Blue AcademIA
Blue AcademIA est une marque de BlueMarketing
contact@blueacademia.fr
www.blueacademia.fr

Programme de formation : Concevoir et déployer des architectures RAG et des agents IA

Les architectures RAG et les agents IA permettent de connecter les modèles d'intelligence artificielle aux données, outils et processus de l'entreprise afin de créer des assistants métier performants et fiables. Passer du prototype à la production nécessite toutefois de maîtriser les bonnes pratiques d'architecture, d'évaluation, d'orchestration et de déploiement.

Cette formation permet aux développeurs, data scientists et ingénieurs IA d'acquérir les méthodes et compétences nécessaires pour concevoir, évaluer et industrialiser des solutions IA avancées intégrant RAG, agents IA et connexions aux systèmes de l'entreprise.

Résolument orientée pratique, cette formation s'appuie sur des ateliers de développement, des cas d'usage concrets et l'application des concepts au cas d'entreprise des participants. L'objectif est de transformer les acquis de la formation en solutions directement exploitables dans leur environnement professionnel.

Durée : 35.00 heures ou jour(s) si formation en présentiel.

Lieu : À distance ou en présentiel

Profils des stagiaires

- Développeurs back-end / full-stack
- Data scientists
- Ingénieurs ML/IA
- Architectes logiciels
- Lead techs souhaitant industrialiser des solutions IA

Prérequis

- Maîtrise de Python (POO, async, gestion de packages).
- Notions REST/API.
- Familiarité avec Git, Docker et une IDE moderne (VS Code, PyCharm).
- Une expérience préalable d'appel à un LLM via API est un plus.
- Les pré-requis sont validés lors d'un échange avec le commanditaire du projet de formation et retranscrits dans le document de recueil des besoins. Il s'agit d'un document collaboratif partagé systématiquement avec le commanditaire à l'issue du premier entretien pour co-construire l'action de formation. Par ailleurs, nous adressons à chaque apprenant, 2 semaines avant le début de la formation, un questionnaire de préparation de la formation pour adapter au mieux la formation à ses besoins.

Objectifs pédagogiques

- Comprendre en profondeur le fonctionnement interne des LLM (tokenisation, attention, fenêtre de contexte).
- Concevoir une architecture RAG production-ready : chunking, embeddings, retrieval, re-ranking.
- Évaluer rigoureusement un système RAG (Ragas, DeepEval, métriques métiers).
- Construire des agents IA modulaires avec LangGraph, CrewAI ou Pydantic AI.
- Implémenter le Model Context Protocol (MCP) pour connecter des LLM à des systèmes internes.
- Sécuriser, observer et déployer une solution IA en production (CI/CD, monitoring, coûts).

Contenu de la formation

- Fondations des LLM modernes : ce qu'un ingénieur doit vraiment savoir
  - Architecture Transformer décodeur, attention, tokenisation BPE — ce qui impacte vos coûts.
  - Fenêtre de contexte (jusqu'à 1 M tokens en 2026), trade-offs latence / qualité / prix.
  - Panorama 2026 : Claude Opus 4 / Sonnet 4.5, GPT-5, Gemini 2.5, Mistral Large 3, modèles open-source (Llama 4, Mistral, Qwen).
  - Sampling, température, top-p, structured output (JSON mode, tool calling natif).
  - Comparer 4 LLM sur une même tâche métier (qualité, latence, coût/1 k tokens) et justifier un choix.
  - Cas réel : choisir entre Claude Opus 4 et un modèle open-source self-hosted pour un assistant juridique interne.
  - Implémenter un appel API avec structured output Pydantic pour extraire des données d'un PDF de contrat.
- Embeddings, recherche vectorielle et bases vectorielles
  - Modèles d'embeddings 2026 : text-embedding-3-large, voyage-3, BGE-M3, comparatif latence/qualité.
  - Bases vectorielles : Pinecone (managé), Weaviate, Qdrant, Chroma, pgvector — choisir selon le contexte.
  - Similarité cosinus, dot product, distances, métriques de récupération (recall@k, MRR, nDCG).
  - Recherche hybride : sparse (BM25) + dense, fusion RRF (Reciprocal Rank Fusion).
  - Indexer 10 000 documents internes (procédures qualité ISO) dans Qdrant avec pgvector en backup.
  - Implémenter une recherche hybride BM25 + embeddings pour un moteur de FAQ entreprise.
  - Benchmark : comparer 3 modèles d'embeddings sur un corpus d'entreprise et mesurer le recall@5.
- RAG naïf vs RAG avancé : techniques de production
  - Stratégies de chunking : taille fixe, sémantique, hiérarchique, parent-document.
  - Query transformation : HyDE, multi-query, query decomposition, step-back prompting.
  - Re-ranking avec cross-encoders (Cohere Rerank, BGE-reranker) — gains de précision typiques.
  - RAG agentique, Self-RAG, CRAG (Corrective RAG), GraphRAG (Microsoft).
  - Construire un RAG juridique : recherche sur 50 000 articles de jurisprudence avec citation des sources.
  - Ajouter un re-ranker Cohere et mesurer le gain de précision sur un benchmark interne.
  - Implémenter un RAG corrigé (CRAG) capable d'aller chercher sur le web si la réponse interne est insuffisante.
- Évaluation, observabilité et qualité d'un système RAG
  - Métriques RAG : faithfulness, answer relevancy, context precision/recall (framework Ragas).
  - Évaluation avec LLM-as-judge : DeepEval, Trulens, méthodologies, biais courants.
  - Constituer un golden dataset interne, gérer les régressions à chaque modification.
  - Observabilité : LangSmith, Langfuse, OpenTelemetry, alerting sur dérive de qualité.
  - Construire un golden dataset de 100 paires (question, réponse attendue) sur la base documentaire métier.
  - Lancer une évaluation Ragas comparant 3 versions d'un RAG et présenter le rapport aux décideurs.
  - Mettre en place LangSmith en production : tracing complet, détection automatique des hallucinations.
- Agents IA : du tool calling au ReAct pattern
  - Tool calling natif (Claude / OpenAI) : structurer ses outils, gérer les erreurs.
  - Patterns d'agents : ReAct, Plan-and-Execute, Reflexion, Tree-of-Thoughts.
  - Mémoire des agents : short-term, long-term, summary memory, vector memory.
  - Garde-fous : timeouts, max iterations, validation pydantic, fallback déterministe.
  - Construire un agent qui interroge une API interne SAP + une base SQL + le CRM pour répondre aux ventes.

    - Agent support client : qualifie un ticket, cherche dans la base de connaissance, ouvre un Jira si besoin.
    - Agent analyste : lit un Excel financier, calcule des ratios, produit un commentaire textuel auditable.
- Systèmes multi-agents modulaires avec LangGraph & CrewAI
    - LangGraph : graphe d'états, edges conditionnels, checkpoints, human-in-the-loop.
    - CrewAI : rôles, tâches, processus séquentiels et hiérarchiques.
    - Pydantic AI : agents typés, validation forte, intégration FastAPI native.
    - Patterns : superviseur-workers, hierarchical, swarm, collaborative.
    - Système multi-agent veille concurrentielle : agent collecteur web + agent analyste + agent rédacteur.
    - Pipeline de génération de propositions commerciales : agent qualif → agent rédac → agent pricing → agent QA.
    - Workflow data : agent SQL + agent visualisation + agent commentaire, orchestré par LangGraph avec validation humaine.
- Model Context Protocol (MCP) et intégrations entreprise
    - Spécification MCP : serveurs, ressources, outils, prompts — pourquoi c'est devenu le standard.
    - Implémenter un serveur MCP (Python / TypeScript) exposant une base interne.
    - Sécurité MCP : auth, scopes, audit, sandbox d'exécution.
    - Cas d'intégration : SharePoint, Confluence, Jira, Snowflake, GitHub via MCP.
    - Développer un serveur MCP qui expose les tickets Jira de l'entreprise à un agent Claude.
    - Connecter Claude Desktop / Cursor à votre data warehouse Snowflake via MCP avec contrôle d'accès.
    - Serveur MCP custom : exposer le référentiel produit interne à plusieurs agents (vente, support, marketing).
- Sécurité, déploiement et FinOps des solutions IA
    - Prompt injection, jailbreaks, data exfiltration : OWASP Top 10 for LLM Apps 2025.
    - Architecture cloud : API Gateway, caching (prompt cache Claude / OpenAI), batching.
    - FinOps IA : budgets, alerting coût, choix Sonnet vs Opus, caching agressif.
    - Conformité AI Act, RGPD, hébergement souverain (Mistral, Scaleway, OVH).
    - Mettre en place le prompt caching Claude sur un assistant interne : – 70 % de coût observé en pratique.
    - Auditer un agent existant contre les 10 risques OWASP LLM et produire un plan de remédiation.
    - Déployer un RAG en architecture souveraine (Mistral hébergé EU + Qdrant Cloud EU) conforme AI Act.
- Projet de synthèse : RAG + agents en production
    - Conception d'architecture : choix techniques justifiés et chiffrés.
    - Implémentation d'un système combinant RAG + 2 agents minimum + serveur MCP.
    - Mise en place de l'évaluation continue et du monitoring.
    - Présentation et revue de code en pair-programming avec le formateur.
    - Projet « Assistant Achats » : RAG sur les contrats fournisseurs + agent négociation + serveur MCP ERP.
    - Projet « Copilote Support N2 » : RAG sur base de tickets + agent diagnostic + escalade humaine.
    - Projet « Analyste Financier » : RAG sur rapports annuels + agent calcul + agent visualisation.

Organisation de la formation

Equipe pédagogique
Spécialiste de l'IA générative, des agents IA, des architectures RAG, de la data science et de l'automatisation, Daouda accompagne les organisations dans l'adoption de l'intelligence artificielle. Fort d'expériences au CNRS et chez Saint-Gobain, il allie expertise technique, vision stratégique et pédagogie pragmatique. Ses formations, fondées sur des cas réels et la mise en pratique, permettent d'intégrer l'IA de manière concrète, responsable et créatrice de valeur.

Moyens pédagogiques et techniques

- Repo GitHub fourni avec dépôts Docker prêts à l'emploi, accès LangSmith Pro inclus.
- Supports pédagogiques commentés et partagés à distance
- Apports méthodologiques structurés

- Missions d'application intersession sur le cas réel de l'apprenant
- Bibliothèque de templates d'automatisation prêts à l'emploi
- Cas pratiques
- Etudes de cas concrets
- Quiz
- Mise à disposition en ligne de documents supports à la suite de la formation
- Pour les formations en présentiel, des salles de formation agréables, lumineuses, spacieuses, modernes, équipées de connexion très haut débit, de grands écrans, et à proximité immédiate des transports.
- Pré-requis matériels : disposer d'un ordinateur. En cas de formation à distance, disposer également d'une connexion internet.
- Une semaine avant le début de la formation, nous donnons, à chaque participant, un accès sécurisé à notre plate-forme de formation en ligne (LMS : Learning Management System). L'apprenant accède via son extranet au livret d'accueil, au tuto de l'espace apprenant et au règlement intérieur.

Dispositif de suivi de l'exécution de l'évaluation des résultats de la formation

- Suivi des missions d'application réalisées entre les séances
- Feuilles de présence
- Formulaires d'évaluation de la formation
- Mises en situation
- Certificat de réalisation de l'action de formation
- 10 jours avant l'entrée en formation : Nous transmettons à chaque stagiaire un questionnaire de préparation de la formation permettant d'adapter au mieux la formation à ses besoins.
- Pendant la formation : Le contrôle de connaissances et des acquis se font de façon continue via des quiz en ligne avec notation instantanée et des cas pratiques à réaliser. En fin de formation, les stagiaires restituent leurs travaux avec feedback du formateur.
- A l'issue de la formation : Nous adressons aux stagiaires un questionnaire de satisfaction à chaud et d'auto-diagnostic. Nous mesurons ainsi leur satisfaction et constatons directement l'évolution des compétences en comparaison avec les données recueillies avant la formation. Nous adressons un questionnaire d'évaluation au formateur.
- 15 jours après la fin de la formation : Nous adressons un questionnaire d'évaluation au financeur (OPCO).
- Un mois après la formation : Nous adressons un questionnaire d'évaluation aux différentes parties prenantes : commanditaire /manager.
- 3 mois après la formation : Nous transmettons au stagiaire un questionnaire d'évaluation à froid pour évaluer l'impact de la formation dans son activité professionnelle.

Tarif de la formation

À partir de 10500.00 €HT
TVA non applicable, art. 261-4-4 du CGI