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

Mis à jour le 04/05/2026

S'inscrire

Formation DSPy : maitriser les coûts des LLms

3 jours (21 heures)

Présentation

DSPy permet de concevoir des pipelines LLM plus fiables et moins coûteux en remplaçant le prompt “au feeling” par une approche programmable, mesurable et optimisable. Vous apprendrez à réduire la facture tokens et la latence tout en améliorant la qualité sur des cas d’usage comme RAG, extraction structurée et assistants métier.

La formation se concentre sur la maîtrise des coûts : choix de modèles, stratégies de réduction de contexte, cache, routage, et optimisation automatique des prompts/chaînes via les modules DSPy. Chaque notion est reliée à des métriques concrètes (tokens, temps, taux d’erreur) et à des critères de qualité.

Approche 100% pratique : ateliers guidés, démos reproductibles et mini-projet. Livrables : un repo d’exemples DSPy, une grille d’évaluation, et un pipeline optimisé (baseline vs version optimisée) avec rapport coûts/qualité.

Comme toutes nos formations, celle-ci vous présentera **la dernière version stable** de la technologie et ses nouveautés.

Objectifs

- Instrumenter et mesurer coûts, latence et qualité d’un pipeline LLM.
- Construire des modules DSPy (Predict, ChainOfThought, RAG) et les composer.
- Définir des datasets, métriques et tests de non-régression.
- Optimiser automatiquement prompts et paramètres avec teleprompting.
- Mettre en place cache, routage de modèles et réduction de contexte.

Public visé

- Développeurs Python intégrant des LLM en production
- Data scientists / ML engineers
- Tech leads et architectes applicatifs
- Product engineers travaillant sur des assistants/RAG

Pré-requis

- Maîtrise de Python (fonctions, classes, environnements)
- Notions d'API REST et formats JSON
- Compréhension de base des LLM (tokens, contexte, température)
- Notions de Git et ligne de commande

Pré-requis techniques

- Machine avec 16 Go RAM recommandés (8 Go minimum)
- Windows (WSL2 conseillé), macOS ou Linux
- Python 3.10+, venv/poetry, et un éditeur (VS Code/PyCharm)
- Accès à une API LLM (clé fournie par votre organisation) et connexion Internet stable

Programme de notre formation DSPy : maitriser les coûts des LLms

[Jour 1 - Matin]

Fondamentaux DSPy et mesure des coûts LLM

- Comprendre le positionnement de DSPy : programmation déclarative de pipelines LLM et optimisation
- Identifier les postes de coût : tokens (prompt/completion), latence, taux d'erreur, retries
- Mettre en place une instrumentation : comptage tokens, durée, coût estimé par requête et par scénario
- Définir des métriques qualité/coût : exactitude, couverture, hallucinations, coût par succès
- Atelier pratique : Instrumenter un mini-pipeline DSPy et produire un tableau coût/latence/qualité.

[Jour 1 - Après-midi]

Réduire les tokens : signatures, prompts compacts et sorties contraintes

- Concevoir des Signatures DSPy (inputs/outputs) pour limiter l'ambiguïté et la verbosité
- Appliquer des stratégies de compression : consignes courtes, exemples minimaux, suppression du bruit
- Contraindre la sortie : formats stricts (JSON), champs obligatoires, longueur maximale
- Mettre en place des garde-fous : validation de sortie et relance ciblée plutôt que re-prompt complet

- Atelier pratique : Refactorer un prompt “long” en signature DSPy + validation, puis mesurer le gain en tokens.

[Jour 2 - Matin]

Optimisation DSPy : compilation, téléprompting et choix de modèles

- Structurer un programme DSPy : modules, chaînage, séparation extraction/raisonnement/rédaction
- Comprendre la compilation DSPy : objectifs, jeux de données, métriques et contraintes de coût
- Utiliser le teleprompting pour améliorer la qualité à coût constant (ou réduire le coût à qualité constante)
- Stratégies de model routing : petit modèle par défaut, escalade vers un modèle plus grand en cas d’échec
- Atelier pratique : Compiler un module DSPy sur un dataset et comparer coût/qualité avant/après.

[Jour 2 - Après-midi]

Réduire les appels : cache, batching et contrôle des retries

- Mettre en place un cache (clé par signature + paramètres + version) et gérer l’invalidation
- Réduire les appels via batching et regroupement de requêtes homogènes
- Limiter les retries : backoff, seuils, classification des erreurs (transitoires vs logiques)
- Détecter et corriger les boucles coûteuses : relances inutiles, sur-chaînage, prompts redondants
- Atelier pratique : Ajouter cache + politique de retries et mesurer l’impact sur un flux à forte volumétrie.

[Jour 3 - Matin]

Qualité sous contrainte : évaluations, tests et budgets

- Construire un jeu d’évaluation : cas nominaux, cas limites, données “pièges” et critères d’acceptation
- Mettre en place des evals automatisées : scoring, seuils, régressions et suivi dans le temps
- Introduire des budgets : coût max par requête, coût max par session, timeouts et dégradations contrôlées
- Approche “fail fast” : validations intermédiaires pour arrêter tôt les traitements non conformes
- Atelier pratique : Créer une suite d’evals + budget de coût et intégrer un contrôle de régression.

[Jour 3 - Après-midi]

Mise en production : observabilité, gouvernance et optimisation continue

- Mettre en place une observabilité opérationnelle : logs structurés, traces, coûts par endpoint et par client
- Définir des alertes : dérive de coût, hausse de latence, baisse de qualité, hausse de taux d'échec
- Versionner prompts/signatures/compilations : rollbacks, A/B tests, canary releases
- Plan d'optimisation continue : revue hebdo des top coûts, priorisation et quick wins
- Atelier pratique : Déployer un pipeline DSPy "budget-aware" avec dashboard coût/qualité et stratégie de rollback.

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.