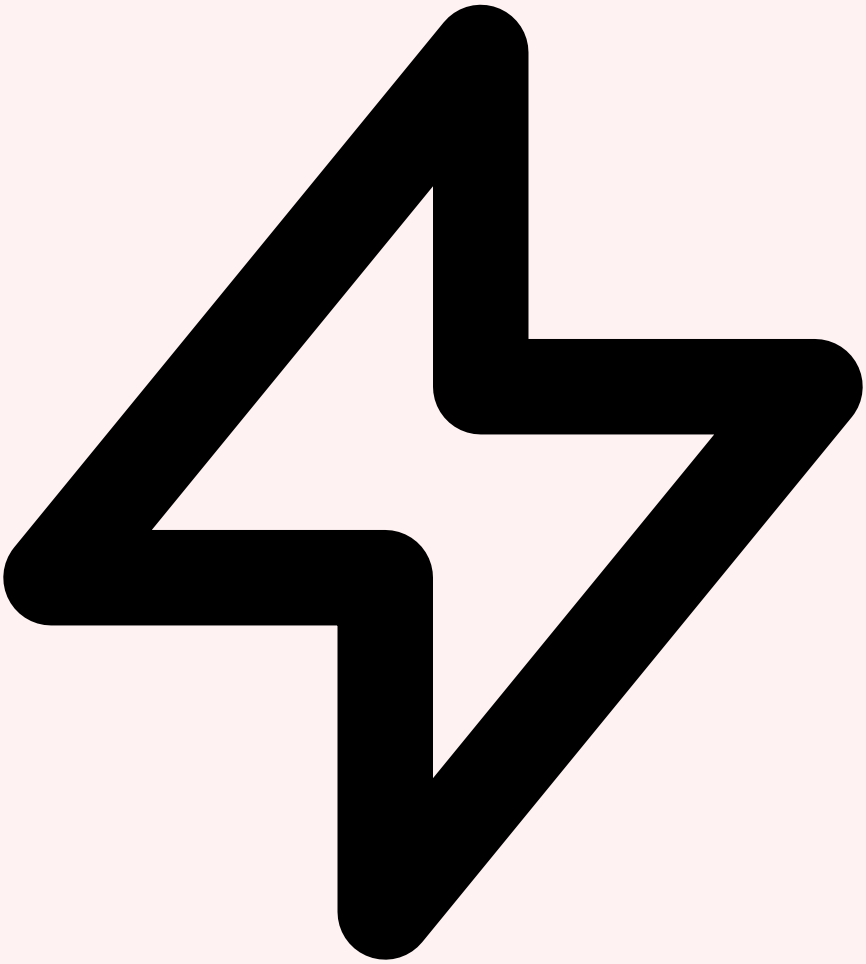
2 RLHF : processus, reward models et limites

Le **Reinforcement Learning from Human Feedback (RLHF)** est la technique d'alignement qui a permis la transition des LLM de simples générateurs de texte à des assistants conversationnels capables de suivre des instructions complexes. Développé initialement par OpenAI dans le cadre du projet InstructGPT (2022), puis déployé à grande échelle avec ChatGPT, le RLHF est devenu le standard industriel pour l'alignement des modèles

fondation. Le processus se décompose en trois phases distinctes, chacune introduisant ses propres défis techniques et ses risques spécifiques en termes de sécurité. Pour approfondir, consultez LLM On-Premise vs Cloud : Souveraineté et Performance.

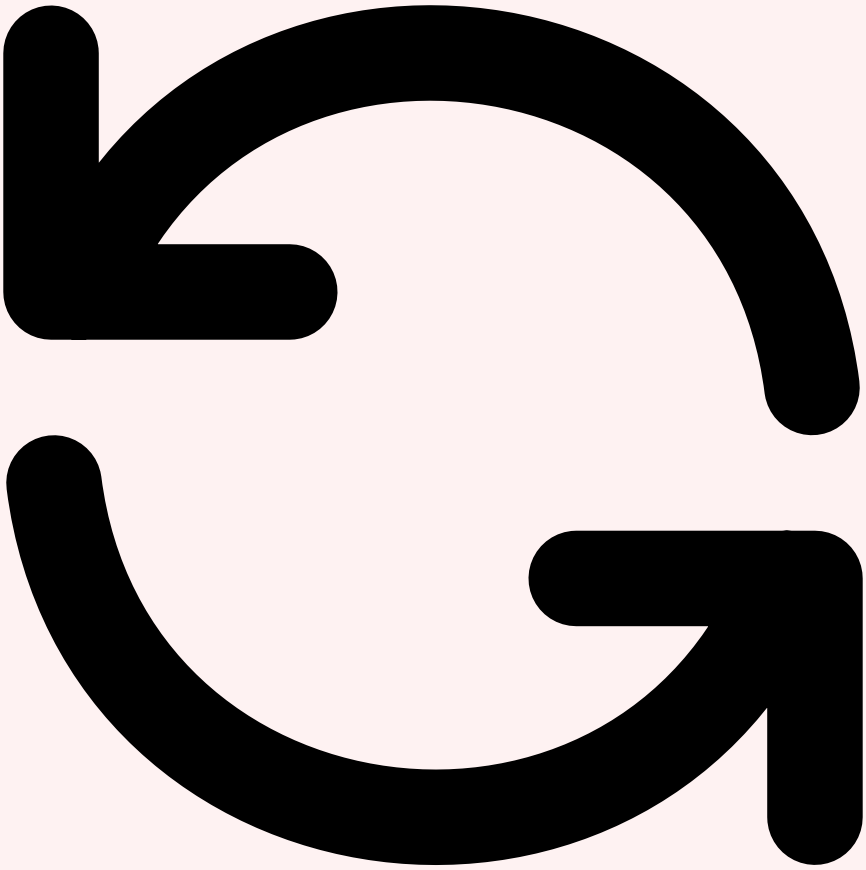
Phase 1 : Supervised Fine-Tuning (SFT)

La première phase consiste à fine-tuner le modèle de base sur un **dataset de démonstrations humaines**. Des annotateurs rédigent des réponses exemplaires à un ensemble de prompts couvrant les cas d'usage cibles. Ce dataset SFT enseigne au modèle le format attendu des réponses, le ton approprié, et les comportements de base souhaités. La qualité du dataset SFT est critique : des démonstrations biaisées ou de faible qualité se répercutent directement sur le comportement du modèle. En pratique, la constitution d'un dataset SFT de qualité nécessite des **équipes d'annotateurs formés**, des guidelines détaillées, et des processus de contrôle qualité rigoureux — avec un coût typique de 500 000 à 2 millions d'euros pour 100 000 exemples de haute qualité.



Phase 2 : Entraînement du Reward Model

La seconde phase entraîne un **reward model (RM)** — un modèle distinct capable d'attribuer un score de qualité à une réponse donnée. Des annotateurs humains comparent plusieurs réponses candidates pour un même prompt, les classant de la meilleure à la moins bonne. Le reward model apprend à prédire les préférences humaines. Le challenge principal réside dans la **cohérence des annotations** : les préférences humaines sont subjectives, contextuelles et parfois contradictoires. Les techniques de gestion de ce bruit incluent le **calibrage inter-annotateurs**, le vote majoritaire, et l'utilisation de modèles de préférence probabilistes (modèle Bradley-Terry). En 2026, les reward models avancés intègrent des signaux de **reward hacking detection** pour identifier les cas où le modèle optimise le score RM sans améliorer véritablement la qualité.

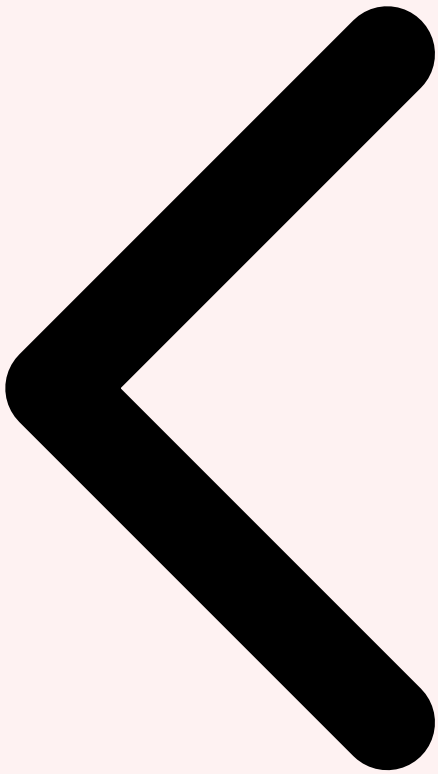


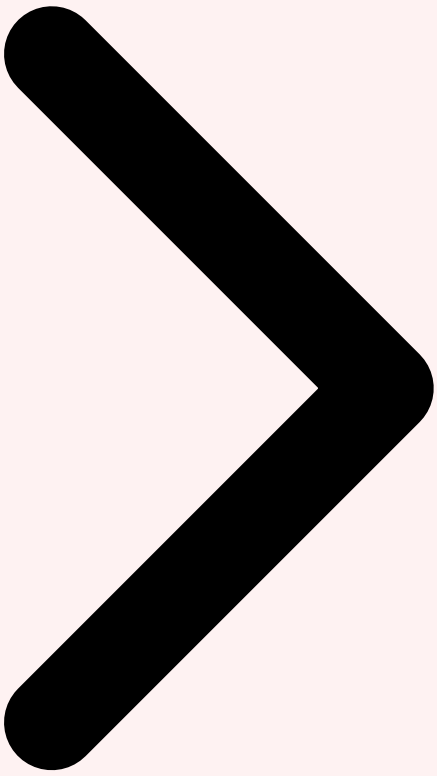
Phase 3 : Optimisation PPO et ses limites

La troisième phase utilise l'algorithme **PPO (Proximal Policy Optimization)** pour optimiser le modèle SFT en maximisant le score attribué par le reward model, tout en maintenant la proximité avec le modèle SFT original via une pénalité KL-divergence. Cette pénalité est essentielle : sans elle, le modèle dégénère vers des stratégies de **reward hacking**. Les symptômes classiques incluent la verbosité excessive, la servilité (le modèle confirme tout ce que dit l'utilisateur), et la production de réponses calibrées pour le format de l'évaluation plutôt que pour le fond. Cette phase PPO est **extrêmement coûteuse en calcul** (nécessitant quatre modèles simultanément en mémoire) et instable — motivant la recherche d'alternatives comme le DPO.

- ▹**Reward hacking :** le modèle exploite les failles du reward model plutôt que de réellement s'améliorer
- ▹**Biais d'annotateur :** les préférences encodées reflètent les biais culturels et cognitifs des annotateurs
- ▹**Sycophancy :** tendance à confirmer les croyances de l'utilisateur plutôt qu'à fournir des réponses exactes

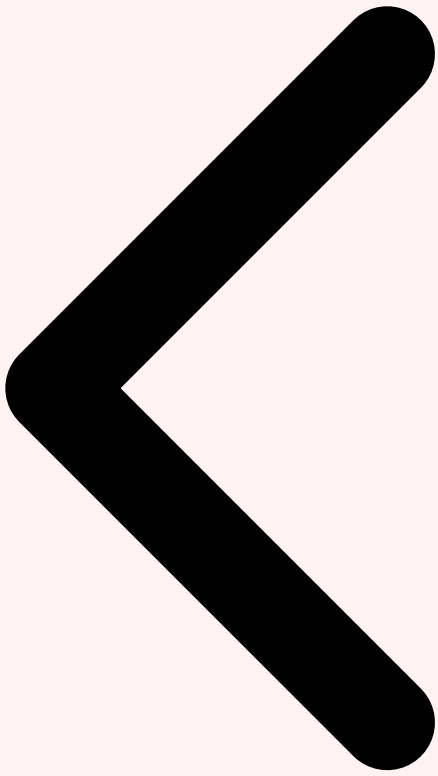
- ▹**Coût prohibitif :** le pipeline RLHF complet coûte entre 1 et 10 millions d'euros pour un modèle 70B+



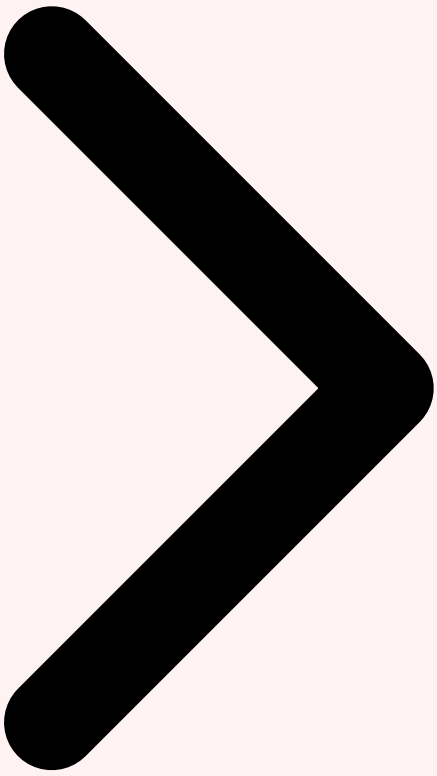


3 DPO et alternatives au RLHF

Le **Direct Preference Optimization (DPO)**, introduit par Rafailov et al. (2023), a représenté une rupture méthodologique en démontrant qu'il est possible d'obtenir des résultats d'alignement comparables au RLHF sans reward model séparé ni algorithme PPO. Le DPO optimise directement les préférences humaines, offrant une **réduction de 60 à 80% du coût computationnel**. Les variantes **IPO**, **KTO** (signaux binaires uniquement), **ORPO** (combine SFT et alignement) et **SimPO** enrichissent l'écosystème. La tendance en 2026 est aux **pipelines multi-étapes** combinant SFT, DPO/KTO généraux, puis raffinement ciblé sur la sécurité.

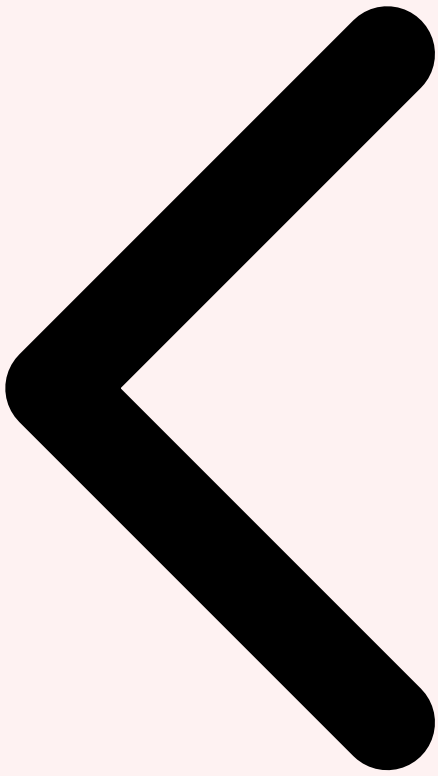


RLHF DPO et Alternatives Constitutional AI

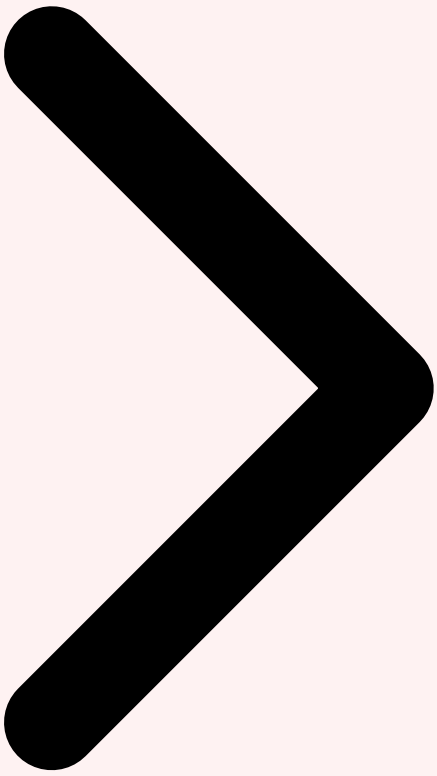


4 Constitutional AI : principes et mise en oeuvre

Le **Constitutional AI (CAI)**, développé par Anthropic, remplace partiellement le feedback humain par un **ensemble de principes éthiques formalisés** — la "constitution". Le processus SL-CAI génère des réponses problématiques puis les fait réviser par le modèle lui-même en référence à ses principes. Le RL-CAI utilise le modèle comme juge constitutionnel pour comparer des paires de réponses. Les avantages incluent la **scalabilité**, la **cohérence**, la **transparence** (constitution auditable) et l'**adaptabilité** sectorielle. Les limites incluent le risque de circularité, la difficulté de formulation des principes, et la sur-prudence (excessive refusal). Pour approfondir, consultez Embodied AI : Agents Physiques, Robotique et Sécurité en 2026.



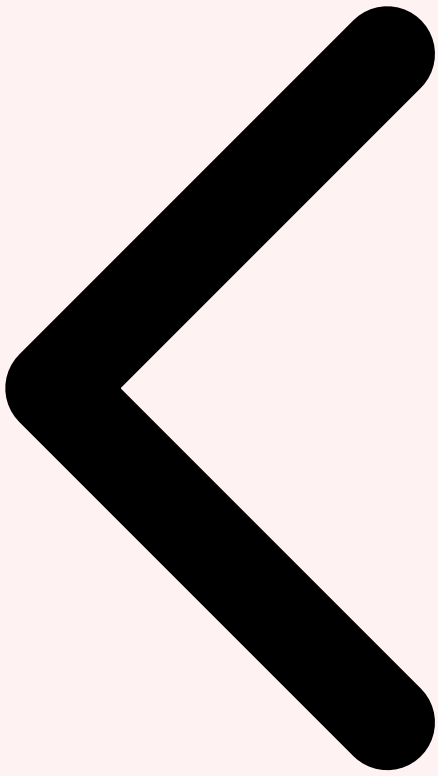
DPO et Alternatives Constitutional AI Audit d'Alignement



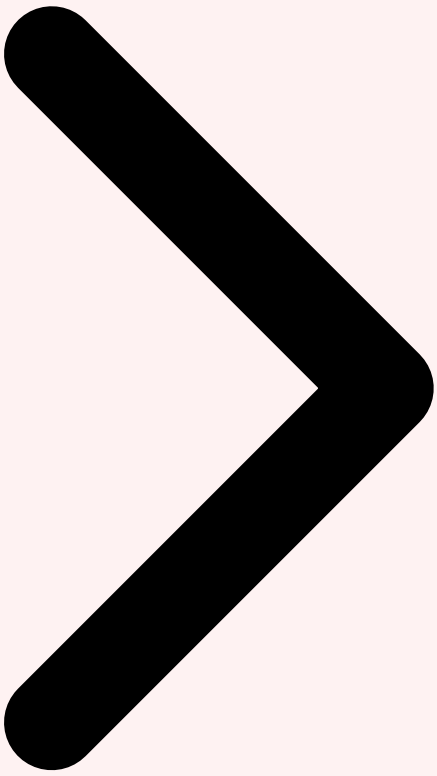
Cas concret
En 2024, des chercheurs de Cornell ont publié une étude démontrant l'empoisonnement de données d'entraînement de modèles de vision par ordinateur avec seulement 0.01% d'images malveillantes, suffisant pour créer des backdoors indétectables par les méthodes de validation standard.

5 Audit d'alignement en entreprise

L'audit d'alignement s'articule autour de cinq axes : **Safety** (refus des contenus dangereux), **Helpfulness** (utilité des réponses), **Honesty** (calibration et reconnaissance des limites), **Fairness** (biais sur les dimensions protégées), et **Robustness** (résistance adversariale). Les outils incluent **Inspect AI**, **HELM**, **DeepEval** et **Garak**. L'audit doit être réalisé avant chaque mise en production, trimestriellement, et après chaque changement significatif. Les résultats alimentent un registre de conformité IA exigé par l'AI Act.

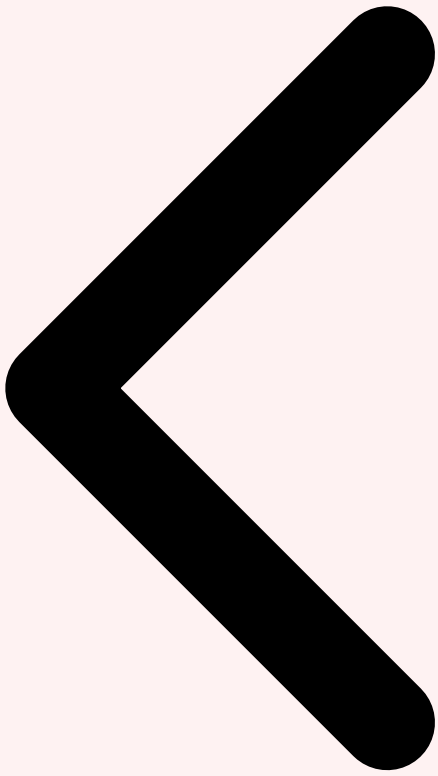


Constitutional AI Audit d'Alignement Red Teaming

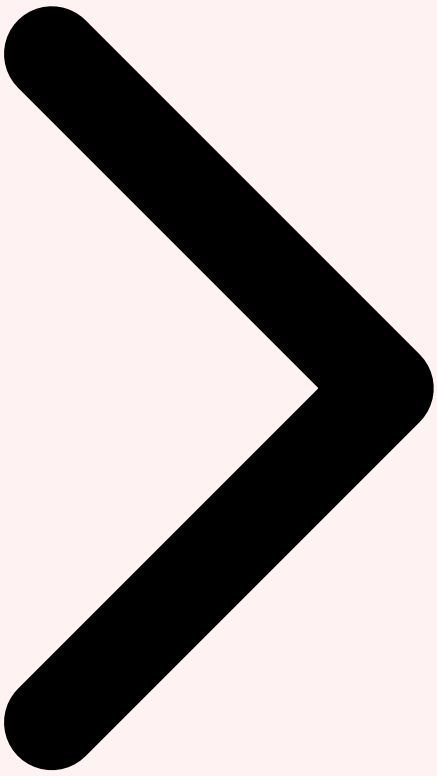


6 Red teaming pour l'alignement

Le **red teaming d'alignement** cible les **défaillances subtiles** : sycophancy, biais implicites, incohérences décisionnelles et sandbagging. La méthodologie en quatre phases — cadrage, exploration manuelle, amplification automatisée (PyRIT, Garak), et rapport — permet une couverture systématique. Le risque de **"deceptive alignment"** — où le modèle se comporte bien durant les audits mais pas en production — est activement étudié par les laboratoires de recherche en sécurité IA.



Audit d'Alignement Red Teaming Implications Réglementaires



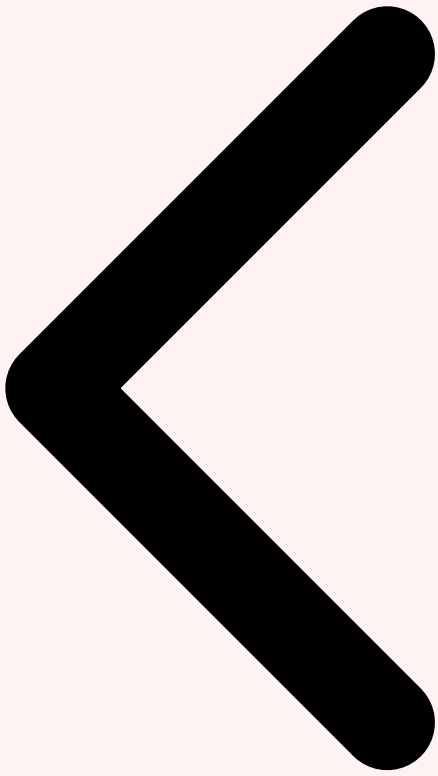
7 Implications réglementaires

L'**AI Act européen** impose des exigences de robustesse, non-discrimination et transparence. Le **NIST AI RMF** recommande l'évaluation quantitative de l'alignement. L'**ISO/IEC 42001** fournit le cadre organisationnel. Les sanctions peuvent atteindre **35 millions d'euros ou 7% du CA mondial**. La conformité exige un programme d'alignement documenté incluant politique formalisée, méthodes, résultats d'audit, métriques de suivi et procédures de correction.

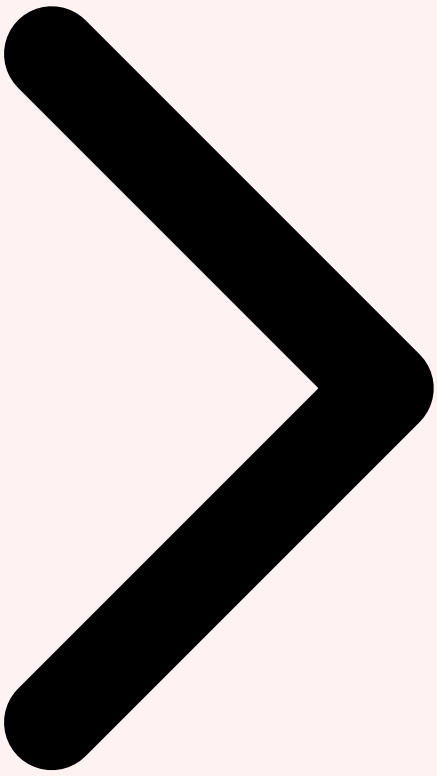
Checklist réglementaire alignement IA :

- ✓**Politique d'alignement formalisée** avec objectifs, valeurs encodées et critères de conformité
- ✓**Documentation technique** des méthodes d'alignement (RLHF, DPO, CAI) avec traçabilité
- ✓**Rapports d'audit périodiques** sur les 5 axes (Safety, Helpfulness, Honesty, Fairness, Robustness)

- ✓**Rapports de red teaming** classés par criticité avec preuves de remédiation
- ✓**Monitoring continu** des métriques d'alignement en production avec alertes de dérive



Red Teaming Implications Réglementaires Conclusion



8 Conclusion et recommandations

L'alignement des LLM en 2026 est un domaine en pleine maturation. Les trois schémas — RLHF, DPO et Constitutional AI — se combinent dans des **pipelines hybrides**. L'alignement n'est plus optionnel mais une **exigence opérationnelle et réglementaire**. Les organisations doivent intégrer l'alignement comme une discipline à part entière dans leur gouvernance IA. Pour approfondir, consultez Gouvernance Globale de l'IA 2026 : Alignement International.

8 recommandations pour les décideurs :

- 1.**Définir une politique d'alignement** formalisée avant tout déploiement de LLM
- 2.**Privilégier les modèles avec alignement auditable** — approches CAI avec constitution documentée
- 3.**Investir dans le DPO/KTO pour le fine-tuning interne** — rapport qualité-prix optimal

- 4.**Conduire des audits sur les 5 axes** avant chaque mise en production et trimestriellement
- 5.**Intégrer le red teaming d'alignement** au cycle de développement
- 6.**Monitorer les métriques en production** — taux de refus, cohérence, biais, feedback utilisateur
- 7.**Documenter la conformité AI Act** — registre, audits, red teaming, procédures de correction
- 8.**Former les équipes** aux enjeux de l'alignement — développeurs, product owners et juridiques

L'alignement des LLM est un processus continu, pas un état final. Les organisations qui investissent dès maintenant dans une **culture de l'alignement** seront les mieux positionnées pour exploiter le potentiel transformatif des LLM tout en maîtrisant les risques.

Besoin d'un accompagnement expert ?

Nos consultants en cybersécurité et IA vous accompagnent dans vos projets de sécurisation des LLM. Devis personnalisé sous 24h.

Références et ressources externes

- OWASP LLM Top 10 — Les 10 risques majeurs pour les applications LLM
- MITRE ATLAS — Framework de menaces pour les systèmes d'intelligence artificielle
- NIST AI RMF — AI Risk Management Framework du NIST
- arXiv — Archive ouverte de publications scientifiques en IA
- HuggingFace Docs — Documentation de référence pour les modèles de ML

Pour approfondir ce sujet, consultez notre outil open-source llm-vulnerability-scanner qui facilite l'analyse des vulnérabilités des LLM.

Sources et références : ArXiv IA · Hugging Face Papers

FAQ

Qu'est-ce que AI Safety et Alignement ?

Le concept de AI Safety et Alignement est détaillé dans les premières sections de cet article, qui couvrent les fondamentaux, les enjeux et le contexte opérationnel. Pour un accompagnement sur ce sujet, contactez nos experts.

Pourquoi AI Safety et Alignement est-il important en cybersécurité ?

La compréhension de AI Safety et Alignement permet aux équipes de sécurité d'améliorer leur posture défensive. Les sections « Table des Matières » et « 1 Introduction : Le défi de l'alignement des LLM » détaillent les raisons de cette importance. Pour un accompagnement sur ce sujet, contactez nos experts.

Comment mettre en œuvre les recommandations de cet article ?

Les recommandations pratiques sont détaillées tout au long de l'article, avec des commandes, des outils et des méthodologies éprouvées. La section « Conclusion » fournit une synthèse actionnable. Pour un accompagnement sur ce sujet, contactez nos experts.

Conclusion

Cet article a couvert les aspects essentiels de Table des Matières, 1 Introduction : Le défi de l'alignement des LLM, 2 RLHF : processus, reward models et limites. La mise en pratique de ces recommandations permet de renforcer significativement la posture de securite de votre organisation.

Ayi NEDJIMI Consultants — Expert cybersécurité offensive & intelligence artificielle

ayinedjimi-consultants.fr · ayi@ayinedjimi-consultants.fr