SOUVERAINETÉ · OPEN-WEIGHT

Formation IA souveraine : déployez Mistral open-weight chez vous

Une IA souveraine en entreprise, c'est un modèle de langage qui tourne sur une infrastructure que vous contrôlez, sans envoyer vos données à un service externe. En deux jours, cette formation apprend à vos équipes techniques à choisir un modèle open-weight de Mistral AI, à vérifier sa licence, à le servir avec vLLM et à le sécuriser.

DURÉE	FORMAT	PUBLIC
2 jours (14 h)	Distanciel, INTRA ou INTER	Équipes infrastructure, DevOps, sécurité

PRÉREQUIS	TARIF	SESSIONS
Administration Linux, conteneurs (Docker), notions de réseau	dès 1 390 € HT / personne	Sur demande

Objectifs pédagogiques

✓ **Choisir un modèle open-weight**
Comparer les modèles Mistral disponibles selon la tâche, la taille et le matériel disponible.

✓ **Dimensionner l'infrastructure**
Estimer la mémoire GPU, le débit et la latence attendus pour votre volume d'usage.

✓ **Sécuriser le service**
Restreindre l'accès réseau, authentifier les appels, journaliser et filtrer les contenus.

✓ **Lire une licence**
Distinguer Apache 2.0 et la licence Mistral Ecosystem Open-Weight, et repérer les cas qui exigent un accord commercial.

✓ **Servir le modèle**
Déployer un serveur d'inférence avec vLLM et exposer une API compatible avec vos applications.

✓ **Exploiter dans la durée**
Mettre à jour un modèle, surveiller la qualité et documenter le système pour l'audit.

Programme détaillé

Proposition de déroulé sur deux jours, adaptée à votre environnement (sur site, cloud privé, poste de test) lors du cadrage.

Demi-journée 1 — Choisir son modèle et son architecture

- Open-weight, open source, API : définitions et conséquences pratiques.
- Panorama des modèles Mistral téléchargeables et de leurs licences.
- Architectures possibles : serveur unique, cluster, cloud privé, offre Enterprise sur site.
- Dimensionnement : taille du modèle, quantification, mémoire GPU, contexte, nombre d'utilisateurs simultanés.
- Quiz d'entrée.

Demi-journée 2 — Déployer le serveur d'inférence

- Récupérer les poids sur Hugging Face (jeton d'accès, conditions d'accès des modèles).
- Servir le modèle avec vLLM, en conteneur Docker.
- Exposer une API compatible OpenAI pour vos applications existantes.
- Alternatives présentées dans la documentation Mistral : TGI, TensorRT-LLM, SkyPilot.
- Atelier : déploiement d'un modèle compact et premiers tests de charge.

Demi-journée 3 — Sécuriser l'IA souveraine

- Fermer le service : réseau privé, reverse proxy, authentification, quotas.
- Filtrer les entrées et les sorties ; présentation de Shieldstral, classifieur de sécurité open-weight de Mistral.
- Injection de prompt, exfiltration via les outils, empoisonnement de documents.
- Journalisation, conservation, protection des données personnelles traitées.

Demi-journée 4 — Exploiter, mettre à jour, documenter

- Brancher une recherche documentaire (RAG) sur vos sources internes.
- Évaluer la qualité avant et après une mise à jour de modèle.
- Supervision : disponibilité, latence, consommation GPU.

- Documentation du système pour la conformité (registre, analyse de risques).
- Cas pratique évalué : dossier d'architecture et plan de déploiement de votre projet.
- Quiz de sortie.

Public et prérequis

- Ingénieurs infrastructure et DevOps chargés d'héberger un modèle.
- Équipes sécurité (RSSI, SOC, architectes sécurité) qui doivent valider le déploiement.
- Architectes et responsables techniques qui comparent IA souveraine et API.

Prérequis : administration Linux, utilisation de Docker, notions de réseau (ports, proxy, TLS). Un accès à une machine de test avec GPU est utile ; à défaut, nous travaillons sur un modèle compact et un environnement fourni pour les ateliers.

Modalités

Public	Infrastructure, DevOps, sécurité, architectes
Prérequis	Linux, Docker, notions de réseau
Durée	2 jours (14 heures), consécutifs ou espacés
Format	Distanciel, en INTRA ou en INTER
Tarif	Dès 1 390 € HT par personne
Prochaine session	Sur demande
Modalités et délais d'accès	Inscription après un appel de cadrage ; accès sous 15 jours après accord de financement
Méthodes pédagogiques	Ateliers techniques guidés, étude d'architecture, revue collective
Évaluation	Quiz d'entrée et de sortie, cas pratique évalué
Documents remis	Support de formation, fichiers de configuration des ateliers, attestation nominative
Accessibilité	Référent handicap joignable à contact@yukisolutions.fr
Indicateurs de résultats	9,1/10 de satisfaction en fin de session ; plus de 100 professionnels formés

Inscription et devis — un appel de cadrage de 30 min, sans engagement, pour adapter le programme à vos cas : http://rel-ia.com/contact/ · contact@yukisolutions.fr