

Livrable 1 : Organisation de l'Équipe et Plan de Travail

ENSAM RABAT

28 novembre 2025

Membres de l'Équipe (Trinôme)

Afin de structurer le travail et d'assurer une couverture complète des modules complexes de ce projet, nous proposons la répartition des rôles suivante :

Composition de l'Équipe (Trinôme)

Le projet sera mené par l'équipe suivante, avec une répartition claire des responsabilités pour les différents modules :

TABLE 1 – Répartition des rôles et des responsabilités		
Rôle	Membre	Missions Principales et Modules Cibles
Chef d'Équipe / NLP Lead	Zaynab ER-REGHAY	Coordination générale, gestion du Git, intégration des modules NLP et fusion (Modules 1, 4, 5).
CV Lead / Gabarits Expert	Lahfari BILAL	Prétraitement des images/PDF, développement du Module Gabarits, implémentation et entraînement du modèle CV hybride (Modules 1, 2, 3).
Architecte Système / MLOps	Malek SAMI	Architecture offline, pipeline principal (<code>main.py</code>), interface utilisateur, gestion des dépendances et optimisation des performances (Modules 1, 5, 6).

Plan de Travail Détaillé (Phase 1 & 2)

Le projet sera découpé en deux phases principales pour assurer une livraison progressive et des tests d'intégration réguliers, conformément aux recommandations du sujet¹.

Phase 1 : Infrastructure & Prototypes Simples (Jours 1-10)

L'objectif est de mettre en place l'environnement 100% offline et de disposer d'un prototype fonctionnel pour chaque tâche de classification.

Stratégie de Données Synthétiques

Étant donné l'impossibilité de collecter un jeu de données réelles suffisant et labellisé (CNIE, relevés bancaires, factures, etc.), nous allons adopter une stratégie de **Génération de Données Synthétiques** combinée à des techniques d'Augmentation de Données pour contourner ce blocage.

TABLE 2 – Détail de la Phase 1

Étape	Tâches Clés	Responsable	Livrables de la Phase 1
1.	Config Offline : Création de <code>setup_offline.py</code> (téléchargement ResNet50, CamemBERT, Tesseract)	Architecte / CV Lead	24cmScript <code>setup_offline.py</code>
2.	Définition de la structure de dossiers <code>models/</code>		
3.	Pipeline Basique : Conversion PDF vers images (Module 6, Étape 1)	22cmArchitecte	24cmClasse <code>OfflineModelManager</code>
4.	Implémentation de la classe <code>OfflineModelManager</code> (chargement simple des modèles) ⁶		
5.	Prototype NLP (Motifs) : Définition des dictionnaires de motifs sémantiques pour les 5 classes ⁷	22cmNLP Lead	24cmDictionnaires de motifs (JSON/Python)
6.	Implémentation du score de classification basé uniquement sur les mots-clés (4.2.1) ⁸		
7.	Prototype CV (ResNet) : Entraînement initial simple de ResNet50 (sans gabarits) pour établir une baseline ⁹	CV Lead	Baseline ResNet50 simple.

1. Génération de Données Synthétiques

L'objectif est de créer des documents qui imitent la structure visuelle et le contenu textuel des 5 classes, sans utiliser de vraies données sensibles.

Pour les Relevés/Factures (Structure Tabulaire) :

- Utiliser des bibliothèques de génération de PDF pour créer des tables avec des champs typiques (Montant, Date, Description).
- Remplir les champs avec des données aléatoires mais cohérentes (ex : dates logiques, montants float).
- Intégrer les **Motifs Sémantiques** définis (ex : "solde", "kWh", "m³") dans les textes générés¹⁰.

Intérêt : Fournit un corpus labellisé pour le fine-tuning de CamemBERT et une base d'entraînement pour la détection de structure tabulaire (Transformée de Hough)¹¹.

Pour les Pièces d'Identité (Format et Features Gabarits) :

- Générer des images avec le ratio d'aspect correct (format carte)¹².
- Placer des zones pour la photo, le numéro d'identité, et la date, pour entraîner les détecteurs de zones (ex : Cascade Classifiers pour la photo)¹³.
- Ajouter des images de fond (comme la carte du Maroc) pour simuler les gabarits visuels¹⁴.

2. Augmentation de Données (Clé de la Robustesse)

Nous appliquerons des techniques d'augmentation agressives aux données synthétiques pour simuler les conditions réelles des documents administratifs¹⁵.

- **Augmentations CV** : Rotations légères, changements de contraste, ajout de bruit, compression JPEG pour simuler une mauvaise qualité de scanner/photo¹⁶.
- **Augmentations OCR** : Dégrader artificiellement les images (flou, faible résolution) pour entraîner le pipeline à mieux gérer les erreurs post-OCR¹⁷.

3. Validation

Un petit jeu de données réelles (anonymisées et ne nécessitant pas d'autorisation spéciale) sera recherché pour la **Validation Finale** et l'évaluation de la Robustesse¹⁸.