

IA Générative – Les modèles de langages massifs (LLMs)

INFORMATIONS GÉNÉRALES

Type de formation : Formation continue

Éligible au CPF : Non

Domaine : IA, Big Data et Bases de données

Action collective : Non

Filière : IA

Rubrique : Microsoft Azure OpenAI Service

Code de formation : MSAI050

€ Tarifs

Prix public : 2050 €

Tarif & financement :

Nous vous accompagnons pour trouver la meilleure solution de financement parmi les suivantes :

Le plan de développement des compétences de votre entreprise : rapprochez-vous de votre service RH.

Le dispositif FNE-Formation.

L'OPCO (opérateurs de compétences) de votre entreprise.

France Travail: sous réserve de l'acceptation de votre dossier par votre conseiller Pôle Emploi.

CPF -MonCompteFormation

Contactez nous pour plus d'information : contact@aston-institut.com

PRÉSENTATION

Objectifs & compétences

À l'issue de cette formation, les participants seront en capacité de :

- Utiliser Azure OpenAI Service
- Appliquer l'ingénierie rapide avec Azure OpenAI Service

Public visé

Chefs de projets
Développeurs
Data scientists

Pré-requis

Une connaissance de base des principes de Machine Learning et de Deep Learning
La maîtrise d'une langue de script type Python est recommandé

📍 Lieux & Horaires

Durée : 14 heures

Délai d'accès : Jusqu'à 8 jours avant le début de la formation, sous condition d'un dossier d'inscription complet

PROGRAMME

1 - Introduction

Un changement de paradigme

Qu'est ce qui change ?

Une publication fondatrice

Une loi d'échelle pour les modèles de langage

Évolution temporelle des LLMs

De nouveaux écosystèmes

L'ère du Post Deep Learning

Personnalisation par Prompts

Personnalisation par enrichissement

Personnalisation par réglage fin

2 - Cas d'usage

Agents conversationnels et assistants virtuels

Génération de code et debuggage

Analyse de sentiments / opinions

Classification de texte et clusterisation

Synthèse de texte ou corpus

Traduction

Génération de contenu

Autres cas d'usages significatifs

LAB : Proof of concept sur cas concrets

3 - Fondations

Le traitement du langage naturel (TAL)

L'architecture disruptive des Transformers

La tokenisation des textes

L'encodeur d'un Transformer

La couche d'embedding

L'encodage de positionnement

Vecteur de positionnement

Le mécanisme d'attention multi-têtes

Points essentiels du mécanisme d'attention

La "spécialisation" des têtes d'attention

Calcul des scores d'attention

Addition et Normalisation des sorties

📅 Prochaines sessions

Consultez-nous pour les prochaines sessions.

Le Décodeur d'un Transformer
L'entraînement d'un Transformer
La couche d'auto-attention masquée
La couche d'attention du décodeur
Les couches supérieures du décodeur

4 - En pratique

Choisir un LLM
Critères de choix
Trois classes de modèles
Modèles à encodeur simple
Focus modèles BERTs
Modèles à décodeur simple
Focus modèles GPTs
Un foisonnement de modèles dérivés
La bataille des LLMs
La course vers des LLMs légers
L'exemple de LLaMa
Trois approches de réduction
Écosystèmes clés
APIs de modèles Fermés
HuggingFace et les modèles ouverts
Écosystèmes applicatifs type LangChain
LLMops et MLFlow
Atelier
Prise en main des écosystèmes LLMs clés

5 - Mise en oeuvre

Choix service / in house / hybrid
In house workflow
Service workflow
Écosystèmes d'entraînement et d'inférence
L'entraînement d'un modèle massif
L'étape d'évaluation des modèles
Le réglage des hyperparamètres
Déploiement d'un modèle
Model fine-tuning
Prompt engineering
MLOps d'un LLMs
LAB : Environnement de déploiement d'un LLM

6 - Le Prompt Engineering

Configuration des paramètres des LLMs
Qu'est ce qu'un token ?
Notion de distribution des probabilités des LLMs
Les échantillonnages Top-K et top-p
La température du modèle
Le réglage des paramètres en pratique
Les composantes d'un prompt
Règles générales
L'approche Few-Shot Learning
Zero, one to Few-shot learning
L'approche Chain of thoughts
L'incitation par chaînes de pensées
Des approches plus avancées
ReAct Prompting
Méthode ReAct
Atelier
Prompt Engineering sur cas concrets

7 - LLMs augmentés

Au delà du prompt, l'enrichissement des LLMs
Ajout de capacité mémorielle
Mémoire tampon (Buffer Memory)
Plusieurs mécanismes de mémorisation
Les mémoires de l'écosystème LangChain
Élargissement des connaissances
Retrieval Augmented Generation (RAG)
Le partitionnement des textes externes
Projection sémantique des documents (Embeddings)
Les bases de données vectorielles
Les algorithmes du search dans les bases vectorielles
Une galaxie d'outils possibles !
Atelier
Mise en oeuvre d'un agent conversationnel

8 - Déploiement de LLMs

Quand le prompt engineering ne suffit plus

Qu'est ce que le réglage fin
Trois techniques classiques
Reinforcement learning by Human feedback (RLHF)
Détails d'un réglage fin Supervisé
Les trois options pour l'ajustement des paramètres
Les approches PEFT (Parameter Efficient Tuning)
La méthode LoRA (Low Rank Adaptation)
Une variante efficiente : QLoRA
Qu'est ce que la mise en service d'un LLM
Journaliser le modèle dans le registre des modèles
Création d'un point de terminaison vers le modèle
Interroger le point de terminaison

MODALITÉS

Modalités

Modalités : en présentiel, distanciel ou mixte . Toutes les formations sont en présentiel par défaut mais les salles sont équipées pour faire de l'hybride. – Horaires de 9H à 12H30 et de 14H à 17H30 soit 7H – Intra et Inter entreprise.

Pédagogie : essentiellement participative et ludique, centrée sur l'expérience, l'immersion et la mise en pratique. Alternance d'apports théoriques et d'outils pratiques.

Ressources techniques et pédagogiques : Support de formation au format PDF ou PPT Ordinateur, vidéoprojecteur, Tableau blanc, Visioconférence : Cisco Webex / Teams / Zoom.

Pendant la formation : mises en situation, autodiagnosics, travail individuel ou en sous-groupe sur des cas réels.

Méthode

Fin de formation : entretien individuel.

Satisfaction des participants : questionnaire de satisfaction réalisé en fin de formation.

Assiduité : certificat de réalisation.

Validations des acquis : grille d'évaluation des acquis établie par le formateur en fin de formation.