+33 (0)1 48 10 09 46
formation@ambient-it.net
140 Avenue Jean Lolive 93500 Pantin - France

Mis à jour le 11/05/2026

S'inscrire

Formation Distilabel

3 jours (21 heures)

Présentation

Distilabel est un framework open source développé par Argilla permettant de concevoir des pipelines de génération de datasets synthétiques pour les modèles de langage de nouvelle génération.

Notre formation Distilabel vous permettra de maîtriser la conception de pipelines de génération de données pour les LLM, l'automatisation des workflows de création de datasets ainsi que l'intégration avec l'écosystème Argilla, Hugging Face et les principaux fournisseurs de modèles d'intelligence artificielle.

Vous apprendrez à produire des datasets conversationnels, créer des jeux de préférences pour l'alignement des modèles, superviser la qualité des données générées et mettre en place des workflows collaboratifs de validation humaine.

À l'issue de cette formation, vous serez en mesure de développer des pipelines Distilabel complets, générer des datasets exploitables pour le fine-tuning et l'évaluation des LLM, intégrer Argilla dans vos workflows IA et industrialiser la génération de données synthétiques dans un environnement moderne orienté LLMOps.

Comme toutes nos formations, celle-ci présente la dernière version stable de Distilabel et privilégie une approche résolument pratique et opérationnelle.

Objectifs

- Comprendre les concepts fondamentaux de Distilabel et d'Argilla
- Créer des pipelines de génération de datasets synthétiques pour les LLM
- Produire des données pour le fine-tuning, le RLHF et le DPO
- Évaluer et valider la qualité des datasets IA générative
- Intégrer Distilabel dans des workflows LLMOps
- Industrialiser des pipelines de génération de données IA

Public visé

- LLM Engineers
- AI Engineers
- Data Scientists orientés IA générative
- ML Engineers
- Ingénieurs MLOps
- Développeurs Python spécialisés IA

Pré-requis

- Connaissances de base en Python
- Notions de Machine Learning et de modèles de langage
- Connaissance générale des APIs IA et de l’écosystème Hugging Face appréciée

Programme de notre formation Distilabel avec Argilla

[Jour 1 - Matin]

Introduction à Distilabel et à l’écosystème Argilla

- Comprendre Distilabel et son rôle dans la génération de datasets synthétiques pour les LLM
- Découvrir l’écosystème Argilla, Hugging Face et Transformers
- Identifier les cas d’usage : fine-tuning, RLHF, DPO, RAG
- Installer et configurer un environnement Python pour Distilabel
- Comprendre les concepts de Steps, Tasks et Pipelines
- Atelier pratique : Création d’un premier pipeline de génération de données conversationnelles.

[Jour 1 - Après-midi]

Construction de pipelines de génération LLM

- Concevoir des pipelines modulaires de génération de données IA
- Utiliser le prompt engineering pour produire des datasets cohérents
- Générer des données multi-turns et scénarios conversationnels
- Exploiter différents fournisseurs de modèles : OpenAI, Mistral, vLLM
- Structurer des workflows réutilisables pour les projets IA
- Atelier pratique : Création d’un pipeline avancé de génération d’instructions.

Préparation des datasets pour le fine-tuning

- Comprendre les formats de données pour le fine-tuning supervisé
- Créer des datasets adaptés aux modèles conversationnels
- Générer des jeux de données spécialisés métier
- Nettoyer et normaliser les données synthétiques

- Organiser les datasets pour l’entraînement et l’évaluation
- Atelier pratique : Création d’un dataset prêt pour le fine-tuning d’un LLM.

[Jour 2 - Matin]

Distilabel et l’alignement des modèles LLM

- Comprendre les concepts de RLHF et DPO
- Générer des datasets de préférences et de ranking
- Créer des données pour l’évaluation comparative de réponses
- Structurer des jeux de données d’alignement
- Automatiser la génération de données de préférence
- Atelier pratique : Création d’un dataset de préférences pour alignement de modèles.

[Jour 2 - Après-midi]

Intégration avec Argilla et validation humaine

- Découvrir les fonctionnalités d’Argilla pour l’annotation IA
- Importer et versionner des datasets dans Argilla
- Organiser les workflows de validation humaine
- Annoter et corriger des sorties générées par les LLM
- Construire un workflow collaboratif de revue de données
- Atelier pratique : Validation et annotation d’un dataset dans Argilla.

Évaluation et qualité des datasets IA

- Mesurer la qualité des données générées automatiquement
- Détecter les hallucinations et incohérences des modèles
- Évaluer la diversité et la pertinence des réponses
- Mettre en place des métriques d’évaluation LLM
- Construire des datasets de benchmark
- Atelier pratique : Audit qualité d’un dataset synthétique généré avec Distilabel.

[Jour 3 - Matin]

Distilabel pour les workflows RAG et IA générative

- Générer des données pour assistants conversationnels et systèmes RAG
- Produire des jeux de questions/réponses contextualisés
- Créer des datasets multilingues et spécialisés
- Structurer des pipelines de génération documentaire
- Exploiter Distilabel dans des architectures IA générative
- Atelier pratique : Génération d’un dataset RAG pour assistant métier.

[Jour 3 - Après-midi]

Industrialisation et automatisation des pipelines

- Automatiser l’exécution des pipelines Distilabel
- Utiliser Docker pour standardiser les environnements
- Versionner les datasets et pipelines IA
- Superviser les workflows et les coûts d’inférence
- Déployer des pipelines de génération en production
- Atelier pratique : Industrialisation d’un pipeline complet de génération de données IA.

Bonnes pratiques et projet fil rouge

- Définir les bonnes pratiques de génération de données IA
- Optimiser les performances et la qualité des pipelines
- Sécuriser les accès API et les données générées
- Construire une stratégie de maintenance des datasets
- Préparer les workflows pour des projets IA à grande échelle
- Atelier pratique : Projet final de génération, validation et exploitation d’un dataset LLM complet.

Sociétés concernées

Cette formation s’adresse à la fois aux particuliers ainsi qu’aux entreprises, petites ou grandes, souhaitant former ses équipes à une nouvelle technologie informatique avancée ou bien à acquérir des connaissances métiers spécifiques ou des méthodes modernes.

Positionnement à l'entrée en formation

Le positionnement à l'entrée en formation respecte les critères qualité Qualiopi. Dès son inscription définitive, l'apprenant reçoit un questionnaire d’auto-évaluation nous permettant d’apprécier son niveau estimé sur différents types de technologies, ses attentes et objectifs personnels quant à la formation à venir, dans les limites imposées par le format sélectionné. Ce questionnaire nous permet également d’anticiper certaines difficultés de connexion ou de sécurité interne en entreprise (intraentreprise ou classe virtuelle) qui pourraient être problématiques pour le suivi et le bon déroulement de la session de formation.

Méthodes pédagogiques

Stage Pratique : 60% Pratique, 40% Théorie. Support de la formation distribué au format numérique à tous les participants.

Organisation

Le cours alterne les apports théoriques du formateur soutenus par des exemples et des séances de réflexions, et de travail en groupe.

Validation

À la fin de la session, un questionnaire à choix multiples permet de vérifier l’acquisition correcte

des compétences.

Sanction

Une attestation sera remise à chaque stagiaire qui aura suivi la totalité de la formation.