



**Université Ibn Tofail Faculté des Sciences, Kénitra**

**FICHE DE STAGE**

**Master Intelligence Artificielle et Réalité Virtuelle**



**NLP - RAG SYSTEM**

## 



Préparé par :

LAHMECH Loubna, KHELLOUFI Naoual



Année universitaire 2024/2025

#### 1. **Informations sur l’entreprise**:

**Nom de la société** : 3D SMART FACTORY

**Activité de la société** : Recherche et Développement

**Service concerné par le stage** : Recherche et développement

**Adresse** : Villa N° 75, Lotissement La Gare 28820 MOHAMMEDIA

**Téléphone** : +212523300446

**Email** : [3dsmartfactory@gmail.com](mailto:3dsmartfactory@gmail.com)

**Représentée par** : Thierry BERTIN GARDELLE

**Encadrant professionnel** : M. **Thierry BERTIN GARDELLE**

#### **Encadrant pédagogique:**

**Nom** : M. Anass Nouri

**Poste** : Enseignant-chercheur

**Établissement** : Université Ibn Tofail

### **Email** : anass.nouri@uit.ac.ma

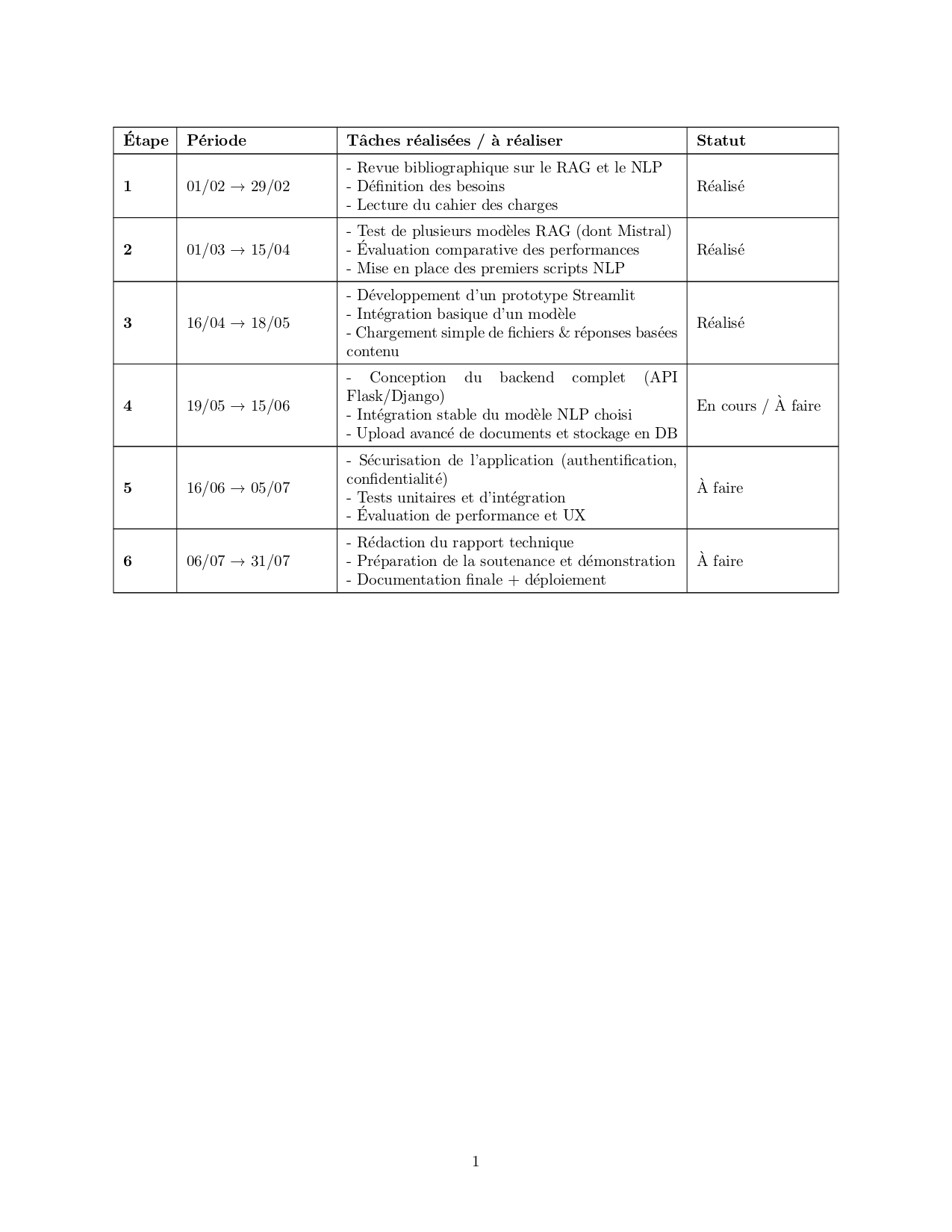
#### **Descriptif du Sujet de Stage:**

Le projet vise la conception et le développement d’un système web intelligent exploitant les techniques de **NLP** et l’architecture **Retrieval-Augmented Generation (RAG)**.  
L’application permet aux utilisateurs d’**uploader des fichiers variés** (PDF, Word, Excel, PowerPoint, CSV, etc.) puis de **poser des questions** sur leur contenu via une interface de **chat dynamique**. Le système extrait automatiquement l'information pertinente à partir des fichiers et génère des réponses contextualisées.

Les principaux objectifs du projet sont :

* Mettre en œuvre un système **d’auto-entraînement de modèle NLP** sur des documents personnalisés.
* Concevoir une **interface utilisateur fluide**, intuitive, avec des barres de dialogue dynamiques.
* Garantir la **sécurité des données**, la **scalabilité** du système, et exploiter **pleinement les ressources GPU** pour des performances optimales.

#### **Travail à Réaliser (Planning Prévisionnel):**



#### **État d’Avancement :**

À ce stade, plusieurs éléments clés ont été mis en place :

* Une interface web fonctionnelle (via Streamlit) permet le téléchargement de fichiers dans différents formats (PDF, TXT, DOCX, CSV...).
* Un module de prétraitement automatique extrait le contenu pertinent de ces fichiers.
* Plusieurs modèles NLP ont été testés, avec une intégration actuelle basée sur Mistral (modèle RAG), bien que des ajustements soient encore nécessaires.
* L’interface de chat permet de poser des questions et d’obtenir des réponses contextualisées à partir des documents importés
* **Problème actuel** :  
  La génération de réponse prend **trop de temps**. Ce ralentissement est dû à une **utilisation sous-optimale du GPU** disponible sur **Google Colab**. Des investigations sont en cours pour **maximiser l’accélération GPU** (gestion des lots, parallélisation, nettoyage mémoire, etc.).

#### **Perspectives et Évolutions Prévisionnelles:**

Les prochaines étapes et évolutions envisagées du projet incluent :

* **Optimisation de la performance du système RAG**, en exploitant mieux le GPU pour réduire le temps de réponse.
* **Implémentation d’un système multi-agents** pour répartir les tâches NLP selon leur nature (extraction, synthèse, recherche…).
* **Création de profils utilisateurs personnalisés** (avec mémoire des préférences, documents récents, etc.).
* **Ajout d’un historique des échanges** pour permettre à l’utilisateur de suivre et reprendre ses discussions précédentes.