New

Applications d’IA générative avec LangChain

Concevoir, orchestrer et déployer des applications d'IA générative avec LangChain

Présentiel ou en classe à distance

3 jours (21 h)

Prix inter : 2.490,00 € HT
Forfait intra : 7.690,00 € HT

Réf.: IA117

La formation **Applications d’IA générative avec LangChain** permet de concevoir des **applications basées sur les LLM** en s'appuyant sur **LangChain**, les architectures **RAG (Retrieval-Augmented Generation)** et les premiers concepts d'**IA agentique**. Elle couvre l'orchestration des chaînes de traitement, l'intégration de sources de données, le **prompt engineering** et les bonnes pratiques de développement pour créer des applications d'IA générative performantes, fiables et prêtes à évoluer vers des environnements de production.

A qui s'adresse cette formation ?

Pour qui

- Développeurs AI, data engineers

Prérequis

- Connaissances en Python et LLM
- Disposez-vous des connaissances nécessaires pour suivre cette formation ? Testez-vous !

Programme

1 - IA générative, LLM et écosystème LangChain

- Panorama de l’intelligence artificielle générative et des grands modèles de langage
- Fonctionnement d’un LLM : tokens, fenêtre de contexte et paramètres de génération
- Présentation de l’écosystème LangChain : composants, modules et principaux cas d’usage
- Comparaison des modèles propriétaires et des modèles ouverts selon différents critères de choix
- Configuration de l’environnement de développement et gestion des clés d’API

Atelier

Configurer l’environnement LangChain et les accès aux modèles de langage

Effectuer des appels à un LLM et ajuster ses paramètres de génération

Comparer le comportement de deux modèles sur une même tâche

Mesurer la consommation de tokens associée à une requête

2 - Prompts, chaînes et sorties structurées

- Conception de prompts et création de gabarits réutilisables avec les prompt templates
- Présentation du langage de composition LCEL, ou LangChain Expression Language
- Construction d’une chaîne de traitement associant un modèle, un prompt et un parseur
- Utilisation de parseurs de sortie pour générer des réponses structurées
- Application des techniques de prompt engineering : few-shot prompting et instructions de rôle

Atelier

Créer des gabarits de prompts paramétrés

Composer une chaîne avec LCEL en combinant un prompt, un modèle et un parseur

Produire des sorties structurées directement exploitables par une application

Enchaîner plusieurs chaînes afin de réaliser une tâche en plusieurs étapes

3 - Mémoire et applications conversationnelles

- Enjeux de la gestion du contexte dans une application conversationnelle
- Conservation de l'historique des conversations et gestion de la mémoire
- Construction d'une chaîne conversationnelle intégrant un historique de messages
- Mise en oeuvre de stratégies de troncature et de résumé pour limiter la taille du contexte
- Gestion des sessions et prise en charge de plusieurs utilisateurs
  Atelier

  Concevoir une chaîne conversationnelle capable de gérer l'historique des échanges

  Mettre en place une stratégie de résumé pour limiter la taille du contexte

  Gérer plusieurs sessions de conversation isolées

  Tester l'application au moyen d'un scénario de dialogue en plusieurs tours

4 - Embeddings, bases vectorielles et recherche sémantique

- Principe des embeddings et de la représentation vectorielle des contenus textuels
- Chargement et découpage des documents avec les document loaders et les text splitters
- Fonctionnement des bases de données vectorielles : indexation et recherche par similarité
- Utilisation des retrievers pour interroger une base de connaissances
- Définition et application des critères de qualité d'une recherche sémantique
  Atelier

  Charger et découper un corpus de documents

  Calculer les embeddings et indexer les représentations obtenues dans une base vectorielle

  Interroger la base avec un retriever et analyser la pertinence des résultats

  Ajuster le découpage des documents et les paramètres de recherche afin d'améliorer la pertinence des réponses

5 - Applications RAG : génération augmentée par la recherche

- Présentation du principe de la génération augmentée par la recherche, ou RAG
- Architecture d'une chaîne RAG : recherche, récupération du contexte et génération de la réponse
- Conception d'un prompt de génération intégrant les documents récupérés
- Citation des sources et mise en place de mécanismes visant à limiter les hallucinations
- Évaluation de la qualité des réponses produites par une application RAG
  Atelier

  Assembler une chaîne RAG combinant un retriever et un modèle de génération

  Concevoir un prompt de génération exploitant les documents récupérés

  Ajouter la citation des sources aux réponses produites

  Évaluer la qualité des réponses à partir d'un jeu de questions de test

6 - Agents IA, outils et orchestration avec LangChain

- Principe de fonctionnement d'un agent IA : raisonnement et sélection des actions
- Mise à disposition d'outils pour un agent : recherche, calcul et appels d'API
- Compréhension du schéma de fonctionnement d'un agent et du mécanisme d'appel d'outils par le modèle
- Orchestration de flux complexes et introduction à LangGraph
- Automatisation d'une tâche générative composée de plusieurs étapes
- Limites des agents, mise en place de garde-fous et supervision
  Atelier

  Définir les outils disponibles et les mettre à la disposition de l'agent

Construire un agent et observer son raisonnement étape par étape

Automatiser une tâche générative nécessitant l'utilisation de plusieurs outils

Mettre en place des garde-fous afin de limiter et de contrôler les actions de l'agent

7 - Qualité, observabilité et évaluation des applications d'IA générative

- Enjeux de fiabilité d'une application d'IA générative déployée en production
- Traçabilité et observabilité des chaînes et des agents avec LangSmith
- Évaluation systématique au moyen de jeux de test et de critères de qualité
- Détection et réduction des hallucinations et des réponses hors sujet
- Validation des sorties et mise en place de garde-fous applicatifs
  Atelier

Instrumenter une application afin de tracer les chaînes de traitement et les agents

Construire un jeu de test et définir des critères mesurables de qualité

Évaluer l'application et analyser les cas d'échec Ajouter des règles de validation des réponses générées

8 - Performance, maîtrise des coûts et sécurité

- Maîtrise des coûts liés à la consommation de tokens et au choix des modèles
- Mise en cache des réponses afin d'améliorer les performances et de réduire la latence
- Diffusion progressive des réponses avec le streaming
- Sécurisation des entrées et protection contre les attaques par injection de prompt
- Gestion des limites de débit et préparation de la montée en charge
  Atelier

Mesurer et réduire la consommation de tokens d'une application

Mettre en place un mécanisme de cache et mesurer le gain obtenu

Activer la diffusion en continu des réponses

Implémenter une protection contre les attaques par injection de prompt

9 - Déploiement d'une application d'IA générative

- Exposition d'une application d'IA générative au moyen d'une API
- Conteneurisation de l'application avec Docker
- Gestion de la configuration par environnement et sécurisation des clés d'API
- Surveillance en production : latence, coûts, taux d'erreur et qualité des réponses
- Stratégies de mise à jour des prompts, des modèles et des bases de connaissances
  Atelier

Exposer l'application au moyen d'une API et tester les requêtes

Conteneuriser l'application avec Docker et gérer sa configuration

Mettre en place la surveillance de la latence, des coûts et des erreurs

Réaliser un test de charge et vérifier le comportement de l'application

Les objectifs de la formation

- Concevoir des applications exploitant les capacités de l'IA générative
- Intégrer des modèles de langage (LLM) dans des applications métier
- Automatiser des processus et des tâches à l'aide de l'IA générative
- Déployer et maintenir des applications intégrant des modèles d'IA générative

Evaluation

- Pendant la formation, le formateur évalue la progression pédagogique des participants via des QCM, des mises en situation et des travaux pratiques. Les participants passent un test de positionnement avant et après la formation pour valider leurs compétences acquises.

Les points forts de la formation

- Une formation couvrant l'ensemble du cycle de vie d'une application d'IA générative, de la conception de chaînes LangChain (LCEL) à la création d'agents intelligents, jusqu'au déploiement et au suivi en production
- Des ateliers pratiques mettant en oeuvre l'écosystème LangChain (LCEL, LangGraph, LangSmith, bases de données vectorielles et agents IA) à travers des cas d'usage concrets
- Un volet dédié à l'industrialisation et à la mise en production des applications d'IA générative, intégrant les bonnes pratiques de sécurité, d'observabilité, d'optimisation des coûts et de déploiement

Dates et villes 2026 - Référence IA117

Dernières places disponibles Session garantie

A distance

du 19 oct. au 21 oct.	du 7 déc. au 9 déc.

Paris

du 19 oct. au 21 oct.	du 7 déc. au 9 déc.