Promotion 2025 – 4e année cycle Ingénieur

Majeure Data & Intelligence Artificielle

|  |  |
| --- | --- |

Safran Aircraft Engines Villaroche

*Rond-point René Ravaud, 77550 Moissy-Cramayel*

*Département Navigabilité*

*Tuteur école :* **RODRIGUES Christophe**

*Tuteur administrative entreprise :* **FOURNIER Olivier**

*Tuteur technique entreprise :* **CURTELIN Raphael**



**Stage Data Science Natural Language Processing pour la Navigabilité**



Large Language Models (LLM) : Tout ce qu'il faut savoir*Etudiant /Stagiaire :* **Enge NOUADJE FOTSO** 

Date de soutenance : 12/09/2024

Table des matières

…

[**1 Remerciements** 5](#_gjdgxs)

[**2 Résumé du stage** 6](#_1fob9te)

[**3**](#_3znysh7)  **Présentation de l’entreprise** 6

[3.1 Le groupe SAFRAN 6](#_2et92p0)

[3.2 Quelques sociétés et joint-ventures du groupe Safran 7](#_3dy6vkm)

[3.3 Initiative RSE 8](#_4d34og8)

[3.4 Safran Aircraft Engines 10](#_17dp8vu)

[3.4.1 Implémentation 11](#_26in1rg)

[3.4.2 Moteurs civils 11](#_35nkun2)

[3.4.3 Moteurs militaires 12](#_44sinio)

[3.4.4 Quelques règlementations et principes de navigabilité 12](#_z337ya)

[3.4.5 Le Département Navigabilité 13](#_3j2qqm3)

[**4 Présentation du stage** 14](#_4i7ojhp)

[4.1 Activités principales du Bureau de Navigabilité 14](#_2xcytpi)

[4.2 Projet 15](#_2bn6wsx)

[4.2.1 Sujet 15](#_qsh70q)

[4.2.2 Les objectifs 16](#_1pxezwc)

[4.2.3 Méthodes et outils 17](#_2p2csry)

[**5 Déroulement du stage** 19](#_3o7alnk)

[**6 Bilan du stage** 32](#_2zbgiuw)

Table des figures

[Figure 1 - La dominance de SAFRAN sur le marché 8](#_1t3h5sf)

[Figure 2 - Schéma de la stratégie climat de Safran 9](#_2s8eyo1)

[Figure 3 - Historique de Safran AE 10](#_3rdcrjn)

[Figure 4 – Implémentation des sites SAE dans le monde 11](#_lnxbz9)

[Figure 5 – CFM-56 : best-seller des moteurs SAFRAN AE 11](#_1ksv4uv)

[Figure 6 – Moteur équipant le Rafale : M88 12](#_2jxsxqh)

[Figure 7 – Les directions de Safran Aicraft Engines 13](#_1y810tw)

[Figure 8 – Exemple de commentaire de catégorie [4- ] 14](#_1ci93xb)

[Figure 9 – Extrait du Tableau des commentaires pour la catégorie orthographe 15](#_3whwml4)

[Figure 10 – Exemple de document Word xCAD-CAD 15](#_3as4poj)

[Figure 11 – Exemple d’une page d’instruction « Dev Recipes » 17](#_49x2ik5)

[Figure 12 – Diagramme de Gantt (Pas final) 19](#_147n2zr)

[Figure 13 - Exemple modèle GGUF 20](#_23ckvvd)

[Figure 14 - Quelques résultats sur les modèles testés 20](#_ihv636)

[Figure 15 - Exemple d'évaluation de modèle 21](#_32hioqz)

[Figure 16 - Légende des évaluations 21](#_1hmsyys)

[Figure 17 - Rapport entre performance et coût 22](#_41mghml)

[Figure 18 – Portail PuTTY vers terminal Xanthe 23](#_2grqrue)

[Figure 19 – Terminal Xanthe 23](#_vx1227)

[Figure 20 - Page de téléchargement des fichiers Mistral Instruct v0.2 24](#_3fwokq0)

[Figure 21 -Transfert de fichiers dans Xanthe via Filezilla 24](#_1v1yuxt)

[Figure 22 - Dossiers CFM56 pour l'entrainement 25](#_4f1mdlm)

[Figure 23 – Exemple de Ficher Word CAD dans un dossier de CFM56 pour l'entraînement 25](#_2u6wntf)

[Figure 24 - Excel généré par le code Python 25](#_19c6y18)

[Figure 25 – Exemples de phrases/commentaires inutilisables 26](#_3tbugp1)

[Figure 26 – Exemple de résultat 27](#_28h4qwu)

[Figure 27 - Méthode d'entrainement LoRA pour les commentaires xCAD 28](#_46r0co2)

[Figure 28- Exemple de différents degrés de quantifications 28](#_nmf14n)

[Figure 29 - Fonctionnement de QLoRA comparé à LoRA et sans méthode 29](#_37m2jsg)

[Figure 30 - Commande Tensorboard pout accéder aux infos de l’entrainement 30](#_2lwamvv)

[Figure 31 - Suivi du train/loss au cours de l'entrainement 31](#_111kx3o)

[Figure 32 – Cross-entropy 31](#_3l18frh)

[Figure 33 - Atelier d'assemblage 33](#_206ipza)

**Liste des symboles**

EASA : European Union Aviation Safety Agency

NLP : Natural Language Processing

LLM : Large Language model

IATA : International Air Transport Association

Safran AE : Safran Aircraft Engines

BdN : Bureau de navigabilité

CAD : Changement à la définition

xCAD : CAD ou Changement à la définition normatif (NCAD) ou réparation (RCAD)

DOA : Design Organization Approval

NCAD : Changement à définition normatif

RCAD : Changement à la définition type réparation

GGUF : GPT-Generated Unified Format

DD : Développement Durable

RSE : Responsabilité Sociétale des entreprises

CPU : Central Processing Unit

GPU : Graphics Processing Unit

PEFT : Parameter efficient fine tuning

ESILV : École supérieure d'ingénieurs Léonard-de-Vinci

SNECMA : Société Nationale d’Etude et de Construction de Moteurs d’Aviation

SAGEM : Société d’Applications Générales d’Electricité et de Mécanique

## **1 Remerciements**

## **2 Résumé du stage**

*Mots clés : Navigabilité, Fine-Tuning, Natural Language Processing, LLM, Python, Excel, Entrainement*

Dans le cadre de ma 4e année de cycle ingénieur généraliste majeure Data & Intelligence Artificielle à L'École supérieure d'ingénieurs Léonard-de-Vinci (ESILV), j’ai été amené à effectuer un stage de 20 semaines allant d’avril à août 2024. Intéressé par le domaine de l’aéronautique, j’ai choisi de faire un stage au Bureau de Navigabilité de Safran Aircraft Engines Villaroche portant sur le sujet suivant : Natural Language Processing (NLP) pour le traitement des documents Navigabilité. Ce sujet concerne l’appropriation de documents reçus et traités par la Navigabilité, l’identification et la construction d’une base d’exemples d’apprentissage et d’une base de validation, la sélection de quelques modèles de langage et de méthodes de « Fine-Tuning » pour entraîner les modèles sur la base d’apprentissage et de validation, l'entraînement des modèles et les tests de performances sur la base d’apprentissage et de validation.

Au cours de ce stage, j’ai pu tester différents modèles NLP que ce soit des LLM GGUF avec une i7 ou Regular avec une NVIDIA GPU A100 80GB pour une première identification des modèles utilisables dans le projet. Après avoir sélectionné le modèle le plus adéquat pour un entrainement au vu des ressources disponibles chez Safran AE et des test faits avec les modèles non entraînés, j’ai travaillé sur l’identification des phrases ou mots commentés et de leurs commentaires navigabilité associés pour une utilisation en entraînement de modèle NLP. Avec Python j’ai pu extraire environ 2000 ensembles de phrases ou mots commentés et leurs commentaires associés que j’ai stockés dans Excel pour ensuite les nettoyer et les mettre dans le format adéquat à l'entraînement du modèle. Suite à cela, j’ai entraîné le modèle choisi en manipulant les différents paramètres en essayant ainsi d’avoir la meilleure cohérence possible des sorties.

## **3 Présentation de l’entreprise**

### 3.1 Le groupe SAFRAN

Créé en 2005 par la fusion de SAGEM et SNECMA, Safran est un groupe industriel de premier plan qui opère dans trois domaines clés : l'aéronautique, la défense et la sécurité.

Leader mondial ou européen sur plusieurs marchés, Safran excelle dans :

- La propulsion et les équipements aéronautiques, concevant, produisant, commercialisant, entretenant et vendant des moteurs et des équipements pour les aéronefs civils et militaires.

- La défense, où il fournit des équipements de navigation et de guidage, des systèmes optroniques, des systèmes d'avionique, des systèmes de drones et des équipements aéroterrestres pour les marchés civils, militaires et spatiaux.

### 3.2 Quelques sociétés et joint-ventures du groupe Safran

Le groupe Safran est composé de onze entreprises spécialisées dans différents domaines et jouant un rôle majeur à l'échelle internationale. En 2016, plusieurs entreprises du groupe ont intégré le nom de Safran dans leur dénomination, renforçant ainsi l'image de marque du groupe à travers le monde. Les activités principales du groupe se répartissent en trois domaines :

1. Domaine de la propulsion / Motoristes :

* Safran Aircraft Engines : Fabrication de moteurs pour avions civils et militaires.
* Safran Helicopter Engines : Conception de turbomoteurs pour hélicoptères civils et militaires, groupes auxiliaires de puissance et systèmes propulsifs pour missiles.
* Safran Aero Boosters : Spécialisation dans les compresseurs basse pression et les équipements destinés à l'aéronautique et l'espace.
* Safran Aerosystems : Conception de solutions de haute technologie qui optimisent la performance des aéronefs et la sécurité des vols comme des systèmes d’évacuation d’urgence, des systèmes oxygène pour l'équipage ou des systèmes carburant et fluides.
* Safran Engineering Services : Offre de services d'ingénierie de haute technologie pour les secteurs aéronautique, automobile et ferroviaire.
* Safran Power Units : Production de groupes de puissance auxiliaire et de systèmes de démarrage pour l'aéronautique civile et militaire, ainsi que de turboréacteurs pour missiles et engins cibles.

1. Equipementiers :

* Safran Electrical & Power : Développement de systèmes électriques pour le marché aéronautique, couvrant l'ensemble de la chaîne électrique comprenant la génération, la distribution, la conversion, les câblages, la gestion de la charge et la ventilation.
* Safran Landing Systems : Fabrication de systèmes d'atterrissage et de freinage aéronautiques.
* Safran Nacelles : Conception de structures et de nacelles complètes pour moteurs d'avions.
* Safran Transmission Systems : Fourniture de systèmes de transmission de puissance mécanique pour moteurs d'avions et d'hélicoptères.
* Safran Ceramics : Production de matériaux destinés à la défense, l'aérospatiale, l'aéronautique, la sécurité automobile et l'industrie.
* Safran Aerosystems : Conception de solutions de haute technologie qui optimisent la performance des aéronefs et la sécurité des vols comme des systèmes d’évacuation d’urgence, des systèmes oxygène pour l'équipage ou des systèmes carburant et fluides.
* Safran Seats : Acteur majeur sur le marché international des sièges passagers et techniques.

1. Domaine de la défense et de la sécurité :

- Safran Electronics & Defense : Fourniture d'équipements et de systèmes en optronique, avionique, navigation, électronique et logiciels pour les marchés civils et de défense.

- Safran Optics 1 : Leader mondial dans les systèmes électro-optiques et de navigation aux États-Unis.

Ces différentes entreprises contribuent à la diversité des activités de Safran et à sa présence dans les secteurs de la propulsion, des équipements, de la défense et de la sécurité. La principale activité du groupe reste l’aérospatiale, qui génère environ 80 % de son chiffre d'affaires, avec 53 % provenant des entreprises de propulsion et 27 % des équipementiers aéronautiques.

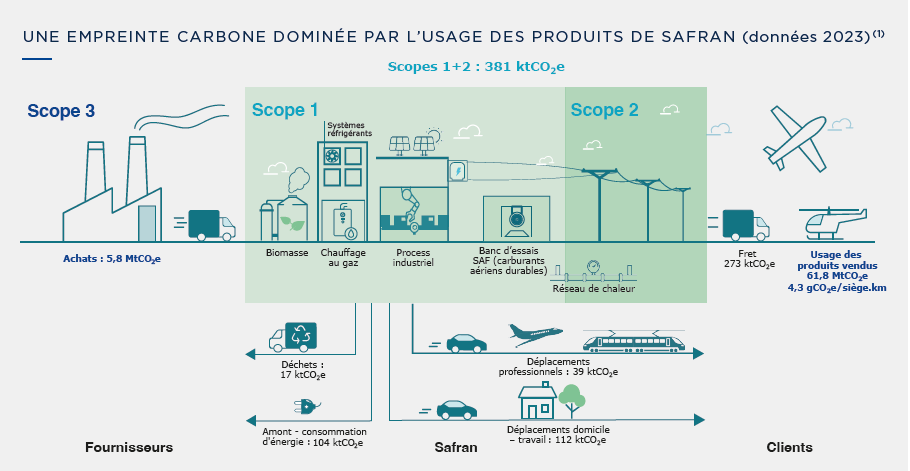
En 2023, Safran compte près de 92 000 collaborateurs répartis sur tous les continents. En France, le groupe possède 75 établissements. Voici quelques chiffres illustrant la domination de Safran sur les marchés mondiaux et européens [1]:



*Figure 1 - La position de SAFRAN sur le marché*

### 3.3 Initiative RSE

Safran s'est engagé à mettre en œuvre une stratégie RSE ambitieuse, visant à décarboner l'aéronautique, à être un employeur exemplaire, à incarner l'industrie responsable et à affirmer son engagement citoyen, pour contribuer à un avenir plus durable et responsable. [2]



*Figure 2 - Schéma de la stratégie climat de Safran*

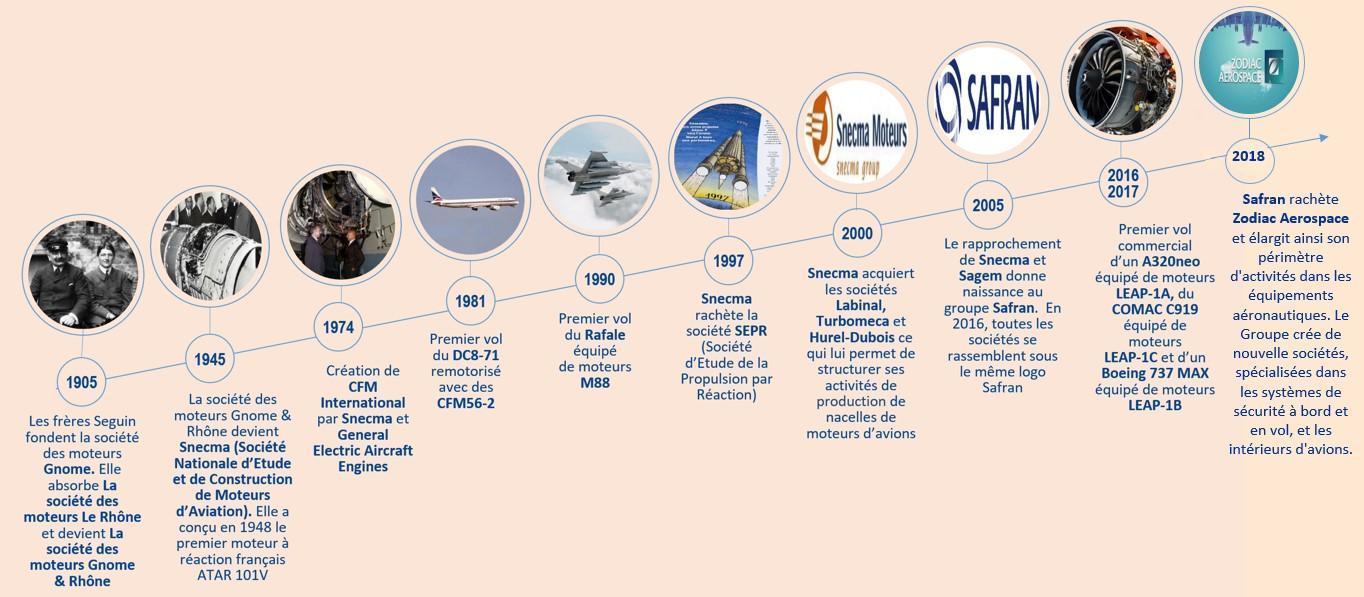
Pour réduire les émissions liées à ses installations et sa consommation d’énergie (scopes 1 & 2), Safran agit sur un ensemble de leviers, entre autres :

* + - Le déploiement de sources de substitution au chauffage par gaz naturel en ayant recours à des réseaux de chaleur, l’électrification de la production de chaleur ou encore la récupération de chaleur fatale.
    - La production et l’autoconsommation d’électricité sur site : en 2023, Safran a inauguré des centrales photovoltaïques, dédiées à l’autoconsommation sur ses sites en France, au Maroc, aux États-Unis et au Royaume-Uni.
    - L’incorporation de carburants durables (à hauteur de 20 % en moyenne) dans le kérosène utilisé pour les essais de réception de moteurs d’avions et d’hélicoptères. Par ailleurs, le plan de sobriété énergétique lancé en 2022 a permis à Safran de réduire la consommation énergétique de ses sites en Europe de 21 % en 2023.
    - Safran agit également pour réduire les émissions indirectes liées à ses activités (scope 3). Safran a poursuivi en 2023 l’objectif d’embarquer ses 400 principaux fournisseurs en termes d'émissions sur une démarche de plans d’actions compatibles avec l’Accord de Paris d’ici 2025 en s'appuyant sur une grille de maturité proche de celle du CDP. Le Groupe applique également le prix interne du carbone dans l’évaluation des offres des fournisseurs.

Ces faits marquants soulignent l'engagement de Safran en matière de durabilité et de responsabilité sociale, témoignant de leurs efforts pour réduire leur empreinte environnementale et contribuer à un secteur de l'aéronautique plus durable.

### 3.4 Safran Aircraft Engines

Safran Aircraft Engines, qui est la plus ancienne société de fabrication de moteurs aéronautiques au monde, a été formée au fil des années par la fusion de plusieurs entreprises leaders dans les secteurs de l'aéronautique, de l'espace et de la défense. Autrefois connue sous le nom de SNECMA, cette société est aujourd'hui une filiale du groupe français Safran. Voici un bref résumé de l'évolution historique de la société depuis sa création en 1905 :



*Figure 3 - Historique de Safran AE*

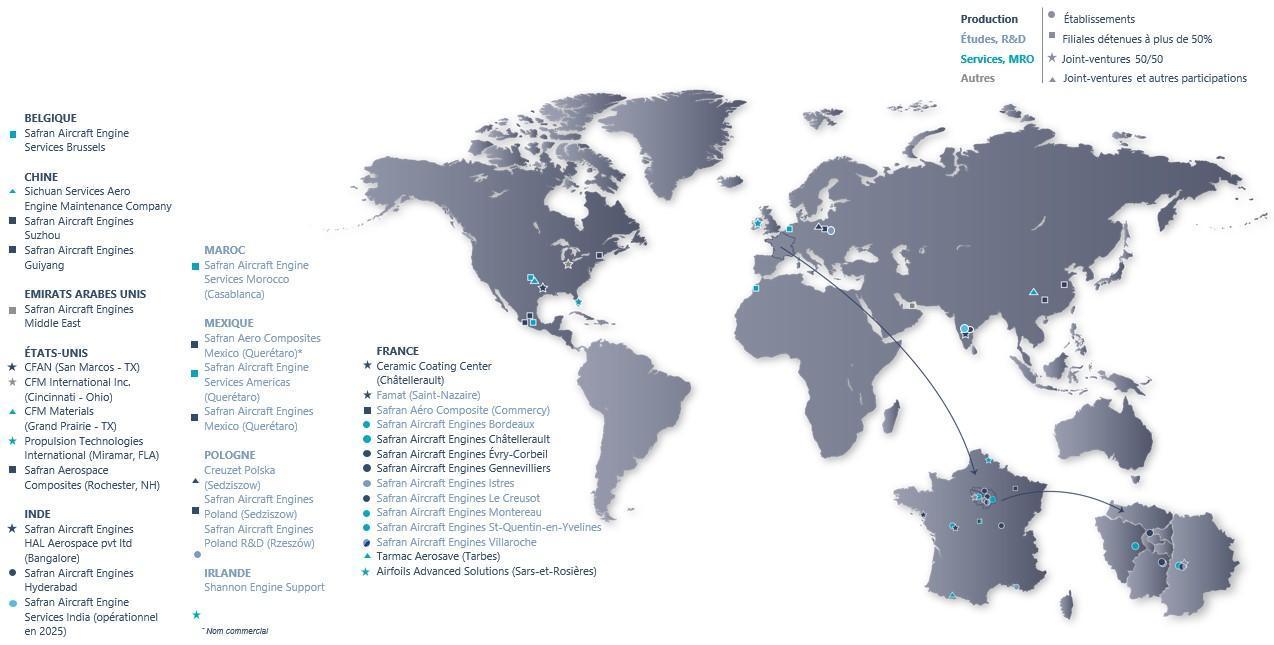
La société conçoit, développe, produit et commercialise des moteurs d'avions militaires et civils. Parmi ses moteurs les plus populaires, on retrouve le CFM-56: « Dans le monde, toutes les 2 secondes un avion équipé de moteurs CFM décolle ». En 2023, le chiffre d'affaires de Safran Aircraft Engines s'élève à 11,9 milliards d'euros.

En plus de la fabrication de moteurs, Safran Aircraft Engines offre une gamme de services aux compagnies aériennes, aux opérateurs d'avions et aux forces armées. Ces services comprennent le support, la maintenance et la réparation des moteurs, ainsi que la gestion de flotte et l'optimisation des opérations. L'objectif de ces services est de garantir la disponibilité opérationnelle maximale des flottes et de fournir un soutien constant à plus de 570 clients et opérateurs dans le monde. Safran Aircraft Engines assure la couverture complète du cycle de vie des moteurs, de leur mise en service à leur retrait, en proposant une gamme complète de produits et de services adaptés à chaque type de moteur.

Safran AE reste à disposition 24h/7j pour répondre aux différents besoins, fournir des conseils, des recommandations de maintenance ou de gestion de flottes et exploiter des données techniques de vol. En outre, l’entreprise propose à ses clients des visites techniques régulières et obligatoires qui peuvent durer plusieurs mois en fonction du type de moteur et du nombre d’heures de vol.

#### 3.4.1 Implentation

Safran AE compte plus de 16 400 salariés dont 13 000 en France repartis sur plus de 30 sites



*Figure 4 – Implentation des sites Safran AE dans le monde*

#### 3.4.2 Moteurs civils

Liste des moteurs et ensembles propulsifs pour avions commerciaux destinés aux compagnies aériennes en coopération essentiellement avec General Electric:

| — Famille CFM56 : A320ceo, B737 classique, A340  — LEAP-1A (A320neo)  — LEAP-1B (B737max)  — LEAP-1C (C919 – Comac)  — Silvercrest (Moteur en développement pour un avion d’affaires)  — SaM146 (Superjet 100 – Sukhoi)  — CF6 (A330 – Airbus)  — GE90 (777 – Boeing)  — GP7200 (A380 - Airbus) |  |
| --- | --- |

*Figure 5 – CFM-56 : best-seller des moteurs SAFRAN AE*

#### 3.4.3 Moteurs militaires

Liste des moteurs d’avions de combat, d’entrainement et de transport pour des clients étatiques comme l’Armée de l’Air et de l’Espace, l’Aéronautique Navale et les clients à l’export :

| — M88 (Rafale – Dassault Aviation)  — TP400 (A400M – Airbus) coopération via EPI  — M53 (Mirage 2000 – Dassault Aviation)  — ATAR (Mirage F1/Super Etendard/Mirage 50 – Dassault Aviation)  — LARZAC (Alphajet – Dassault Aviation)  — Tyne (Transall C160/Atlantic 1/Atlantique 2) |  |
| --- | --- |

*Figure 6 – Moteur équipant le Rafale : M88*

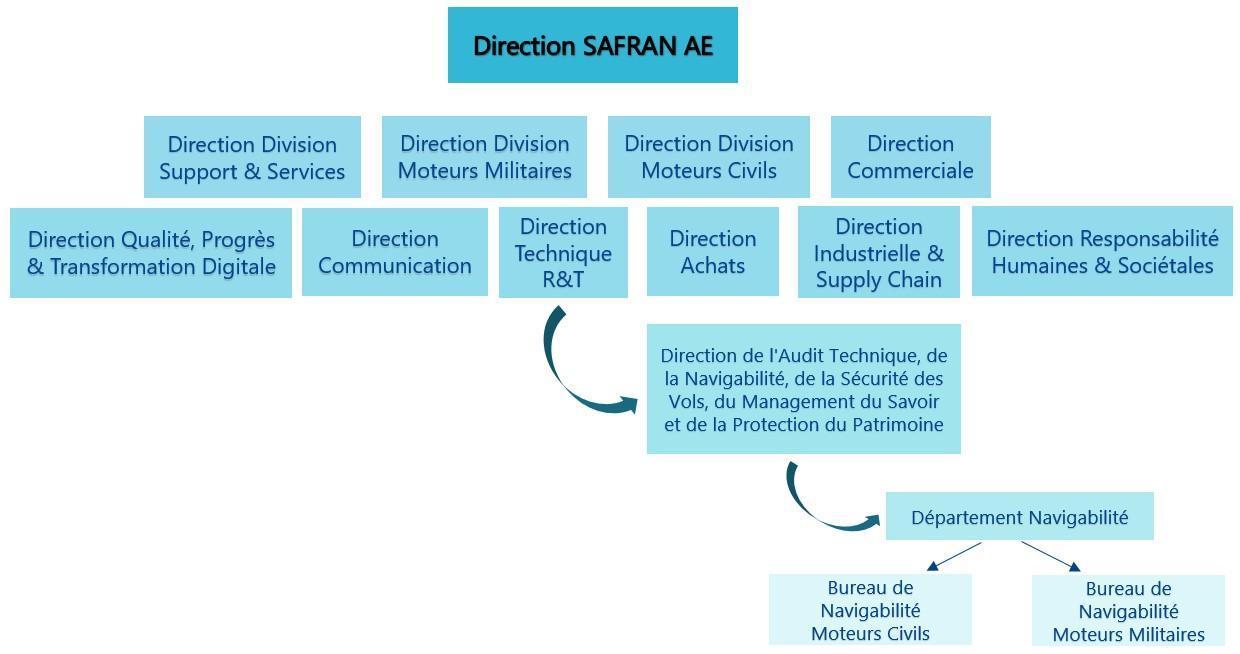
#### 3.4.4 Quelques règlementations et principes de navigabilité

#### 3.4.5 Le Département Navigabilité

Safran Aircraft Engines est une entreprise qui comprend plusieurs directions, dont la Direction Technique. Mon stage s'est déroulé au sein du Département Navigabilité, qui compte environ une vingtaine de personnes. Ce département est rattaché à la Direction de l'Audit Technique, Navigabilité, Management du Savoir et Protection du Patrimoine, basée sur le site de Villaroche, elle-même rattachée à la Direction Technique.

En tant que "Stagiaire Data Scientist - Natural Language Processing", j'ai été affecté à des missions spécifiques au sein du département Navigabilité. Les détails de ces missions sont présentés plus loin au chapitre 5 du rapport.

Le Département Navigabilité joue un rôle crucial pour garantir la navigabilité des moteurs tant pour les moteurs civils que militaires. Il travaille en étroite collaboration avec d'autres départements et directions au sein de Safran Aircraft Engines pour assurer que les moteurs remplissent les exigences de navigabilité et de sécurité en réalisant au nom des autorités de sécurité aérienne (comme l'EASA), des missions déléguées pour la certification, le maintien de la certification et le suivi de la navigabilité des moteurs d'avions en service. Il gère également les instructions pour le maintien de la navigabilité, telles que les manuels et les bulletins de service après-vente. Le département Navigabilité joue aussi un rôle de formation réglementaire des personnels habilités pour exercer des tâches de vérification indépendante des documents de navigabilité.



*Figure 7 – Les directions de Safran Aicraft Engines*

## **4 Présentation du stage**

### 4.1 Activités principales du Bureau de Navigabilité

### 4.2 Projet

Dans le cadre de mon stage, j'avais pour objectif de réaliser

#### 4.2.1 Les objectifs

Les objectifs principaux du projet comprenaient les points suivants :

**1. La compréhension des documents** :

**2. La collecte de dossiers** :

**3. La préparation des données** :

**4. La sélection de modèles de NLP** :

**5. L’entraînement et test des modèles** :

**6. Conclure sur la faisabilité du POC**:

Tous ces objectifs sont à réaliser en respectant les règles de sécurité de Safran AE afin de prévenir toute attaque ou fuite de données.

Pour garantir la sécurité des opérations, Safran AE utilise un gestionnaire de paquets logiciels appelé Artifactory. Ce gestionnaire permet de télécharger de manière sécurisée les dépendances logicielles nécessaires au développement en Python, Java, .NET, Javascript, Docker, Linux, et plus encore. Cela garantit que seuls les composants logiciels approuvés et sécurisés sont utilisés dans le développement du projet.

De plus, une conférence sur la sécurité informatique et des formations en ligne à la sécurité des données lors de l’utilisation de la « Generative AI » a été mis à disposition à travers le site CoorpAcademy pour les nouveaux arrivants.

J'ai participé à des réunions régulières avec des membres du département navigabilité de SAE (YCN) et du département Data Science (YOD). Ces réunions nous ont permis de discuter de l'avancement du stage, des problèmes rencontrés et des actions à entreprendre par la suite tout en respectant les règles de sécurité.

#### 4.2.2 Méthodes et outils

Pendant toute la durée de mon stage, j'ai utilisé plusieurs outils et méthodes :

* + Suite Office 365 : J'ai utilisé Excel, Word, PowerPoint, Outlook et Teams pour la gestion des tâches et la communication au quotidien.
  + Environnement de développement Python : J'ai installé et utilisé en interne Python 3.11 ainsi que l'IDE PyCharm 2023.3.3 comme environnement de développement pour écrire mes codes Python.
  + Bibliothèques Python : J'ai employé différentes bibliothèques telles que datasets, transformers, peft, trl, torch, pandas, os, re, csv et win32com.client pour des tâches de manipulation de données, de traitement du langage naturel et de gestion de fichiers.
  + HuggingFace : J'ai utilisé la bibliothèque HuggingFace pour télécharger et tester divers modèles de traitement du langage naturel (NLP).
  + PuTTY :
  + Supercalculateur Xanthe
  + Outils de développement et de gestion de fichiers : J'ai utilisé Venv pour créer des environnements virtuels, Git/Gitlab pour la gestion du code source et Filezilla pour le transfert de gros fichiers entre Xanthe et le PC local de Safran.

Ces diverses méthodes et outils ont été essentiels pour accomplir les objectifs de mon stage et pour collaborer efficacement avec les différentes équipes.

Par ailleurs mon tuteur de stage m’a permis de participer à des réunions avec des professionnels du département navigabilité, pour obtenir des informations supplémentaires sur le « workflow » des documents xCAD

## **5 Déroulement du stage**

### 5.1 Prompt engineering avec llama-cpp

ent.

### 5.2. Préparation à l'entraînement sur Xanthe

### 5.3 Réglage fin (Fine-Tuning)

L’entrainement de Mistral 7B Instruct s’est fait avec la méthode de réglage fin (Fine-Tuning) QLoRA (Quantized Low Rank Adaptation) de la bibliothèque PEFT (Parameter efficient Finetuning) s'appuyant sur les bases posées par LoRA (Low Rank Adaptation).

L’Adaptation en Bas-Rang (Low-Rank Adaptation) :

Plutôt que d’entrainer directement des matrices de poids de grande taille, LoRA décompose ces matrices en deux matrices de rang plus bas (matrices A et B). La multiplication de ces deux matrices donne une approximation des ajustements nécessaires.

Pendant l'entraînement, seules les matrices de rang inférieur sont ajustées, les poids originaux du modèle restent inchangés. Les matrices A et B sont ajoutées aux poids d'origine pour ajuster les résultats du modèle sans altérer les paramètres initiaux.

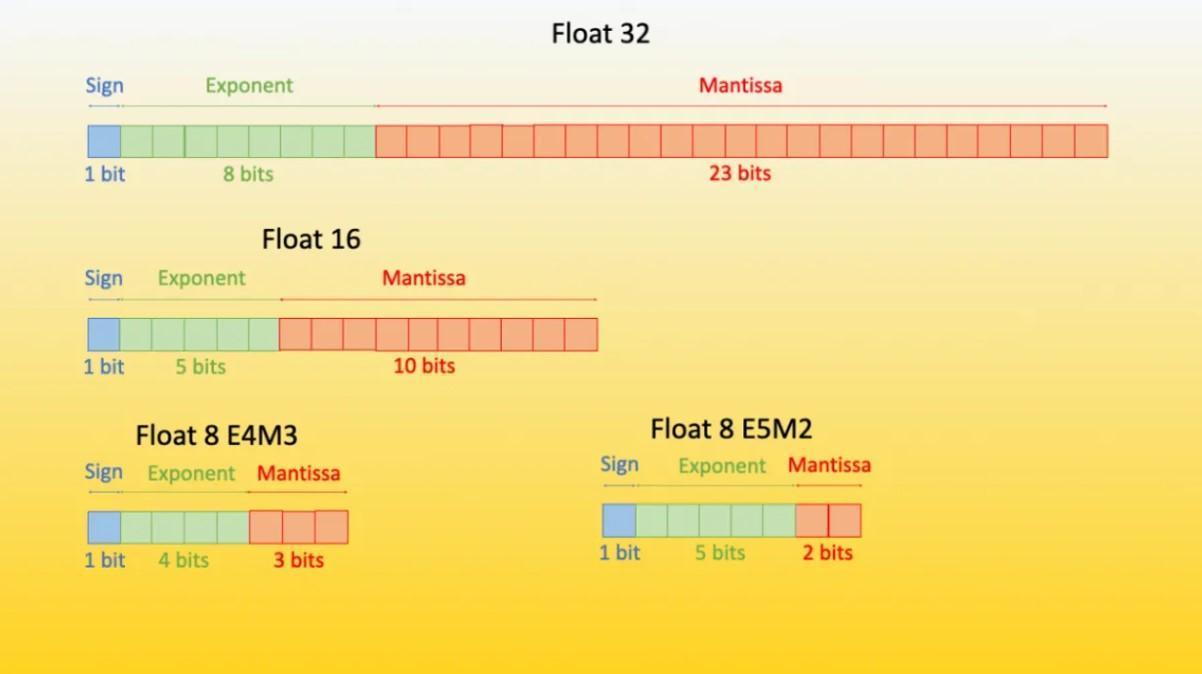
Comme A et B sont de rang inférieur (beaucoup plus petits que les matrices originales), cela réduit considérablement la quantité de mémoire nécessaire pour l'entraînement.

L’Adaptation en Bas-Rang Quantifié (Quantized Low-Rank Adaptation) :

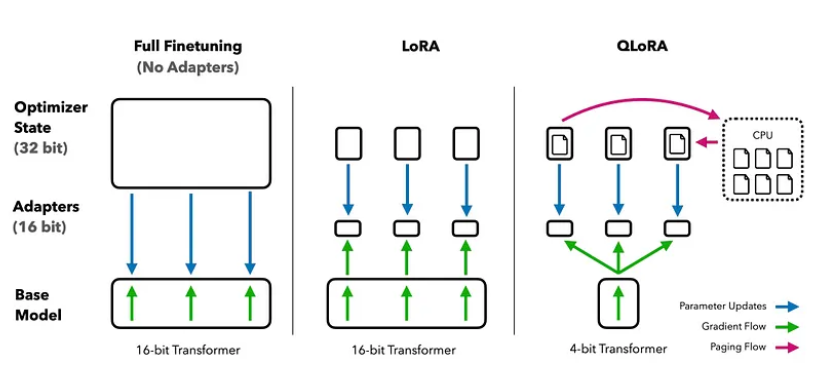
La méthode utilisée lors de l'entraînement est QLoRA qui est similaire à LoRA mais utilisant en plus la quantification (normalisée avec zéro en valeur centrale) et des optimiseurs paginés, qui gèrent les pics de mémoire en déchargeant des parties de l'état de l'optimiseur (State Optimizer) vers la mémoire du CPU si nécessaire.

La quantification consiste à réduire la précision des poids du modèle, généralement en passant de la précision en 32 bits (float32) à une précision plus faible, comme 4 bits (nf4). Cette réduction permet d'économiser de la mémoire et d'accélérer les calculs sans sacrifier de manière significative la performance du modèle.

En utilisant des techniques avancées de quantification, il est possible de maintenir une précision élevée tout en réduisant les besoins en mémoire.



*Figure 28- Exemple de différents degrés de quantifications [6]*



*Figure 29 - Fonctionnement de QLoRA comparé à LoRA et au « Full Finetuning » [4]*

Définitions :

L’« Optimizer State » fait référence à l'ensemble des variables que l'optimiseur maintient en mémoire et met à jour lors du processus d'optimisation. Lors du réglage fin des LLMs, l'optimiseur est responsable de mettre à jour les poids du modèle afin de minimiser la fonction de perte (loss function).

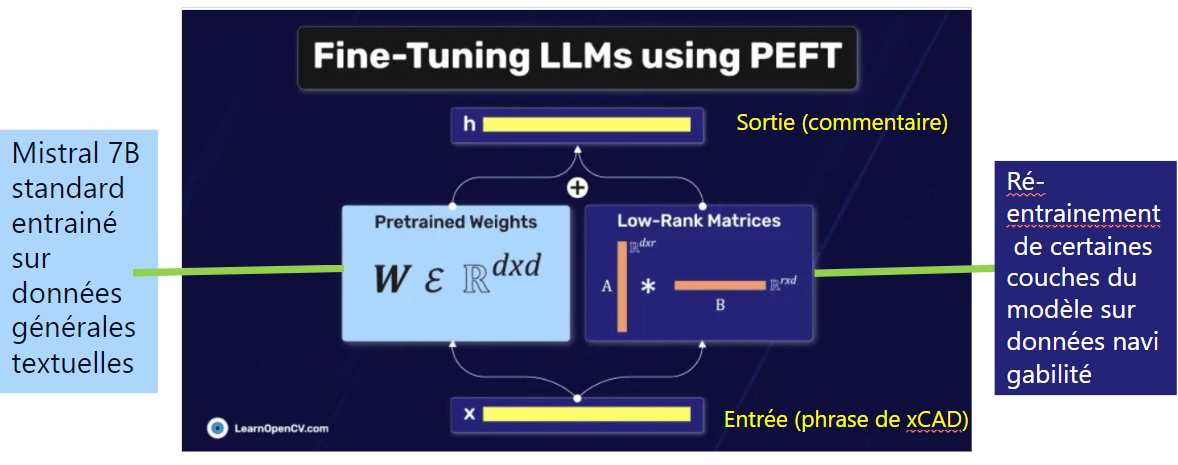
Les « Adapters » sont les seules parties du modèle mises à jour lors du réglage fin. Le modèle de base reste figé. Au lieu de « fine-tuner » tous les paramètres du modèle, les adaptateurs introduisent des couches spécifiques à la tâche qui sont insérés dans l'architecture existante du modèle pré-entraîné.

Les trois méthodes :

Pour le « Full finetuning », on utilise des blocs de mémoire de 32 bits pour stocker l' « Optimizer State » en pleine précision, notamment les gradients. Contrairement aux techniques utilisant des « adapters », ici, tous les paramètres du modèle sont mis à jour directement, sans recourir à des composants supplémentaires. Le modèle de base, quant à lui, conserve ses poids en précision 16 bits. Cette méthode traditionnelle de réglage fin implique la mise à jour exhaustive de tous les paramètres du modèle, nécessitant ainsi des ressources mémoire importantes en raison de la taille considérable de l'état de l'optimiseur.

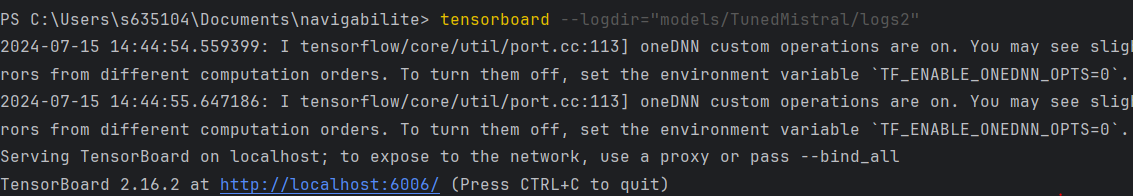
Pour le réglage fin LoRA, l'« Optimizer State » se compose de plusieurs petits blocs, indiquant que seuls les paramètres des « adapters » sont optimisés, ce qui réduit considérablement la mémoire nécessaire par rapport au réglage fin complet. Les « adapters », stockés en précision 16 bits, sont les seules parties mises à jour durant le réglage fin, tandis que le modèle de base reste figé en précision 16 bits. LoRA améliore l'efficacité en mémoire en mettant à jour uniquement un ensemble restreint de poids d’« adapters » plutôt que l'ensemble du modèle, réduisant ainsi la mémoire requise pour l'état de l'optimiseur.

Le réglage fin QLoRA suit une approche similaire à LoRA, mais avec des optimisations supplémentaires. L'« Optimizer State » reste composé de plusieurs petits blocs de 32 bits, avec l'ajout de mémoire paginée, illustrée dans la Figure 29 par des flèches roses pointant vers le CPU, permettant de stocker temporairement des parties de l’« Optimizer State » pour mieux gérer les pics de mémoire pendant l'entraînement. Les « adapters » sont mis à jour de la même manière qu'en LoRA, en précision 16 bits. Le modèle de base est quantifié en précision 4 bits, ce qui réduit significativement son empreinte mémoire, optimisant ainsi les ressources nécessaires pour l'entraînement.

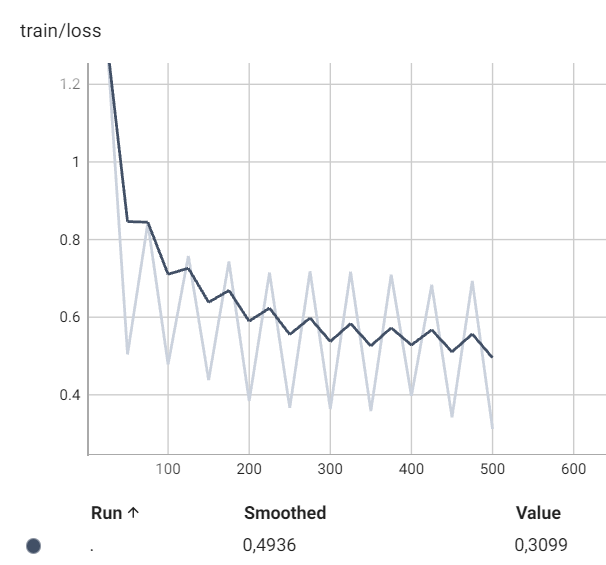


*Figure 27 - Méthode d'entraînement QLoRA pour les commentaires xCAD*

Après le réglage fin nous avons utiliseé l’outil Tensorboard de la bibliothèque Tensorflow permettant de fournir les mesures et les visualisations nécessaires au cours du « workflow » d'apprentissage automatique à partir du fichier « logs » généré à la fin de l'entraînement qui donne la valeur de la fonction de perte au cours de l'entraînement. Il permet de suivre les métriques d'expérience telles que la perte et la précision et de visualiser le graphique du modèle. On peut considérer que le modèle a convergé lorsque le « train/loss » atteint une asymptote.



*Figure 30 - Commande Tensorboard pout accéder aux infos de l’entrainement*



*Figure 31 - Suivi du train/loss au cours de l'entraînement*

Voici ci-dessus le graphe obtenu pour l'entraînement d’environ 2000 exemples sous forme de prompts d’apprentissage, l’évaluation de « train loss » se fait en calculant la moyenne de la perte toutes les 25 étapes par « Cross-entropy » qui est une méthode quantifiant ici la différence entre les distributions de probabilité d’apparition des mots en sortie du modèle et les mots réels des données d'entraînement.

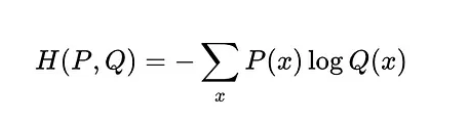


Figure 32 – Cross-entropy

Ainsi chaque 25 étape on obtient le valeur **(H)** de « Cross-entropy loss» en comparant la distribution de probabilité des données réels **(P)** à la distribution de probabilité obtenue par le modèle **(Q)**. [5]

L'entraînement s’est fait sur 1 « epoch » (un seul passage de chaque donnée pour l'entraînement) de 500 « steps ». Un niveau trop élevé « d’epoch » pourrait entraîner un surentrainement sur les données rentrées (Conseil du secteur Data Science). Le « batch size » (nombre de lignes d’exemple que l’on rentre chaque étape) étant de 4. La courbe bleue foncé obtenue précédemment est « smoothed » (lissé) ce qui nous permet de visualiser la tendance de la courbe brute en bleu clair [4]

Les résultats obtenus :

| **Phrase à commenter :** |
| --- |
|  |
| **Réponse du modèle standard non entrainé :** |
|  |
| **Réponse du modèle entraîné** |
|  |

Sur cet exemple on tire 2 enseignements :

* + Le modèle standard (non entrainé) noie le commentaire utile « fissures » dans une réponse trop longue et hors sujet (colle du BA veut dire colle de Bord d’Attaque en aéronautique et ne désigne pas une epoxy).
  + Le modèle entraîné répond le strict nécessaire et s’arrête. Il emploie des mots clés « Mots Commentés », « Commentaire effectué » qui seraient utilisables pour un post-traitement Python avec la bibliothèque Openpyxl pour réinjecter le commentaire dans le fichier word d’origine.

Le « Fine-Tuning » est donc prometteur mais il reste à savoir sur un ensemble de cas si les réponses du modèle sont pertinentes en :

* + Ré-entraînant avec des données nettoyées et améliorée
  + Scorant un nombre significatif (>15) de sorties obtenues avec la base d’entrainement
  + Scorant un nombre significatif (>15) de sorties obtenues avec la base de test

### 5.4 Autres activités

En plus du projet navigabilité, durant mon stage j’ai eu l’opportunité de suivre plusieurs formations enrichissantes et variées. Grâce à CoorpAcademy, j’ai pu accéder à des cours en ligne portant sur l’IA et divers autres sujets, mis à disposition pour les employés de Safran. J’ai également assisté à une conférence sur la sécurité informatique, ce qui m’a permis de mieux comprendre les enjeux et les bonnes pratiques dans ce domaine critique. De plus, j’ai participé aux présentations mensuelles de projets Data Science organisées par le groupe de travail Machine Learning de Safran. Ces sessions étaient l’occasion d’explorer de nouvelles techniques et avancées en Data & IA. Une introduction à Amazon Bedrock et Sagemaker par un expert d’Amazon a également été très instructive, ajoutant une dimension pratique à mes connaissances théoriques. Enfin, la visite du site d’assemblage de SAE Villaroche, où j’ai pu observer différents ateliers de fabrication des pièces de moteurs, y compris celles traitées dans les documents xCAD-CAD, a été une expérience particulièrement marquante et concrète.



*Figure 33 - Atelier d'assemblage*

## **6 Bilan du stage**

Durant mon stage, j'ai eu l'opportunité de développer diverses compétences techniques et professionnelles. J'ai acquis une expertise en data scrapping, en utilisant Python pour automatiser l’extraction de l’ensemble (chemin du fichier source, nom du fichier, date de création, phrase commentée, mots commenté, commentaire) de fichiers Word. En matière de data cleaning, j'ai maîtrisé l'utilisation de la bibliothèque Pandas et identifié des patterns pour le nettoyage et le traitement des données. L'utilisation du système de versionnage Git et la consultation de ressources sur Stackoverflow ont également été essentielles.

De plus, j'ai exploré les librairies et plateformes liées aux modèles de langage, comme Huggingface et Transformers, renforçant ainsi mes compétences en intelligence artificielle. Parallèlement, j'ai amélioré mes compétences d'adaptation, d'organisation, de motivation et d'autonomie, ainsi que mes compétences en communication, en interagissant avec divers professionnels spécialisés en Data Science ou en Navigabilité.

Pendant ce stage, j'ai aussi appris à utiliser Excel de manière avancée, notamment pour la gestion des données, la création de tableaux et les macros. J'ai acquis des compétences spécifiques à l'apprentissage automatique et aux modèles de langage, incluant les « TrainingArguments », les « Checkpoints », llama\_cpp, les modèles NLP GGUF et le Prompt engineering. J'ai également acquis des connaissances concernant Parameter Efficient Finetuning (PEFT), Low Rank Adaptation (LoRA), et la quantification. J'ai aussi utilisé l’outil Tensorboard pour le suivi et l'analyse des performances des modèles lors de l'entraînement.

Les compétences acquises pendant ce stage sont en grande partie en lien direct avec mon parcours académique, notamment en data scrapping, data cleaning, utilisation de Git, exploration des Transformers, et entraînement de modèles de NLP. Ces compétences renforcent mon profil et enrichissent mon expérience pratique dans le domaine de la science des données. Cependant, certaines compétences restent à développer pour aligner complètement mon profil avec les exigences de mon parcours. Il s'agit principalement des compétences en développement logiciel, gestion des API, et déploiement de solutions.

Un des principaux problèmes rencontrés au début du stage était le manque de ressources informatiques adéquates. Les restrictions sur les droits d'installation de logiciels sur les PC de Safran, comme CUDA pour l'utilisation de la carte graphique, ont été un frein important. Les spécifications limitées des anciens PC Dell, tels que les processeurs i5, ont également allongé les temps d'inférence pour les modèles NLP GGUF. Suite à une réunion avec le département Data Science, j'ai découvert Xanthe, un système qui a résolu plusieurs de nos problèmes liés aux ressources informatiques limitées. Après des tests concluants, nous avons pu utiliser Xanthe efficacement pour l'entraînement de modèles NLP Regular.

Les limitations de temps et de coût lors de l'entraînement et des inférences ont été atténuées par l'utilisation de techniques de quantification, permettant une optimisation des ressources disponibles et ainsi une limitation de la consommation d’énergie.

Sur le plan des compétences relationnelles, j'ai pris contact avec le département Data Science, présenté mon travail en cours et mes problèmes à plusieurs professionnels Data Science, et j'ai présenté mes résultats à la direction de l'audit, au pôle navigabilité moteurs civils et au pôle navigabilité moteurs militaires. En termes de compétences techniques, j'ai réussi à entraîner un modèle LLM dans le département navigabilité avec accès à Xanthe, obtenant quelques résultats prometteurs, et j'ai pu faire du nettoyage de données adapté aux objectifs.

Pour finir ce stage a été une expérience très enrichissante, me permettant de renforcer mes compétences techniques et d'apprendre à naviguer dans un environnement professionnel complexe. Les défis rencontrés ont été des opportunités d'apprentissage qui m'ont préparé à des situations similaires à l'avenir.

Les résultats obtenus en conclusion de ce POC confortent l’idée d’une utilisation de modèles NLP afin de traiter les documents Word xCAD en considérant une standardisation des commentaires par les membres du BdN. Il est également recommandé de poursuivre l'exploration du POC NLP avec une seconde étude portant spécifiquement sur les commentaires de cohérence de navigabilité [4-X].

**Sources**

[1] : page web Safran : <https://www.safran-group.com/fr>

[2] : Rapport annuel de Safran 2023 : [SAFRAN RAPPORT INTÉGRÉ 2023 (safran-group.com)](https://web.safran-group.com/ri2023/#page=18)

[3] : <https://medium.com/@ayyucedemirbas/qlora-4444944c20bd>

[4] : <https://medium.com/@pouyahallaj/how-to-tame-noisy-training-loss-in-deep-learning-strategies-and-tips-c68213c6b6b9>

[5] : <https://medium.com/ai-assimilating-intelligence/cross-entropy-in-large-language-models-llms-4f1c842b5fca>

[6] : [Transformer的4BitQLoRA（4-bit base model + LoRA）量化模型越来越方便了 - 知乎 (zhihu.com)](https://zhuanlan.zhihu.com/p/638814235)

**ANNEXE 1 : Principaux Codes**

**Code inférence de modèles GGUF et création de fichier Excel (P1P2P3P4.py)**

**Code inference de modèles GGUF et creation de fichier Excel**

**(P2bisP3bis.py)**

**Code « Data Scrapping » des fichiers Word CFM56 et LEAP (data\_scrapping.py)**

**Code « Data Cleaning » des fichiers de données commentaires (data\_cleaning.py)**

**Code Python Entrainement PEFT x Lora de Mistral 7B Instruct (finetuning.py)**

**Code Python Inférence PEFT x Lora de Mistral 7B Instruct (inférence.py)**

**ANNEXE 2 : Prompts Testés**

**Type de prompt PX avec X = 1, 2, 3, 4, 2bis, 3bis**

**Annexe 3 : Résultats complets des tests de modèles standards non entrainés Mistral, Llama 3 et Secured ChatGPT (Temps en s)**

**Llama3 8B GGUF :**

**Mistral 7B GGUF :**

**Secured ChatGPT :**