Modèles de Langage Spécialisés par Domaine (Domain-Specific LLMs) : fine-tuning & déploiement

Lien : https://innov-systems.com/formation/modeles-de-langage-specialises-par-domaine-domain-specific-llms-fine-tuning-deploiement

DURÉE
8 jours (56h)

RÉFÉRENCE
BDI214

CATÉGORIE
Intelligence Artificielle Générative (IAG)

Innov Systems

OBJECTIFS DE LA FORMATION

À l'issue de cette formation, vous serez capable de :

- Comprendre les principes des Domain-Specific LLMs et leurs cas d'usage
- Sélectionner le modèle de base adapté à un domaine métier
- Concevoir et préparer des jeux de données de qualité pour le fine-tuning
- Maîtriser les techniques de fine-tuning (SFT, LoRA, QLoRA, PEFT)
- Optimiser les performances et les coûts d'entraînement
- Évaluer les performances d'un modèle spécialisé à l'aide de métriques pertinentes
- Déployer un LLM en environnement cloud, hybride ou on-premise
- Intégrer les modèles dans les applications et processus métiers
- Mettre en œuvre une gouvernance, une sécurité et une IA responsable
- Superviser les modèles en production et gérer leur cycle de vie
- Optimiser les modèles pour une inférence à grande échelle
- Réaliser un projet complet de fine-tuning et de déploiement d'un Domain-Specific LLM

POUR QUI ?

- Ingénieurs en intelligence artificielle
- Machine Learning Engineers
- Data Scientists
- MLOps Engineers
- AI Architects
- Développeurs IA générative
- Data Engineers
- Architectes Cloud
- DevOps et DevSecOps Engineers
- Consultants en intelligence artificielle
- Responsables innovation et transformation numérique
- Toute personne souhaitant concevoir, entraîner et déployer des modèles de langage spécialisés pour des applications métiers

Innov Systems

Programme détaillé

1 / Introduction aux Domain-Specific LLMs

- Différences entre LLMs généralistes et spécialisés
- Cas d’usage sectoriels (santé, finance, juridique, industrie, télécoms, secteur public)
- Critères de choix d’un modèle de base (open source ou propriétaire)

2 / Architecture des LLMs et principes du fine-tuning

- Transformers, embeddings et mécanismes d’attention
- Fine-tuning complet, PEFT, LoRA et QLoRA
- Comparaison entre fine-tuning, RAG et prompt engineering

3 / Constitution et préparation des jeux de données

- Collecte, nettoyage et annotation des données métier
- Qualité, représentativité et équilibrage des corpus
- Protection des données sensibles et anonymisation

4 / Environnements d'entraînement et optimisation

- Infrastructure GPU et accélération matérielle
- Frameworks (PyTorch, Hugging Face Transformers, DeepSpeed)
- Optimisation des coûts, performances et consommation mémoire

5 / Fine-tuning des modèles spécialisés

- Définition des hyperparamètres
- Supervised Fine-Tuning (SFT)
- Validation, checkpoints et gestion des versions des modèles

Innov Systems

6 / Évaluation et benchmarking des modèles

- Métriques de performance (perplexité, précision, F1, BLEU, ROUGE)
- Benchmarks métier et jeux de tests spécialisés
- Évaluation humaine et validation fonctionnelle

7 / Sécurité, gouvernance et IA responsable

- Gestion des biais et réduction des hallucinations
- Sécurisation des modèles et protection contre les attaques
- Conformité réglementaire, auditabilité et traçabilité

8 / Déploiement des Domain-Specific LLMs

- Déploiement on-premise, cloud et hybride
- Conteneurisation avec Docker et Kubernetes
- Inférence optimisée et mise à l'échelle des services

9 / Intégration avec les applications métier

- APIs REST, gRPC et microservices
- Intégration avec ERP, CRM, GED et outils métiers
- Agents IA et workflows intelligents

10 / Monitoring, observabilité et amélioration continue

- Surveillance des performances en production
- Détection de dérive (model drift et data drift)
- Réentraînement et amélioration continue

11 / Optimisation avancée des modèles

- Quantification, distillation et pruning
- Optimisation de la latence et du débit

Inov Systems

- Gestion des coûts d'inférence à grande échelle

12 / Cas pratiques et projet de bout en bout

- Fine-tuning d'un LLM pour un domaine métier spécifique
- Déploiement d'un service d'IA générative en production
- Bonnes pratiques, retours d'expérience et feuille de route industrielle

Approche pédagogique

- Support Ecrit et Projection
- Exposés Interactifs, Podcasts et Vidéos
- Brainstorming et Jeux de Rôle
- Cas Pratiques et Labs inclus pour leur impact opérationnel
- Test de Validation des Acquis des Connaissances

Prochaines dates programmées

08 au 17 Fév. 2027	Présentiel - Casablanca
05 au 14 Avr. 2027	Présentiel - Casablanca
31 Mai au 09 Juin 2027	Présentiel - Casablanca
26 Juil. au 04 Août 2027	Présentiel - Casablanca
20 au 29 Sep. 2027	Présentiel - Casablanca

Autres dates possibles sur demande. Contactez-nous pour organiser une session intra-entreprise.

Innov Systems

Réservation & Renseignements

Téléphone : +212 522 247 210
Email : contact@innov-systems.com
Web : https//www.innov-systems.com

Innov Systems — Tous droits réservés