



RAPPORT DE STAGE 2A

ANTONIN LARVOR

PROMOTION 2025
DU 22 AVRIL AU 22 AOÛT 2024

RAG : Retrieval Augmented Generation

Tuteur de Stage :
Filip GINTER

Professeur encadrant :
Cyril PRISSETTE



Engagement de non plagiat



Engagement de non plagiat.

Je soussigné,Antonin LARVOR.....

N° carte d'étudiant : 22205754.....

Déclare avoir pris connaissance de la charte des examens et notamment du paragraphe spécifique au plagiat.

Je suis pleinement conscient(e) que la copie intégrale sans citation ni référence de documents ou d'une partie de document publiés sous quelques formes que ce soit (ouvrages, publications, rapports d'étudiants, internet, etc....) est un plagiat et constitue une violation des droits d'auteur ainsi qu'une fraude caractérisée.

En conséquence, je m'engage à citer toutes les sources que j'ai utilisées pour produire et écrire ce document.

Fait le ..10/06/2024

Signature(s)



Ce document doit être inséré en première page de tous les rapports, dossiers et/ou mémoires.

Document du chapitre 10 annexe 5, issu de la Charte des examens adoptée en Conseil d'Administration le 11 juillet 2013 après avis du CEVU du 27 juin 2013 - Délibération N°2013-73 – Modifié suite au CFVU du 12/03/2015.



Remerciements

Je tiens tout particulièrement à remercier Filip Ginter, d'abord pour avoir accepté de m'accueillir au sein de TurkuNLP, et ensuite pour avoir été un formidable tuteur qui m'a accordé toute la confiance nécessaire pour mener à bien ce stage.

Je veux aussi remercier Cyril Prissette d'avoir accepté d'être mon professeur référent à Seatech.

Merci à Cassandra Leddins, ma partenaire, et amie, française au sein du laboratoire, de m'avoir si bien accueilli et d'avoir partagé quelques pauses café avec moi. Je veux également remercier mes super collègues, Ilari, Jenna, Joonatan, Li-Hsin, Maryam, Otto, Parisa, Risto, Siiri, Veronika et tous les autres, d'avoir été aussi accueillants et de m'avoir donné plein de conseils, autant au travail qu'en dehors, pour découvrir des endroits en Finlande ou m'accommorder à la culture finlandaise par exemple.

Merci à Pierre-Thomas, Théotime, et tous les super amis que je me suis faits lors de ce séjour. Sans vous, ce stage aurait été beaucoup moins animé.

Merci à ma famille pour avoir toujours été là pour moi.



Résumé

Durant ces 4 mois de stage, j'ai participé au sous-projet de Retrieval Augmented Generation (RAG) dans le cadre du projet Smart European Shipbuild (SEUS). J'ai effectué toute la programmation du projet dont l'objectif était de créer une interface permettant à un utilisateur de faire des recherches avancées dans une base de données personnelle au travers d'un chatbot. Les différentes techniques mises en oeuvre appartiennent au domaine du Natural Language Processing (NLP).

Abstract

During these 4 months of internship, I participated in the Retrieval Augmented Generation (RAG) sub-project as part of the Smart European Shipbuild (SEUS) project. I carried out all the programming for the project, which aimed to create an interface enabling a user to perform advanced searches in a personal database through a chatbot. The various techniques implemented are part of the Natural Language Processing (NLP) field.

Mots clés

Français :

- Intelligence Artificielle
- Traitement du Langage Naturel
- Modèle d'Embedding
- Modèle de Langage à Large Échelle
- Retrieval Augmented Generation

Anglais :

- Artificial Intelligence
- Natural Language Processing
- Embedding model
- Large Language Model
- Retrieval Augmented Generation



Abréviations

CNN	Réseaux de Neurones Convolutifs (Convolutional Neural Network)
CSC	Centre de Technologie et de l'Information pour la Science
GPT	Generative Pre-trained Transformer
IA	Intelligence Artificielle
LLM	Large Language Model (Modèle de Langage à Grande Échelle)
MMR	Maximal Marginal Relevance (Pertinence Marginale Maximale)
NLP	Natural Language Processing (Traitement du Langage Naturel)
OCR	Optical Character Recognition (Reconnaissance Optique de Caractères)
RAG	Retrieval Augmented Generation
RNN	Réseaux de Neurones Récursifs (Recurrent Neural Network)
SEUS	Smart European Shipbuild



Table des figures

1	Le Transformer - architecture du modèle	11
2	Scaled Dot-Product Attention	12
3	Basic RAG pipeline	16
4	Data Indexing	17
5	Data Retrieval and Generation	18
6	Première version de l'application web	19
7	Deuxième version de l'application web	20
9	Photos du workshop	29
10	Filip Ginter présentant le système de RAG	29
11	Schéma expliquant le principe du MRAG comparé au RAG classique	30
12	Où est Charlie ? <i>Version Antonin Larvor</i>	31
13	Recherche d'une formule dans la base de données SEUS	34
14	Recherches dans la base de données d'Archéologie	35
15	Recherches dans la base de données locale	36



Sommaire

1	Introduction	7
2	Présentation de TurkuNLP	8
3	Le stage	9
3.1	Introduction au Traitement du Langage Naturel (NLP)	9
3.1.1	Embeddings	9
3.1.2	Transformer et Modèles de Langage à Grande Échelle (LLM)	10
3.2	Conception d'un système de Retrieval Augmented Generation	15
3.2.1	Architecture du système	16
3.2.2	Mise en oeuvre du système	18
3.2.3	Gestion des données avec ChromaDB	18
3.2.4	Création de l'application web	19
3.2.5	Résultats	22
3.2.6	Perspectives	23
3.3	Connaissances générales	25
3.3.1	Les Superordinateurs	25
3.3.2	Acquérir des connaissances	26
3.3.3	Programmation	26
3.4	Autres expériences	27
3.4.1	Conditions de travail	27
3.4.2	SEUS Workshop	28
3.4.3	Journal Club	29
3.4.4	Iron Age Seminar	30
3.4.5	Summer Days	30
4	Conclusion	32
5	Bibliographie	33
6	Annexes	34



1 Introduction

Le stage que j'ai effectué au sein du laboratoire TurkuNLP de l'Université de Turku s'inscrit dans un contexte technologique marqué par un essor sans précédent de l'intelligence artificielle (IA). Ces dernières années, l'IA a connu un développement spectaculaire, notamment grâce aux avancées dans le domaine du Traitement du Langage Naturel (NLP). Cette discipline, qui se concentre sur les interactions entre les ordinateurs et le langage humain, bénéficie de progrès technologiques tels que les modèles de langage à grande échelle (LLM) et les techniques d'embeddings. Ces innovations ouvrent de nouvelles possibilités pour la création de systèmes capables de comprendre, analyser et générer du langage naturel de manière autonome.

Dans ce cadre, ma mission principale durant ce stage était de développer un système de Retrieval Augmented Generation (RAG). Le RAG est une technologie émergente qui combine les capacités de génération de texte des modèles d'IA avec des systèmes avancés de récupération d'informations. Ce type de système permet de fournir des réponses contextualisées et précises aux requêtes des utilisateurs, en s'appuyant sur une base de données de documents textuels enrichie par des techniques de NLP. Le projet s'insère dans l'initiative plus large du Smart European Shipbuilding (SEUS), qui vise à moderniser le secteur de la construction navale en Europe par l'intégration de technologies intelligentes.

Ce rapport présente les différentes étapes de la conception et de la mise en œuvre de ce système de RAG, les défis techniques rencontrés, ainsi que les compétences et connaissances acquises au cours de cette expérience. Je discuterai également des événements académiques et sociaux auxquels j'ai participé, qui ont enrichi mon stage tant sur le plan professionnel que personnel. Enfin, je proposerai des perspectives d'amélioration et des pistes de recherche future pour le développement de systèmes RAG plus robustes et efficaces. Ce stage a non seulement été l'occasion de contribuer à un projet innovant dans un domaine en pleine expansion, mais m'a aussi permis de me familiariser avec les pratiques de recherche en NLP et d'explorer les applications potentielles de l'IA dans divers secteurs industriels.



2 Présentation de TurkuNLP

TurkuNLP est la branche spécialisée dans le Traitement du Langage Naturel (NLP) au sein du Département d’Informatique de l’Université de Turku, installée au quatrième étage du bâtiment Agora. Le département de linguistique est situé deux étages plus bas, ce qui favorise une coopération étroite entre les deux équipes.

Le groupe est dirigé par le Professeur Filip Ginter, qui a été mon tuteur et principal contact pendant mon stage. Autour de lui, une équipe de chercheurs, assistants chercheurs et post-doctorants travaille avec passion. TurkuNLP est reconnu comme un centre d’excellence dans le domaine du NLP.

Le groupe se distingue par ses recherches variées en NLP, notamment dans les domaines biomédical, biologique, clinique et dans la détection de registres linguistiques sur le web. Cette diversité reflète l’expertise et la polyvalence de l’équipe. TurkuNLP collabore également régulièrement avec le département des humanités numériques de l’Université d’Helsinki dans le cadre du projet HPC-HD (High Performance Computing for the Detection and Analysis of Historical Discourses).

TurkuNLP cherche sans cesse à repousser les limites de la compréhension automatisée du langage. Ses travaux contribuent à des avancées importantes dans des secteurs clés comme la santé et les sciences humaines. Le groupe s’engage également à développer des modèles de langage pour le finnois, une langue peu représentée, avec des créations notables comme FinBERT et FinGPT.

En plus de ses travaux en NLP, TurkuNLP participe à plusieurs projets européens et internationaux, collaborant avec des institutions prestigieuses et prenant part à des conférences majeures. Cette implication montre l’influence et la reconnaissance internationale de TurkuNLP dans le domaine du traitement du langage naturel.



3 Le stage

La mission qui m'a été confiée consistait à développer un système de Retrieval Augmented Generation (RAG) et à le rendre accessible via une application web. Dans un premier temps, je vais expliquer ce système et tout ce que sa mise en place implique. Ensuite, je détaillerai les éléments clés qui ont influencé la conception de cette application, en contribuant à son avancement ou en provoquant des retards, ainsi que ceux qui, bien que n'étant pas directement liés au projet, m'ont personnellement enrichi. Enfin, je décrirai les événements qui ont su rompre la routine quotidienne de mon travail sur cette application, faisant de ce stage une véritable aventure à la fois professionnelle et humaine.

3.1 Introduction au Traitement du Langage Naturel (NLP)

Le Traitement du Langage Naturel, ou Natural Language Processing (NLP) en anglais, est une branche de l'intelligence artificielle qui se concentre sur les interactions entre les ordinateurs et les langues humaines. L'objectif du NLP est de permettre aux machines de comprendre, interpréter et surtout répondre au langage humain de manière utile et significative. Étant donné que le NLP englobe tout ce qui est relatif au langage humain, ses applications sont vastes, allant des chatbots et assistants virtuels aux systèmes de traduction automatique et aux moteurs de recherche avancés. Le système de RAG, se situant à mi-chemin entre un chatbot et un moteur de recherche, s'inscrit parfaitement dans le domaine du NLP.

3.1.1 Embeddings

Les embeddings sont une technique clé dans le NLP moderne. Popularisés en 2013 par la publication de *Word2Vec* [1], qui propose des algorithmes parallélisables avec des architectures neuronales moins complexes et donc beaucoup moins coûteuses en termes de calcul, leur principe est de rendre lisibles par un ordinateur des mots, des phrases ou même des documents entiers en les représentant par des vecteurs de nombres réels dans un espace continu. Cette représentation vectorielle permet de capturer les similarités sémantiques entre les différents éléments du langage. Par exemple, les mots ayant des significations similaires auront des vecteurs proches dans l'espace d'embedding.

Un aspect fondamental des embeddings est la taille des unités de texte qu'ils représentent, appelées tokens. Un token peut être un mot, une partie d'un mot (préfixe, radical, suffixe...), ou même un caractère, en fonction de la segmentation choisie pour le texte. La taille de la fenêtre de contexte, c'est-à-dire le nombre de tokens pris en compte autour d'un mot cible, joue un rôle crucial dans la qualité des embeddings générés. C'est une notion clé dans la construction de notre système de RAG car une fenêtre de contexte trop petite peut manquer d'informations, tandis qu'une fenêtre trop large peut introduire du bruit et diluer l'importance du contexte pertinent.

Depuis *Word2Vec* [1], d'autres modèles et techniques ont été développés pour améliorer



les représentations des mots. Plus récemment, les modèles d'embeddings contextuels, qui créent des représentations des mots dépendant du contexte dans lequel les mots apparaissent au sein d'une phrase ou d'un document, ont vu le jour. Ces modèles utilisent des mécanismes de type Transformer pour analyser simultanément les mots et leur contexte.

3.1.2 Transformer et Modèles de Langage à Grande Échelle (LLM)

Les Transformers cités précédemment ont ouvert la voie à une nouvelle génération de modèles de Langage à Grande Échelle, ou Large Language Models (LLM) en anglais. Ces modèles sont capables de comprendre et de générer du texte de manière très précise. Le plus connu, et qui a mis en lumière le NLP auprès du grand public, est bien évidemment ChatGPT et son modèle GPT-3 [2], Generative Pre-trained Transformer 3, développé par OpenAI. Il utilise l'architecture Transformer pour générer du texte de manière cohérente et contextuellement appropriée, permettant de traduire, répondre à des questions et générer du texte en exploitant les connaissances acquises durant son entraînement sur de vastes corpus de données.

Les Transformers ont été introduits en 2017 dans l'article *Attention is All You Need* [3]. Contrairement aux architectures précédentes comme les réseaux de neurones récurrents (RNN) [4] et les réseaux de neurones convolutifs (CNN) [5], les Transformers encodent les données séquentielles directement via l'embedding et permettent ainsi de paralléliser les calculs plutôt que de les traiter séquentiellement comme les RNN. Ils introduisent également le mécanisme d'attention, qui permet au modèle de se concentrer sur différentes parties de l'entrée séquentielle simultanément. Plutôt que de traiter les données de manière linéaire, le mécanisme d'attention permet à chaque position d'une séquence d'accéder directement à toutes les autres positions. Pour chaque token d'entrée, le Transformer calcule un score d'attention pour chaque autre mot de la séquence, et ces scores déterminent l'importance de chaque mot dans le contexte de celui en cours de traitement. Dans le cas des modèles génératifs comme GPT-3, le Transformer prend en entrée la question et génère la réponse en prédisant le prochain mot dans la séquence jusqu'à ce que la réponse soit complète.

Le Transformer étant la notion clé du NLP moderne et étant au cœur du fonctionnement du système de RAG, je vais désormais proposer une explication de son architecture basée sur la publication *Attention is All You Need* [3] et sur les deux merveilleuses vidéos proposées par Grant Sanderson alias 3Blue1Brown [6] [7] qui tient une chaîne Youtube dédiée à l'enseignement des mathématiques supérieures d'un point de vue visuel. Je recommande à quiconque d'un peu curieux par les mathématiques, ou en l'occurrence l'IA, de consulter sa chaîne Youtube qui est un bijou de pédagogie et d'enseignement. Je tiens à préciser que lors de mon stage je n'ai pas été amené à travailler sur l'architecture d'un quelconque modèle mais directement avec les modèles eux-mêmes. Toutefois il me semble important, notamment d'un point de vu personnel, d'essayer de détailler cette architecture fondamentale. Par soucis de commodité et de pratique scientifique certains termes techniques sont laissés en anglais dans les prochains paragraphes.



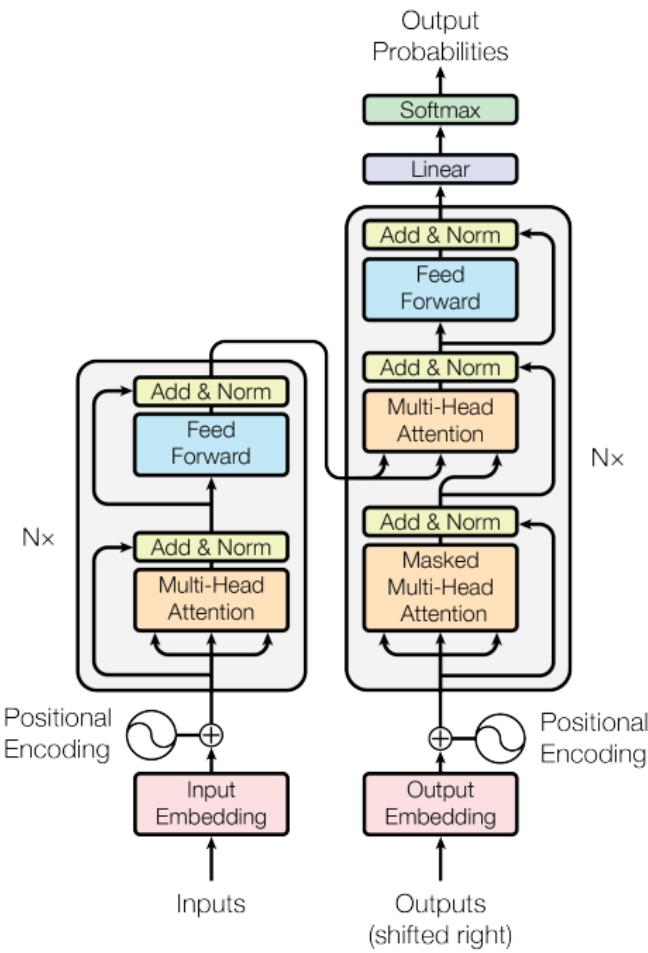


FIGURE 1 – Le Transformer - architecture du modèle

Le Transformer suit l'architecture globale illustrée sur la figure 1 que nous allons détailler ci-dessous.

• Word Embedding et Positional Encoding

D'abord, les mots sont transformés en vecteurs à l'aide de word embeddings, comme détaillé section 4.1.1 : Embeddings. On ajoute ensuite des informations sur la position de chaque mot dans la phrase. Dans la structure de base, ce positional encoding est orchestré par des fonctions sinusoïdales qu'il ne me semble pas pertinent de détailler ici.

• Attention

Rentrons dans le vif du sujet avec le mécanisme d'attention. Vous l'aurez compris, l'objectif derrière tout ça est de pouvoir interpréter le contexte des requêtes et de différencier l'utilisation d'un même mot dans deux phrases différentes. Ainsi, deux phrases comme "le vieux roi anglais était présent au défilé" et "il a gagné avec une paire de roi" contiennent toutes les deux le mot "roi" mais dans des contextes différents. Le but des



mécanismes d'attention est d'ajuster les poids du mot "roi" selon son contexte dans chaque phrase pour que le premier "roi" soit identifier comme le roi du Royaume-Uni dans la première phrase, et que le deuxième soit identifier comme le roi des cartes à jouer.

Scaled Dot-Product Attention

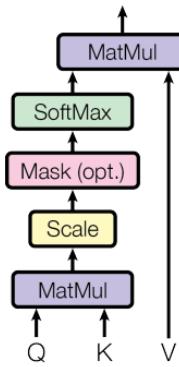


FIGURE 2 – Scaled Dot-Product Attention

La 2 représente le Scaled Dot-Product Attention ou "Attention par produit scalaire normalisé". L'entrée se compose de queries (requêtes) et de keys (clés) de dimension d_k , et de values (valeurs) de dimension d_v . La matrice de sortie d'un bloc d'attention est donnée par la formule suivante que l'on va détailler :

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V$$

• Matrice Q : Queries

Il existe une matrice q contenant des queries, représentées sous forme d'embeddings, que le modèle pose à chaque mot pour tenter de le qualifier. On multiplie la matrice q par les embeddings de chaque mot pour obtenir les vecteurs de queries associés au contexte. Tous ces vecteurs sont ainsi regroupés dans la matrice Q .

Si on reprend notre exemple "le vieux roi anglais était présent au défilé", pour le moment "roi" est interprété par notre système uniquement comme tel. Cependant, la première question qu'on pourrait se poser est "Y a-t-il des adjectifs devant/derrière ce nom ?" pour le qualifier. Il faut noter que chaque query est posée à chaque mot de l'entrée et qu'il existe toute sorte de query propre au modèle.

• Matrice K : Keys

Il existe une matrice k contenant des keys, représentées sous forme d'embeddings, que le modèle pose à chaque mot pour tenter de le qualifier. On multiplie la matrice k par les embeddings de chaque mot pour obtenir les vecteurs de keys. Tous ces vecteurs sont

ainsi regroupés dans la matrice K . La matrice K contient les vecteurs keys associés au contexte, qui peuvent être vus comme les réponses aux queries contenues dans la matrice Q .

Toujours dans notre exemple, cette matrice associerait les vecteurs keys respectifs de "vieux" et "anglais" à des adjectifs et "roi" à un nom.

- **Produit Scalaire pour Correspondance**

Maintenant qu'on a nos matrices, pour mesurer la correspondance entre chaque key et chaque query, on effectue le produit scalaire (dot product) entre chaque paire key-query. Plus les paires sont corrélées, plus le résultat est élevé. Cela donne une carte de la pertinence de chaque mot par rapport aux autres, avec des valeurs allant de très négatives pour les paires les moins en relation à très positives pour les paires contextuellement liées.

Dans notre exemple, les vecteurs keys de "vieux" et "anglais" représentent des adjectifs et sont donc alignés avec le vecteur query de "roi" représentant la question "Y a-t-il des adjectifs devant/derrière ce nom?". Ainsi, leur produit scalaire a une valeur positive élevée, signifiant la corrélation entre ces mots.

- **Normalisation avec Softmax**

La fonction *softmax* [8] convertit un vecteur de K nombres réels en une distribution de probabilités sur K choix. Plus précisément, un vecteur $\mathbf{z} = (z_1, \dots, z_K)$ est transformé en un vecteur $\sigma(\mathbf{z})$ de K nombres réels strictement positifs et de somme 1. La fonction est définie par :

$$\sigma(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}} \quad \text{pour tout } j \in \{1, \dots, K\},$$

c'est-à-dire que la composante j du vecteur $\sigma(\mathbf{z})$ est égale à l'exponentielle de la composante j du vecteur \mathbf{z} divisée par la somme des exponentielles de toutes les composantes de \mathbf{z} .

On applique donc la fonction *softmax* pour normaliser cette matrice, transformant les valeurs réelles en probabilités.

- **Matrice V : Values**

Enfin, on introduit une matrice V , similaire à K mais avec plus de dimensions. Le processus de construction est le même que pour Q et K . Cette matrice définit quels embeddings doivent être ajoutés aux autres pour actualiser leur contexte via une dernière multiplication par la matrice des distributions obtenue après le *softmax*.

- **Application des Poids**

Pour terminer avec notre exemple, si les adjectifs "vieux" et "anglais" se réfèrent au



nom "roi" avec une importance obtenue après la fonction *softmax* de 0,5 chacun et que tous les autres mots sont à 0, cette dernière étape ajoute les vecteurs values de "vieux" et "anglais" au vecteur d'embedding de "roi" avec un coefficient de 0,5.

Ainsi, si le modèle est bien entraîné, il peut identifier "le vieux roi anglais" comme étant le *roi Charles III du Royaume-Uni*.

Voici le fonctionnement général d'une seule tête d'attention au travers d'un exemple bien particulier. Cet exemple a pour vocation de proposer une illustration du concept d'attention, qui est en réalité un peu plus abstrait que l'exemple explicitement cité puisqu'il est interprété par une machine.

En effet il est important de comprendre que l'attention opère à un niveau bien plus complexe. Le modèle, à travers des processus d'apprentissage sur de larges corpus de données, développe des représentations internes sophistiquées qui capturent des relations subtiles entre les mots. Ces relations ne sont pas explicitement régies par des règles humaines, mais sont plutôt découvertes par le modèle au fur et à mesure de son entraînement. Ainsi, l'attention peut capter des nuances de langage qui échappent parfois même à nous les humains, en analysant des milliers de contextes différents et en apprenant à ajuster les poids de manière optimale.

Ce qu'il faut retenir du bloc d'attention, c'est qu'il déplace de l'information stockée dans un vecteur d'embedding vers un autre vecteur d'embedding pour enrichir ce nouveau vecteur et lui fournir le contexte dans lequel il se situe [9].

• Multi-Head Attention

Le transformer met en parallèle différents blocs d'attention, ce qui permet au modèle de se concentrer conjointement sur des informations provenant de différents sous-espaces de représentation à différentes positions. On appelle ce processus l'attention multi-head.

• Feed Forward

Après les mécanismes d'attention, chaque couche du transformer contient une couche feed forward [10] positionnée de manière identique et appliquée de manière indépendante à chaque position. Ces couches feed forward sont des réseaux de neurones entièrement connectés et servent à transformer les représentations vectorielles issues des mécanismes d'attention pour ajouter plus de complexité et de non-linéarité au modèle. Notamment :

- Complexité et Non-linéarité : Les mécanismes d'attention permettent de pondérer l'importance des différents mots entre eux, mais ils restent des opérations linéaires. Ainsi, Les couches feed forward viennent introduire de la non-linéarité, permettant au modèle de capturer des relations plus complexes entre les mots.
- Transformation des Représentations : Les feed forward layers permettent de transformer les représentations vectorielles intermédiaires en de nouvelles représentations plus riches. Cela aide le modèle à extraire des caractéristiques plus abstraites.



- **Décodage**

Une fois la séquence d'entrée passée à travers les différentes couches d'attention et de feed forward on obtient le vecteur d'embeddings final contenant tout le contexte de l'entrée. Ce contexte va être utilisé pour générer la séquence de sortie par le décodeur.

Le décodeur génère la séquence de sortie un token à la fois. À chaque étape, le modèle prédit la probabilité de chaque mot possible et choisit le mot le plus probable comme prochain token dans la séquence de sortie.

Le décodeur utilise lui aussi le mécanisme d'attention avec un principe de masque qui garantit que les prédictions pour la position i ne peuvent dépendre que des sorties connues aux positions inférieures à i .

Enfin, pour chaque token généré, le décodeur utilise une matrice d'unembedding pour convertir les représentations vectorielles finales en scores de probabilité pour chaque mot du vocabulaire. Le mot avec la probabilité la plus élevée est sélectionné comme prochain token dans la séquence.

En résumé, le Transformer est une architecture qui interprète et contextualise des séquences de texte. D'abord, les mots sont transformés en vecteurs grâce aux word embeddings, et des informations de position sont ajoutées. Le mécanisme central de l'architecture est l'attention, qui permet au modèle de comprendre le contexte dans lequel un mot est utilisé en ajustant dynamiquement les poids des mots en fonction de leur contexte. Le modèle utilise ensuite plusieurs têtes d'attention pour capturer différentes perspectives du texte et y ajoutent de la complexité en introduisant des transformations non linéaires avec les couches feed forward. Enfin, le décodeur génère une séquence de sortie en prédisant les tokens un par un, en s'assurant que chaque prédition repose uniquement sur les tokens précédents.

Le Transformer se distingue par sa capacité à capter des relations complexes entre les mots d'une phrase grâce à ses mécanismes d'attention. Cette architecture permet une compréhension fine du contexte, indispensable pour des tâches de traduction, de génération de texte, et d'autres applications en traitement du langage naturel. L'utilisation simultanée de multiples têtes d'attention et la structure en couches permettent de traiter des informations complexes de manière efficace, rendant le Transformer particulièrement performant par rapport aux approches précédentes.

3.2 Conception d'un système de Retrieval Augmented Generation

Le Retrieval Augmented Generation (RAG) est une technique appartenant au domaine du Traitement du Langage Naturel (NLP) qui associe la puissance des modèles d'intelligence artificielle (IA) génératifs à celle des systèmes de recherche d'informations.

La notion de RAG est une innovation récente en matière d'IA générative et de Machine Learning, qui promet de révolutionner la récupération d'informations telle que nous la connaissons [11]. Cette technologie permet de maintenir à jour le flux continu de nou-



velles données tout en améliorant la compréhension des requêtes des utilisateurs grâce aux avancées en traitement du langage et en analyse sémantique. Contrairement aux méthodes traditionnelles de recherche d'informations, le RAG ne se limite pas à la simple récupération de documents pertinents mais inclut également la génération de réponses précises et contextuelles aux requêtes des utilisateurs.

Ainsi, l'utilisateur n'a plus besoin de chercher manuellement parmi des dizaines de sites web proposés par une recherche classique ni de reformuler sa requête en cas de résultats insatisfaisants. Le système de RAG permet d'obtenir une réponse précise en combinant des systèmes de recherche avancés, basés sur des embeddings et des modèles de génération de texte comme GPT. Le contexte des questions précédentes est pris en compte, permettant au modèle de RAG d'affiner ses réponses pour correspondre parfaitement aux attentes de l'utilisateur.

Ce projet a vu le jour au sein d'une initiative de plus grande ampleur : Smart European Shipbuilding (SEUS). Le projet SEUS vise à créer un cadre de construction navale intelligente, permettant de gagner considérablement du temps dans l'ingénierie, l'assemblage et la construction dans les chantiers navals européens grâce à une intégration et une utilisation efficaces des outils informatiques inclus dans ce cadre.

Le laboratoire Turku NLP participe ainsi à la partie "intelligente" du projet SEUS en proposant des travaux de recherche sur le système de recherche avancée RAG.

Ce système présente un grand avantage pour les entreprises et va devenir un outil essentiel dans les prochaines années. En effet, il est possible de faire fonctionner ce système de manière totalement autonome et locale en utilisant des modèles open-source, garantissant ainsi qu'aucune donnée ne fuit. Cet aspect est crucial pour des entreprises travaillant avec des données sensibles qui doivent rester confidentielles.

3.2.1 Architecture du système

On peut schématiser la chaîne de traitement (pipeline) d'un système de RAG simple [12] tel qu'on peut le voir sur la figure 3. On peut décomposer cette chaîne en deux, d'un côté la préparation des données, de l'autre le traitement des données avec la recherche et la génération de réponse.

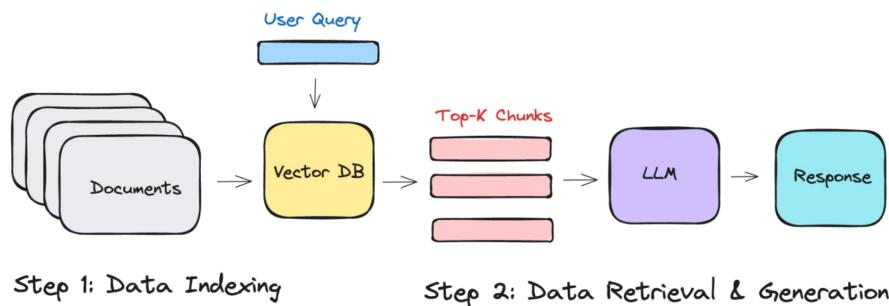


FIGURE 3 – Basic RAG pipeline

- Étape 1 : Préparation des données

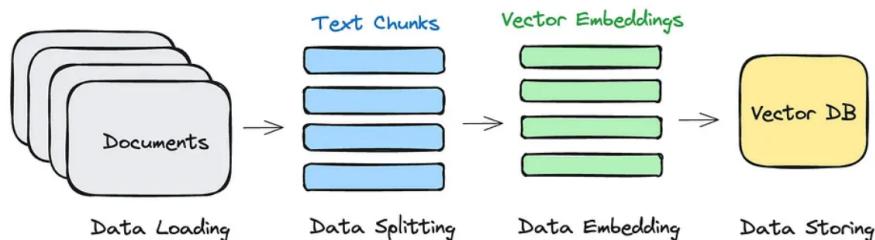


FIGURE 4 – Data Indexing

La première étape, schématisée par la figure 4, consiste à constituer la base de données du système avec des documents textuels. N’importe quelle extension peut être utilisée tant qu’il existe un outil pour lire le texte avec cette extension.

Un document unique peut contenir des centaines de pages, des milliers de caractères et couvrir une multitude de sujets différents. Pour que les documents soient lisibles et interprétables par notre système, il faut les diviser en morceaux de texte, que nous appellerons chunks par la suite. Ces chunks sont ensuite représentés par des vecteurs d’embeddings qui capturent le contexte contenu dans ces chunks et permettent la recherche avancée par la suite. La création de ces chunks est cruciale pour l’efficacité du système, car une fenêtre de contexte trop petite peut manquer d’informations, tandis qu’une fenêtre trop large peut introduire du bruit et perdre la partie la plus importante du contexte.

Ces vecteurs d’embeddings associés aux différents chunks sont ensuite stockés dans une autre base de données, cette fois-ci vectorielle, interne au système.

En résumé, la préparation des données consiste à lire les documents, les séparer en petits morceaux de texte, associer à chaque morceau une représentation vectorielle au sein de l’espace d’embeddings, et stocker cette représentation dans une base de données vectorielle.

- Étape 2 : Recherche et Génération

Une fois que la base de données est constituée, l’utilisateur peut effectuer des requêtes, ou queries en anglais, le processus est détaillé figure 5.

Dans un premier temps, la requête de l’utilisateur est lue par le modèle d’embeddings afin de lui attribuer un vecteur dans l’espace d’embeddings. Ce vecteur est comparé aux vecteurs associés aux chunks stockés dans notre base de données, et le système fait ressortir les chunks dont le contexte est le plus proche de la requête de l’utilisateur dans l’espace d’embeddings.

Dans un second temps, la requête de l’utilisateur ainsi que les chunks trouvés lors de la première étape sont donnés au LLM pour générer une réponse à la question de l’utilisateur en se basant uniquement sur le contexte relevé, à l’aide d’un prompt spécifique.



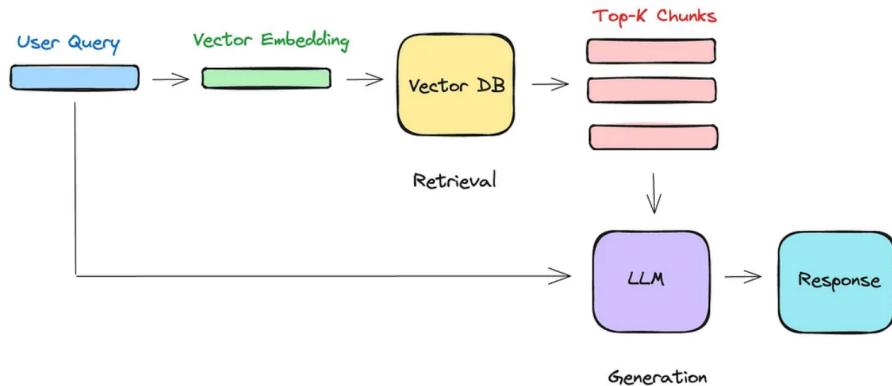


FIGURE 5 – Data Retrieval and Generation

3.2.2 Mise en oeuvre du système

Pour mettre en oeuvre ce système de RAG, j'ai utilisé Python. En réalité, il est relativement rapide de créer une chaîne aussi basique que celle décrite précédemment, en raison des nombreuses documentations disponibles et de l'intérêt croissant pour ce sujet. Il existe de nombreuses bibliothèques Python open source pour les tâches d'IA et de NLP, telles que PyTorch, Transformer ou LangChain, qui offrent des outils puissants et bien développés pour implémenter rapidement ce type de pipeline. Une fois les concepts expliqués précédemment maîtrisés et les outils appropriés appréhendés, il est possible d'obtenir rapidement des premiers résultats.

3.2.3 Gestion des données avec ChromaDB

ChromaDB est une bibliothèque Python conçue pour la gestion et la manipulation de bases de données vectorielles. Elle est particulièrement utile dans le domaine de la recherche d'informations. ChromaDB permet de stocker des représentations vectorielles de données textuelles ou multimédias, ce qui permet de réaliser des recherches basées sur la similarité vectorielle, essentielles pour les applications de NLP et de machine learning. En utilisant ChromaDB, on peut indexer et interroger des documents de manière efficace, en exploitant des techniques de clustering et de classification. Une technique de clustering, ou de regroupement, est une méthode d'apprentissage automatique non supervisée utilisée pour organiser un ensemble de données en groupes homogènes appelés clusters. Les objets au sein d'un même cluster sont plus similaires entre eux qu'avec ceux des autres clusters.

ChromaDB est conçu pour être performant et efficace en termes d'utilisation de l'espace. Toutefois, l'espace de stockage requis dépend principalement de la quantité de données qu'on y stocke et de la complexité des représentations vectorielles utilisées.

Dans le cadre de mon stage, j'ai donc utilisé ChromaDB pour créer le système de RAG.

3.2.4 Cration de l’application web

Le plus grand defi, notamment durant le premier mois du stage, etait de programmer une premire demonstration du systme de RAG afin de la presenter lors du workshop annuel du projet SEUS, qui s’est tenu du 22 au 24 mai 2024 a Turku.

J’ai donc reussi a deployer une premire version fonctionnelle, qui, bien que pas des plus esthtiques, permet a quelqu’un d’tranger au domaine de la programmation et du NLP de comprendre l’objectif de l’application RAG.

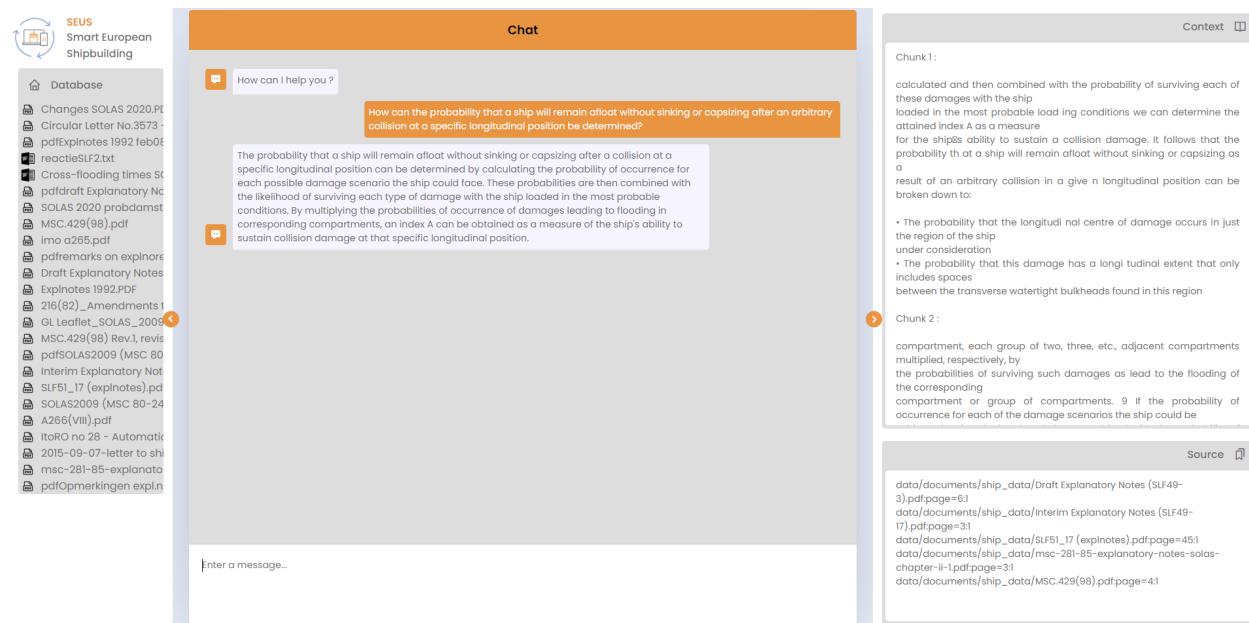


FIGURE 6 – Premire version de l’application web

Cette premire version m’a permis de me familiariser avec les langages de programmation web HTML, CSS et JavaScript, que je n’avais jamais utiliss auparavant, et de poser des bases solides pour une deuxime version.

Dans cette premire version, sur laquelle je ne vais pas trop m’attarder, on retrouve a gauche la base de donnees, au centre la bote de dialogue, et a droite un menu permettant d’afficher les chunks identifis comme intressants par le systme et ayant servi a genrer la reponse, ainsi que leur source.

La base de donnees chargee sur la figure 6 comporte des documents sur les regulations et les normes de construction de bateaux. La question posee sur la capture d’cran avait ete cree a partir d’un des documents de la base de donnees, et le document a bel et bien ete retrouu, ainsi que le passage correspondant. Le systme a egalement trouu d’autres passages pertinents pour repondre a la question dans d’autres morceaux de textes. L’utilisateur peut ensuite consulter ces sources pour verifier ou etofer la reponse proposee par le LLM.

Aprs les resultats extrêmement satisfaisants de la premire version de l’application web et le besoin d’améliorer l’interface graphique, j’ai decide de repartir de zero pour une deuxime

version. La première version étant ma toute première interface graphique web, j'ai préféré repartir sur des bases solides en exploitant ce que j'avais appris durant cette première itération. L'idée était de créer une maquette solide pour ma collègue qui reprendrait le projet après mon départ.

La deuxième version, illustrée sur la figure 7, se veut plus minimaliste, ergonomique et pratique pour l'utilisateur. La structure générale reste la même avec un chat central et deux menus latéraux que l'utilisateur peut fermer s'ils ne lui sont pas utiles.

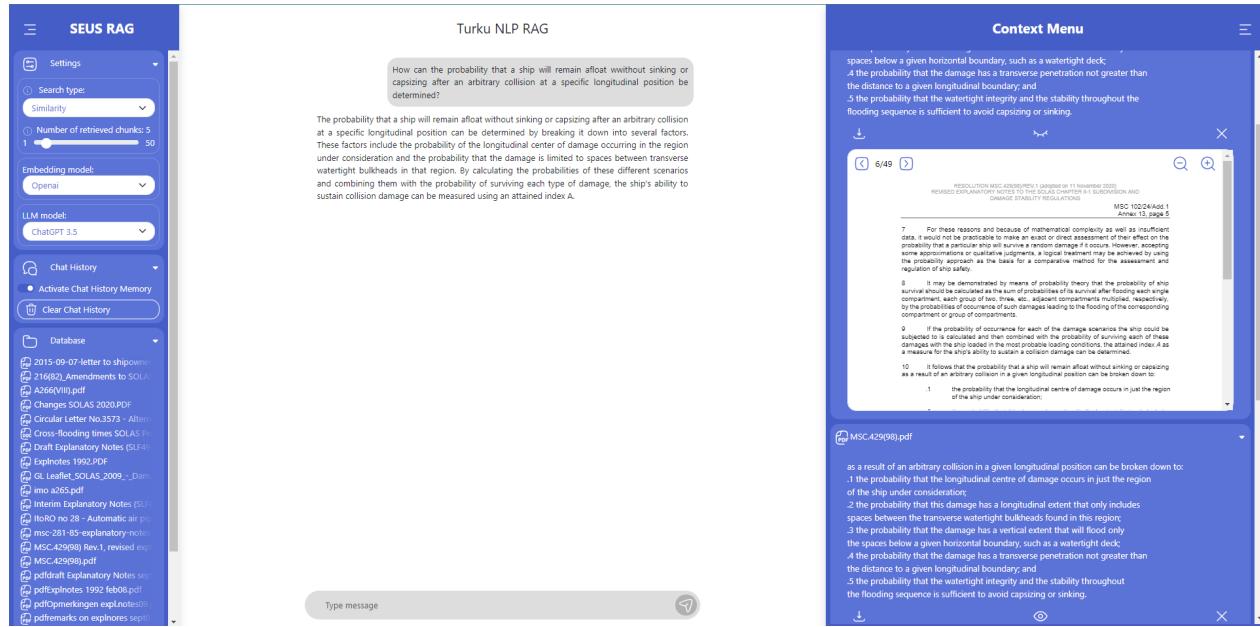


FIGURE 7 – Deuxième version de l'application web

La principale nouveauté est l'introduction de paramètres permettant à l'utilisateur de configurer le système RAG selon ses besoins. Le menu de droite, contenant les chunks de contexte extraits, a été nettement amélioré. Chaque chunk est désormais associé directement à sa source, plutôt que d'être séparé en deux fenêtres comme dans la première version. L'utilisateur peut choisir d'afficher ou non le contenu des chunks, chaque chunk étant un élément déroulant. L'utilisateur peut visualiser directement le chunk dans le document. De plus, il peut directement télécharger le document depuis ce menu sans devoir le chercher dans la base de données. L'utilisateur peut supprimer un chunk qui ne l'intéresse pas dans ses recherches avec la croix. Chaque requête est séparée par une balise *search + numéro*. Les paramètres permettent de modifier le système de différentes manières, notamment :

- **Search Type :** permet de modifier la méthode de sélection des chunks.
 - *Similarité* : sélectionne les chunks les plus similaires à la question. Une barre de défilement permet d'ajuster le nombre de documents à inclure.
 - *Seuil de Score de Similarité* : sélectionne les documents ayant un score de similarité supérieur à un seuil fixé par l'utilisateur, qui fixe également un nombre maximal de documents à relever. Ce paramètre permet de sélectionner un nombre conséquent de documents en garantissant leur pertinence par rapport à la question.



- *Maximal Marginal Relevance (MMR)* [13] : tente de réduire la redondance des résultats tout en maintenant leur pertinence. Il prend en compte la similarité de la question avec le document, ainsi que la similarité des phrases déjà sélectionnées.

$$MMR = \operatorname{Arg\,max}_{D_i \in R \setminus S} \left[\lambda(\operatorname{Sim}_1(D_i, Q) - (1 - \lambda) \max_{D_j \in S} \operatorname{Sim}_2(D_i, D_j)) \right]$$

- Q = Requête
- D = Ensemble de documents liés à la requête Q
- S = Sous-ensemble de documents déjà sélectionnés dans R
- $R \setminus S$ = Ensemble de documents non sélectionnés dans R
- λ = Constante dans la plage [0-1], pour la diversification des résultats
- $Sim1$ est la métrique de similarité utilisée pour la récupération de chunks et le classement de leur pertinence par rapport à la requête. $Sim2$ peut être la même que $Sim1$ ou une autre métrique.

En fixant λ à 0,5, on obtient un équilibre optimal entre diversité et précision dans les résultats. La valeur de λ peut être ajustée selon le cas d'utilisation et l'ensemble de données.

- **Embedding model** : permet de changer le modèle d'embedding utilisé.

- Certains modèles d'embedding peuvent être spécialement entraînés pour des domaines spécifiques (médical, juridique, financier, etc.), offrant ainsi des représentations vectorielles plus pertinentes et efficaces pour ces applications.
- Chaque modèle ayant sa propre représentation vectorielle des contextes, il peut être intéressant de changer de modèle pour observer les différences dans les chunks extraits.
- Les modèles d'embedding évoluent pour couvrir une plus grande variété de langues et de dialectes, ce qui est crucial pour des applications multilingues ou localisées.
- Certains modèles peuvent offrir des interfaces ou des API plus compatibles avec les infrastructures modernes, facilitant ainsi leur intégration dans les systèmes existants.

- **LLM model** : permet de changer le modèle de LLM utilisé.

- Changer de modèle peut permettre d'adapter les réponses aux préférences linguistiques, stylistiques ou de ton des utilisateurs, rendant l'interaction plus naturelle et agréable.
- Si un modèle présente moins de biais que celui actuellement utilisé, cela peut conduire à des interactions plus équitables et inclusives, améliorant l'expérience utilisateur pour un public diversifié.
- Un modèle spécialisé dans un domaine particulier (comme la médecine, le droit, la finance) peut offrir des réponses plus pertinentes et informatives pour les utilisateurs ayant des besoins spécifiques dans ces domaines.



- Certains modèles peuvent être plus efficaces en termes de ressources computationnelles, ce qui peut se traduire par des réponses plus rapides et une meilleure utilisation des ressources, améliorant ainsi l'expérience utilisateur en termes de réactivité.
- **Chat History** : entre la première et la deuxième version du site, j'ai implémenté la possibilité pour le système de suivre l'historique du chat. C'est un élément essentiel pour un chatbot car il permet de simuler une véritable conversation entre l'utilisateur et le chatbot. Cela permet d'affiner les réponses de question en question et de mentionner des éléments des questions ou réponses précédentes de manière indirecte. Par exemple, question 1 : "*Qui est le président de la France ?*", question 2 : "*Depuis quand est-il au pouvoir ?*"; la question 2 ne serait pas interprétable par notre système sans accès à l'historique du chat puisqu'il ne saurait pas à qui "*il*" fait référence. J'ai donc implémenté deux boutons pour gérer l'historique du chat : un pour l'activer ou le désactiver si l'on souhaite faire uniquement des requêtes simples, et un autre pour l'effacer si la conversation devient trop longue ou si l'on veut changer de sujet.

3.2.5 Résultats

Le principal résultat de mon stage est la création du site web finalisé et opérationnel, qui témoigne de lui-même de l'avancée du projet. Un autre résultat que j'aurais aimé pouvoir analyser est le système RAG, notamment la qualité de ses réponses et la pertinence des documents qu'il récupère à chaque requête. Malheureusement, nous n'avons pas encore mis en place de système d'évaluation, comme je le détaille dans la section suivante Perspectives. Je vais néanmoins discuter brièvement des résultats obtenus dans les cas d'applications que j'ai pu expérimenter.

Des captures d'écran de différentes conversations sont disponibles en section 6 : Annexes.

Tout d'abord, la base de données chargée pour le projet SEUS. Cette base contient un échantillon de documents portant sur les normes et régulations concernant la construction de bateaux, fournis par un des collaborateurs du projet. Une des utilisations que j'ai envisagées est la recherche d'une certaine formule mathématique ou physique, comme illustré en figure 13 en annexe. Avec la requête appropriée, il est très facile de retrouver la formule. La réponse du chatbot reste cependant limitée, notamment lorsque la formule utilise des signes mathématiques qui n'ont pas pu être correctement lus lors du traitement de la base de données. Néanmoins, les segments de texte contenant la formule sont bien identifiés, ce qui permet à l'utilisateur d'accéder au document et de consulter la formule directement.

Le projet RAG est né au sein du projet SEUS, mais son application est universelle. C'est pourquoi, durant mon stage, j'ai également collaboré avec le laboratoire d'archéologie de l'université de Turku pour développer un système RAG spécifique à leurs besoins. Depuis le début de la guerre en Ukraine, toutes les relations avec la Russie ont été rompues, y compris les relations universitaires. Cependant, l'histoire de la Finlande est étroitement liée à celle de la Russie, et certains sites archéologiques russes sont d'un grand intérêt pour les chercheurs



finlandais. Ces chercheurs ont identifié des centaines de publications scientifiques russes potentiellement intéressantes, mais inaccessibles sans la connaissance de la langue russe. En plus d'être inaccessibles, ces documents sont vastes et nécessiteraient un travail d'analyse considérable. C'est ici que le système RAG intervient : il traduit et trouve l'information pertinente !

Ce cas d'utilisation semble idéal, encore faut-il savoir quoi chercher... Mais même sans une idée précise, avec un tel concentré d'informations historiques, il est facile de partir d'une notion générale et d'arriver, en quelques requêtes, à un point plus précis qui justifie d'ouvrir le document et de lire quelques paragraphes à ce sujet. Quelques exemples présents en annexe illustrent ce cheminement.

Enfin, j'ai voulu expérimenter le RAG avec des documents que je connaissais. J'ai donc construit une base de données à partir de différents comptes-rendus que j'avais rédigés et de quelques PDF de cours que je possédais sur mon disque dur. Le résultat est impressionnant : rechercher un passage de cours qui nous intéresse devient vraiment facile.

3.2.6 Perspectives

Le projet de RAG en est encore à ses débuts, et j'ai principalement travaillé sur les étapes préliminaires, en mettant en place un système simple mais utilisable via une application web. Avant de commencer le développement de la V2 de l'interface graphique, je me suis beaucoup questionné sur la suite du projet et les améliorations possibles. J'ai donc consulté plusieurs articles et papiers scientifiques qui explorent diverses pistes d'amélioration. Il est également important de noter que le travail que j'ai effectué jusqu'à présent ne constitue pas une recherche académique approfondie. La prochaine étape pour M. Ginter et Maryam sera de choisir un domaine spécifique à approfondir, afin de mener des recherches et de rédiger un article scientifique. Voici une liste non exhaustive des pistes d'amélioration que je considère intéressantes :

- **Méthodes d'évaluation :** Depuis le début de mon stage, une question me préoccupe : comment savoir si mon système est fiable ? Pour le moment, nous nous basons sur notre appréciation subjective des réponses générées. Bien que les résultats soient prometteurs, il est difficile de les quantifier précisément. C'est la raison pour laquelle je n'ai pas présenté de résultats concrets dans les sections précédentes. L'évaluation du système est cruciale pour mesurer son efficacité et orienter les améliorations futures. Actuellement, les méthodes d'évaluation incluent souvent des mesures telles que la précision et le rappel. La précision mesure la proportion de documents pertinents parmi ceux récupérés, indiquant la capacité du système à éviter les faux positifs. Le rappel, quant à lui, évalue la proportion de documents pertinents correctement identifiés parmi tous les documents pertinents disponibles, mesurant ainsi la capacité à éviter les faux négatifs. De nouvelles méthodes de notation, telles que l'utilisation de métriques de cohérence sémantique ou de satisfaction perçue par l'utilisateur, sont également explorées. Un système de notation robuste permettrait non seulement de comparer les versions successives du système, mais aussi de mesurer l'impact des améliorations apportées. Sans



un système d'évaluation solide, les autres points d'amélioration n'ont que peu de valeur.

- **Préparation des données :** La qualité des données est essentielle pour tout système de RAG. Une étape importante serait d'améliorer les processus de préparation et de nettoyage des données. Actuellement, notre système peut être limité par des données bruitées ou mal structurées, ce qui affecte la précision et la pertinence des réponses générées. Une meilleure préparation des données, incluant la normalisation, la déduplication et l'élimination des biais, permettrait de fournir des entrées plus cohérentes et de haute qualité au système. Cela pourrait réduire les erreurs de traitement du langage naturel et améliorer la pertinence des documents récupérés et des réponses générées.
- **Méthodes de découpage :** Comme expliqué précédemment, le découpage des documents en morceaux plus petits (chunking) est une technique couramment utilisée dans les systèmes RAG pour gérer de grandes quantités de texte. Explorer et affiner les méthodes de découpage pourrait significativement améliorer l'efficacité et la précision du système. Différentes stratégies de chunking, telles que le découpage basé sur les paragraphes, les phrases, ou les thèmes, peuvent être comparées pour déterminer laquelle optimise le mieux la balance entre granularité et contexte. Actuellement, le système utilise une fenêtre fixe de 800 caractères avec un chevauchement (overlapping) de 80 caractères. Le chevauchement consiste à inclure une partie du texte déjà présente dans le chunk précédent ou suivant, ce qui peut parfois conduire à des chunks commençant ou se terminant au milieu d'une phrase ou d'un paragraphe. Une meilleure segmentation des documents pourrait permettre une récupération plus ciblée des informations pertinentes et une génération de réponses plus précises et contextuellement appropriées.
- **Intégration des images :** L'intégration de la lecture et de l'analyse des images issues des documents dans le système RAG constitue une perspective prometteuse pour enrichir la base de données et améliorer la pertinence des réponses générées. Actuellement, de nombreux systèmes se concentrent principalement sur le texte, laissant de côté les informations visuelles qui peuvent être cruciales, notamment dans des domaines tels que la recherche scientifique, les rapports techniques, et les documents historiques. L'extraction de texte à partir d'images via la reconnaissance optique de caractères (OCR) permet de convertir des contenus visuels en textes exploitables, tandis que l'analyse d'images peut identifier des éléments visuels importants comme des graphiques, des tableaux ou des schémas. En intégrant ces éléments dans la base de données, le système RAG pourrait offrir des réponses plus complètes et contextuelles, incluant des descriptions d'images ou des explications basées sur des visualisations. Cette capacité à traiter et intégrer des informations visuelles pourrait également permettre une recherche multimodale, où les utilisateurs peuvent interroger le système avec des questions se rapportant à des images spécifiques, augmentant ainsi l'utilité et la robustesse du système. À l'heure actuelle, les images et les figures constituent une source de bruit dans notre système, car seul leur titre ou leur description est lu et intégré à la base de données.
- **Amélioration du système de Récupération :** De nombreuses pistes peuvent être



envisagées pour améliorer le système de récupération lui-même, et les progrès continus des nouveaux modèles d'embeddings contribueront probablement à cette amélioration. Cependant, une perspective particulièrement intéressante est celle présentée dans le papier scientifique *Multi-Head RAG : Solving Multi-Aspect Problems with LLMs* [14]. Ce papier propose d'utiliser l'activation des différentes têtes d'attention en revenant une couche en arrière dans le modèle d'embedding pour améliorer la précision de la récupération de documents pertinents dans les requêtes complexes. Cette méthode permet de répondre à des requêtes comprenant des contextes éloignés dans l'espace d'embedding sans se concentrer uniquement sur le contexte prépondérant, comme le fait notre système actuel qui concatène toutes les têtes d'activation pour ne retourner qu'une seule tête d'attention.

Les systèmes de RAG étant encore nouveaux et très prometteurs, il existe de nombreuses possibilités de recherche pour explorer leur fonctionnement et leurs améliorations potentielles. C'est avec un peu de tristesse que je vais quitter TurkuNLP, car je me suis beaucoup attaché au projet et j'aurais aimé le voir évoluer. J'espère avoir construit une base solide pour mes collègues de TurkuNLP, et peut-être aurais-je l'occasion de retravailler sur ce type de système à l'avenir. À voir ce que le futur nous réserve.

3.3 Connaissances générales

Avant d'arriver au site web opérationnel je suis évidemment passé par des étapes d'apprentissage et de pratique. Je dédis donc cette section à la description des tâches et des éléments qui ne rentraient pas forcément dans la description du système en lui-même mais qui feront de moi, je l'espère, un meilleur ingénieur.

3.3.1 Les Superordinateurs

Une première difficulté que j'ai rencontrée, notamment les premières semaines, concernait l'utilisation des superordinateurs. L'université de Turku a accès aux superordinateurs du Centre de Technologie et de l'Information pour la Science (CSC), une association à but non lucratif finlandaise appartenant à l'État et aux établissements d'enseignement supérieur. J'ai notamment effectué une grande partie de mon travail sur le superordinateur nommé Puthi. L'utilisation de cet environnement virtuel n'était pas évidente au début ; j'avais des difficultés à installer correctement mes packages, la gestion des chemins d'accès était compliquée, et la maniabilité n'était pas aussi confortable que sur un ordinateur personnel. En effet, toute la gestion et la visualisation des fichiers se faisaient via le terminal de commande. Cela m'a donc permis d'apprendre les commandes UNIX usuelles et de me familiariser avec l'invite de commande. J'ai également appris à me connecter à un serveur et à créer des ponts entre serveurs.

Bien que j'avais déjà utilisé des environnements virtuels pour de précédents projets, cette notion est devenue rapidement importante à cause des conflits entre versions des différents



packages que j'utilisais. Cela m'a donc permis d'approfondir cette pratique courante mais essentielle en Python.

3.3.2 Acquérir des connaissances

Après avoir discuté du sujet de mon stage avec Filip Ginter, mon tuteur, le premier jour, j'étais extrêmement excité car le sujet me passionnait, bien que je n'avais que très peu de connaissances à ce propos. Habitué au modèle scolaire, il m'a fallu, pour la première fois, constituer par moi-même les connaissances nécessaires à la maîtrise du sujet. J'ai particulièrement apprécié lire des articles sur le sujet, mais je me suis rapidement perdu dans la masse d'informations disponibles. Avec les nombreuses avancées récentes en NLP et l'intérêt croissant pour ce domaine, il existe une multitude d'articles de qualité variable. Il m'était parfois difficile de déterminer la pertinence de ce que je lisais, et j'avais souvent l'impression que la partie théorique était négligée. Néanmoins, je pense avoir acquis une compréhension globale solide du sujet.

J'ai découvert pour la première fois la lecture de papiers scientifiques et j'ai particulièrement apprécié cette activité. Il m'arrivait fréquemment de consulter les publications récentes liées à mon sujet pour explorer les possibilités d'amélioration de mon système. J'ai consacré du temps à la lecture et à la relecture des papiers fondamentaux qui ont mis le NLP en avant et en ont fait un enjeu majeur pour l'avenir. Je prévois de continuer à suivre les nouvelles publications après la fin de mon stage afin de rester à jour sur ce sujet en constante évolution, et parce qu'il me passionne.

3.3.3 Programmation

J'ai programmé l'intégralité de l'application seul, ce qui m'a permis de travailler avec plusieurs langages et technologies différentes. Le code est d'ailleurs accessible sur GitHub [15]. Une collègue, Maryam Teimouri, a rejoint le projet au cours de mon stage, mais elle était principalement concentrée sur la rédaction de sa thèse et n'a pas eu le temps de m'aider dans la création de l'application. Cependant, elle m'a initié à la collaboration sur GitHub, et je l'en remercie, car savoir utiliser GitHub est un véritable atout, notamment pour la recherche de stage ou d'emploi l'année prochaine.

- **Python :**

Toute la partie backend et IA a été programmée en Python. Pour la partie IA, j'ai exploré de nombreuses technologies pour me familiariser avec l'utilisation des modèles, en particulier les premières semaines. La librairie LangChain s'est avérée être un outil très puissant pour les tâches de NLP et est au cœur du système de RAG.

La connexion entre l'interface graphique et le code Python a été effectuée avec Flask. Flask est une bibliothèque Python qui permet de créer facilement des sites web et des applications en définissant des pages, en gérant les requêtes et en ajoutant des fonctionnalités avec des extensions. Elle est simple à utiliser et idéale pour des projets web



rapides. TurkuNLP ne visant pas à commercialiser cette application à grande échelle, Flask était le choix idéal.

- **HTML, CSS, JavaScript :**

Ces trois langages sont utilisés pour la création d'interfaces graphiques web. Le HTML gère la structure générale de la page et affiche les éléments souhaités. Le CSS permet de mettre en forme la page et de la rendre attrayante. Le JavaScript, quant à lui, permet de mettre en place et de gérer toutes les interactions entre l'utilisateur et le site, ainsi que de les relier au code Python.

Alors que j'avais réalisé la première version du site web en quelques jours, j'ai passé plus d'un mois sur la deuxième version. En effet, le désir de produire une version définitive, flexible, réutilisable et maintenable m'a poussé à peaufiner chaque détail. Cela m'a pris beaucoup plus de temps que je ne l'avais imaginé, et je n'ai pas pu entreprendre tout ce que j'avais prévu. Néanmoins, la version finale semble répondre aux critères précédemment cités, bien que je n'aie pas assez de recul et d'expérience pour évaluer pleinement la maintenabilité et la qualité de mon code.

J'aurais aimé consacrer plus de temps durant mon stage à l'amélioration du système de RAG, notamment la partie backend, mais je suis heureux d'avoir découvert la réalité de la conception d'un site web. En effet, le frontend m'a toujours attiré, et j'aurais pu envisager un futur poste en tant que développeur frontend, mais cette expérience, qui m'a conduit à passer des heures à ajuster de simples alignements en CSS, m'a permis de voir les aspects moins plaisants du frontend. Bien sûr, avec l'expérience, je gagnerai en efficacité pour résoudre ces problèmes, mais au-delà de cela, je n'ai pas ressenti la même satisfaction intellectuelle que celle procurée par la résolution de problèmes algorithmiques, par exemple. Néanmoins, j'ai apprécié de pouvoir programmer à la fois la partie technique (backend) et de concrétiser ma vision (frontend), faisant de cette expérience fullstack un véritable enrichissement.

3.4 Autres expériences

Bien que j'ai consacré une grande partie de mes 4 mois de stage à travailler sur mon application, il y a eu quelques évènements qui sont venus enrichir mon stage au sein de TurkuNLP, faisant de ce stage une expérience professionnellement et humainement enrichissante.

3.4.1 Conditions de travail

Habituellement plein, le bureau de TurkuNLP, figure 8a, présente une agréable atmosphère de travail au sein d'une équipe internationale, figure 8b. Ce fut un pur plaisir de collaborer et de discuter avec tous les membres de l'équipe.





(a) Bureau de TurkuNLP



(b) Membres de TurkuNLP

Concernant mon travail personnel, j'étais vraiment très autonome. M. Ginter me donnait des lignes directrices que je suivais. Régulièrement, je lui montrais les avancées et on pouvait ainsi en discuter et planifier la suite. Le travail était très libre et cela m'a permis notamment de prendre mon temps pour me documenter, lire des articles ou des publications scientifiques.

3.4.2 SEUS Workshop

Un mois après le début de mon stage, s'est tenu le workshop estival du projet SEUS, figure 9, qui avait lieu à Turku. Le projet réunissant des entreprises de plusieurs pays européens, tels que l'Espagne, l'Allemagne, la Norvège ou encore la Finlande, le lieu du workshop alterne chaque année entre ces différents pays. J'ai donc eu la chance d'assister à cette réunion et d'observer son déroulement. Bien que je n'aie pas compris toutes les présentations des différentes entreprises, j'ai appris de nombreuses choses et, surtout, j'ai pu observer les différentes relations entre les participants.

C'était une expérience vraiment enrichissante, d'autant plus que Filip Ginter a présenté le travail que j'avais réalisé devant tous les participants industriels, figure 10. C'était l'une des présentations qui a suscité le plus de réactions, probablement grâce à la démonstration et à la curiosité autour de ces nouvelles technologies d'IA. Cela m'a permis d'échanger avec de nombreuses personnes après la présentation et de répondre à des questions plus ou moins pointues. C'était un bon défi pour tester mes connaissances sur mon sujet.



(a) Photo prise à l'université après les présentations des académiques



(b) Photo prise aux bureaux de Cadmatic après une démonstration de leur application de suivi de chantier avec un casque de réalité virtuelle

FIGURE 9 – Photos du workshop



FIGURE 10 – Filip Ginter présentant le système de RAG

3.4.3 Journal Club

Tous les deux lundis, les membres de TurkuNLP se réunissent pour un journal club au cours duquel certains d'entre eux présentent un ou plusieurs articles scientifiques qui les ont particulièrement intéressés. Cet exercice, bien que rigoureux, se déroulait dans une ambiance détendue, à l'image du laboratoire. Les sujets choisis étaient toujours en lien avec le NLP et mettaient en lumière des avancées récentes dans le domaine. J'ai eu l'opportunité de participer à cet exercice deux mois après mon arrivée, encouragé par M. Ginter, mon tuteur, après lui avoir partagé un article qui avait retenu mon attention. Malgré mon appréhension face à la difficulté de m'exprimer sur un sujet complexe en anglais, j'étais heureux de pouvoir me prêter à cet exercice. J'ai ainsi présenté l'article suivant : *Multi-Head RAG : Solving Multi-Aspect Problems with LLMs* [14]. En résumé, cet article propose d'utiliser l'activation des différentes têtes d'attention en revenant une couche en arrière dans le modèle d'embedding pour améliorer la précision de la récupération de documents pertinents dans les requêtes



complexes. La figure 11 résume très bien le principe général du MRAG.

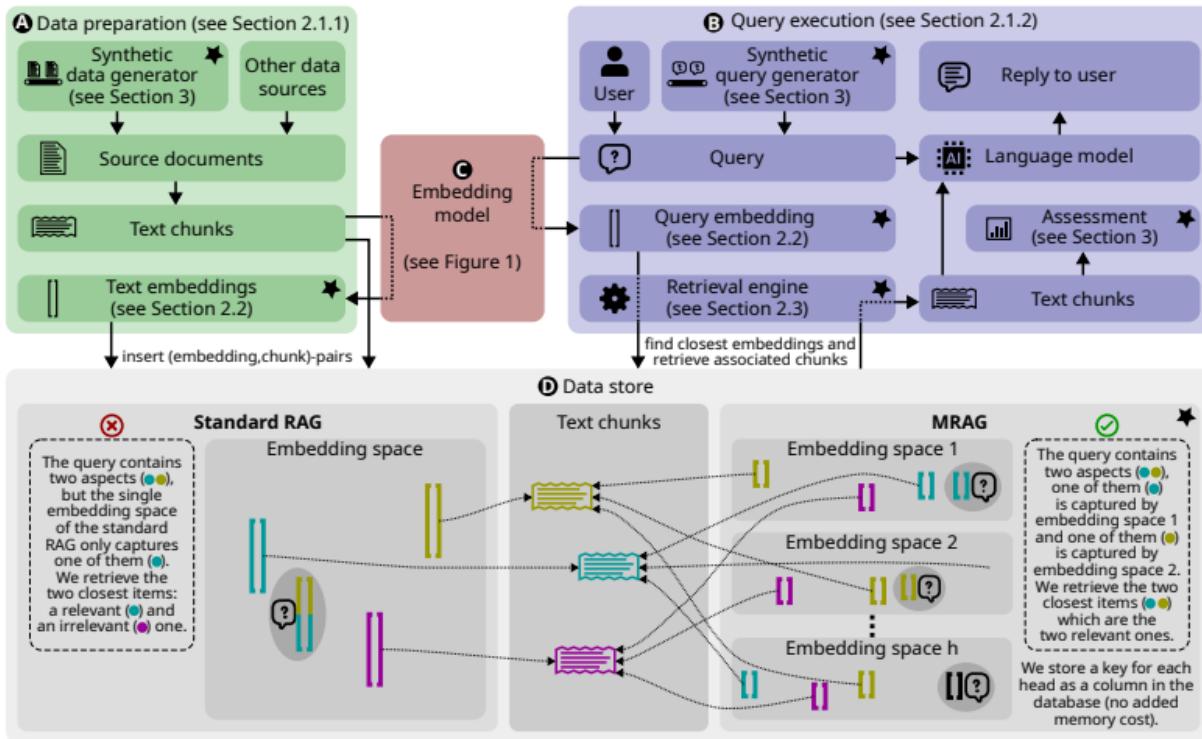


Figure 2: Overview of the MRAG pipeline, consisting of two parts: data preparation **A** and query execution **B**. The embedding model **C** and the data store **D** are used by both parts. The data store **D** contains text embeddings **II** linking to text chunks **■■■** reflecting three different aspects (cyan, magenta, yellow). Blocks marked by a star ***** are a novelty of this work.

FIGURE 11 – Schéma expliquant le principe du MRAG comparé au RAG classique

3.4.4 Iron Age Seminar

C'est malheureusement sans photos à l'appui que je rédige ce paragraphe. M. Ginter m'avait invité à ce séminaire réunissant le laboratoire d'archéologie et les différents acteurs de leurs projets en cours, notamment TurkuNLP, pour une journée sur le thème de l'âge de fer finlandais. Costumes traditionnels, visite d'une ferme et repas en accord avec le thème étaient au programme, entre les présentations des différents projets menés au cours de l'année. Parmi ces présentations, figurait celle de notre système de RAG. Entre tradition et modernité, cette journée m'a permis d'apprendre des choses intéressantes sur l'histoire de la Finlande.

3.4.5 Summer Days

Pour célébrer la fin de l'année académique, le département d'informatique organise chaque année un grand buffet dans un cadre idyllique, sur l'île de Ruissalo. Cette longue île située au sud-ouest de Turku a échappé à la construction d'un aéroport pour devenir un espace naturel protégé. C'est l'occasion pour toutes les équipes du département de se retrouver



pour un moment convivial, figure 12. Je ne m'attendais pas à un tel événement. Entre la nourriture excellente et les fantastiques moments partagés avec mes collègues du laboratoire, cette journée a été un véritable succès.



FIGURE 12 – Où est Charlie ? *Version Antonin Larvor*

Dans la même semaine, tous les membres de TurkuNLP se sont retrouvés pour un barbecue dans le jardin de Veronika Laippala, une membre du laboratoire, que je salue et remercie pour sa gentillesse. Encore une fois, ce fut une merveilleuse journée qui m'a permis d'en apprendre davantage sur mes collègues et amis du laboratoire.

Je termine ces derniers paragraphes sur une note plus légère, mais il me tenait à cœur de souligner l'importance de ces moments, qui transforment un simple stage en une expérience de vie inoubliable.

4 Conclusion

Ce stage au sein de TurkuNLP a été une expérience extrêmement enrichissante, tant sur le plan professionnel que personnel. La réalisation d'un système de Retrieval Augmented Generation (RAG) m'a permis de développer des compétences techniques en programmation et en NLP, tout en découvrant des technologies de pointe telles que les modèles d'embeddings, le transformer et les modèles de Langage à Grande Échelle (LLM). Ce projet m'a permis pour la première fois de gérer un projet dans son intégralité, de la gestion du serveur et de son installation via des commandes UNIX, de l'implémentation des différents programmes et de leur connectivité tels que python html css et javascript ou encore du suivi de version grâce à github.

Au-delà des aspects techniques, j'ai appris à collaborer dans un environnement de recherche académique international, à gérer un projet de manière autonome et à communiquer mes résultats. Les perspectives d'amélioration du système RAG offrent de nombreuses opportunités pour de futures recherches, et je suis enthousiaste à l'idée de voir comment ces travaux évolueront. Enfin, ce stage a non seulement consolidé mes connaissances et compétences, mais a également renforcé ma passion pour le NLP et l'intelligence artificielle.



5 Bibliographie

Références

- [1] Tomas MIKOLOV et al. *Efficient Estimation of Word Representations in Vector Space*. 2013. arXiv : 1301.3781 [cs.CL]. URL : <https://arxiv.org/abs/1301.3781>.
- [2] Tom B. BROWN et al. « Language Models are Few-Shot Learners ». In : (). arXiv : 2005.14165. URL : <https://arxiv.org/abs/2005.14165>.
- [3] Ashish VASWANI et al. « Attention Is All You Need ». In : *CoRR* abs/1706.03762 (2017). arXiv : 1706.03762. URL : <http://arxiv.org/abs/1706.03762>.
- [4] Wikipedia CONTRIBUTORS. *Recurrent Neural Network — Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/wiki/Recurrent_neural_network. Accessed : 2024-07-17. 2024.
- [5] Wikipedia CONTRIBUTORS. *Convolutional Neural Network — Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/wiki/Convolutional_neural_network. Accessed : 2024-07-17. 2024.
- [6] Grant Sanderson / 3BLUE1BROWN. *But what is a GPT? Visual intro to Transformers — Deep learning, chapter 5*. YouTube video. Accessed : 2024-04-23. 2024. URL : <https://www.3blue1brown.com/lessons/gpt>.
- [7] Grant Sanderson / 3BLUE1BROWN. *Visualizing Attention, a Transformer’s Heart — Chapter 6, Deep Learning*. YouTube video. Accessed : 2024-04-23. 2024. URL : <https://www.3blue1brown.com/lessons/gpt>.
- [8] Wikipedia CONTRIBUTORS. *Softmax function — Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/wiki/Softmax_function. Accessed : 2024-08-08. 2024.
- [9] Nelson ELHAGE et al. « A Mathematical Framework for Transformer Circuits ». In : *Transformer Circuits Thread* (2021).
- [10] Wikipedia CONTRIBUTORS. *Multilayer Perceptron — Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/wiki/Multilayer_perceptron. Accessed : 2024-08-08. 2024.
- [11] ELASTIC. *Qu'est-ce que la récupération d'informations ?* URL : <https://www.elastic.co/fr/what-is/information-retrieval>.
- [12] Dr Julija BAINIAKSINA. *How I built a Simple Retrieval-Augmented Generation (RAG) Pipeline*. URL : <https://medium.com/@drjulija/what-is-retrieval-augmented-generation-rag-938e4f6e03d1>.
- [13] Jaime CARBONELL et Jade STEWART. « The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries ». In : *SIGIR Forum (ACM Special Interest Group on Information Retrieval)* (juin 1999). DOI : 10.1145/290941.291025.
- [14] Maciej BESTA et al. *Multi-Head RAG : Solving Multi-Aspect Problems with LLMs*. Juin 2024. arXiv : 2406.05085.
- [15] Antonin LARVOR. *RAG web app*. <https://github.com/TurkuNLP/RAG-web-app.git>. 2024.



6 Annexes

Regulation 7.1

1 The probability of surviving after collision damage to the ship's hull is expressed by the index A . Producing an index A requires calculation of various damage scenarios defined by the extent of damage and the initial loading conditions of the ship before damage. Three loading conditions should be considered and the result weighted as follows:

$$A = 0.4A_s + 0.4A_p + 0.2A_l$$

where the indices s , p and l represent the three loading conditions and the factor to be multiplied to the index indicates how the index A from each loading condition is weighted.

2 The method of calculating A for a loading condition is expressed by the formula:

$$A_c = \sum_{i=1}^{t_m} p_i [v_i s_i]$$

2.1 The index c represents one of the three loading conditions, the index i represents each investigated damage or group of damages and t is the number of damages to be investigated to calculate A_c for the particular loading condition.

2.2 To obtain a maximum index A for a given subdivision, t has to be equal to T , the total number of damages.

3 In practice, the damage combinations to be considered are limited either by significantly reduced contributions to A (i.e. flooding of substantially larger volumes) or by exceeding the maximum possible damage length.

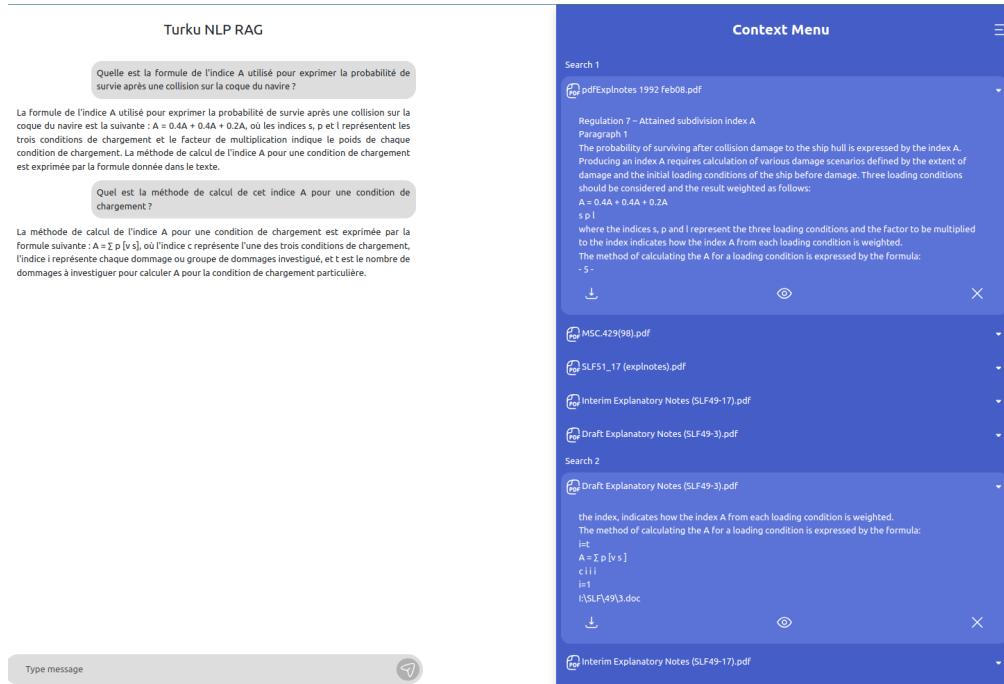
4 The index A is divided into partial factors as follows:

p_i The p factor is solely dependent on the geometry of the watertight arrangement of the ship.

v_i The v factor is dependent on the geometry of the watertight arrangement (decks) of the ship and the draught of the initial loading condition. It represents the probability that the spaces above the horizontal subdivision will not be flooded.

s_i The s factor is dependent on the calculated survivability of the ship after the considered damage for a specific initial condition.

(a) Information ciblée



The screenshot shows a user interface for a Natural Language Processing system. On the left, there is a text input field containing a question in French: "Quelle est la formule de l'indice A utilisé pour exprimer la probabilité de survie après une collision sur la coque du navire ?". Below this, there is a detailed answer in French explaining the formula $A = 0.4A_s + 0.4A_p + 0.2A_l$ and the meaning of the indices s , p , and l .

On the right, there is a "Context Menu" window with a blue header. It contains a search bar at the top. Below it, there is a list of PDF files related to the topic. The first item in the list is "Regulation 7 – Attained subdivision index A" with a preview of the text content. Other items include "MSC.429(98).pdf", "SLF51_17 (explnotes).pdf", "Interim Explanatory Notes (SLF49-17).pdf", "Draft Explanatory Notes (SLF49-3).pdf", and "Interim Explanatory Notes (SLF49-3).pdf". There are also two "Search" sections labeled "Search 1" and "Search 2" with their respective results.

(b) Réponses et informations extraites

FIGURE 13 – Recherche d'une formule dans la base de données SEUS



Turku NLP RAG

Parle moi des rites funéraires dans les bateaux en russie

La tradition des funérailles dans un bateau ou un navire est une pratique funéraire des populations d'Europe du Nord de l'âge du Fer et du début du Moyen Âge en Russie. Ces ritues comprennent l'inhumation, la crémation in situ et la crémation sur le côté, tous unis par l'utilisation d'embarcations ou de leurs parties. En Russie moderne, une douzaine de sites de ce type ont été découverts, principalement à Moscou et dans les environs, avec une prévalence plus grande dans les nécropoles d'élite et les monastères.

Quels sont ces sites qui ont été découverts ?

Certains des sites découverts en Russie comprennent le complexe Ust-Kova, les sites de Chadobets, Sergushkin 3, Otiko, l'estuaire de la rivière Koda, le site de la rivière Taseyev, Skorodumskiy Byk sur le plateau d'Izhora, Pritchude, le lac Kubensky et la rivière Kostroma, ainsi que des sites dans le sud-ouest de Belozero. D'autres sites ont été identifiés le long de la rivière Oka moyenne, dans l'interfluve Volga-Klyazma, et à l'est de la crête Klin-Dmitrovskaya.

Parle moi plus en détail du complexe Ust-Kova

Je ne dispose pas d'informations spécifiques sur le complexe Ust-Kova.

Qu'est-ce que le site d'Ust-Kova ?

Le site d'Ust-Kova était un grand centre commercial et de production entre les 11e et 16e siècles, couvrant une superficie de 15 hectares. Les activités économiques médiévales pratiquées à Ust-Kova comprenaient l'artisanat, le commerce et la guerre. Des preuves matérielles ont également été trouvées concernant les fonctions administratives et défensives du site.

Quel type de commerce pouvait-on trouver sur le site d'Ust-Kova ?

Le site d'Ust-Kova était un centre commercial important qui facilitait le commerce de produits agricoles et artisanaux entre les villes et les zones rurales environnantes. Les villes de la Horde d'Or fournissaient à la campagne des produits artisanaux de haute technologie tels que la céramique irriguée, la verrerie, la Fonte et divers produits métalliques non ferreux. Les échanges commerciaux étaient également facilités par des relations de marchandises-monnaie, comme en témoignent les découvertes de pièces de monnaie de la Horde d'Or dans des sites ruraux et des tumulus nomades.

(a) Conversation 1

FIGURE 14 – Recherches dans la base de données d'Archéologie

Quels sont les sites archéologiques les mieux préservés de Russie ?

Les sites archéologiques les mieux préservés en Russie sont ceux qui ont des zones de protection spécifiques créées pour eux, comme la ville de Zvenigorod, les anciens établissements de Rozhan, Rostislav, Vshchizh, ainsi qu'une série de monuments sur le lac Volgo. Ces zones de protection sont considérées comme le mécanisme le plus efficace pour la préservation du patrimoine archéologique. Le personnel du Département des fouilles de protection a participé à la création de ces zones de protection.

Qu'est-ce qu'on peut trouver à Zvenigorod ?

À Zvenigorod, on peut trouver des antiquités historiques et culturelles, des monuments anciens, des églises et des sites archéologiques. La région de Zvenigorod est riche en découvertes archéologiques, notamment des trouvailles datant des XV-XVI^e siècles. Des études archéologiques ont été menées dans la région pour explorer les sites historiques et culturels de Zvenigorod.

quel genre de trouvailles y ont été faites ?

Des trouvailles archéologiques comprenant des églises anciennes, des nécropoles, des artefacts du XV-XVI^e siècle, des objets religieux et des éléments de l'architecture ancienne ont été découverts à Zvenigorod. Les fouilles ont également révélé des éléments de la culture archéologique moscovite ancienne, des artefacts funéraires et des vestiges de villages princiers anciens. Les découvertes comprennent des objets tels que des pièces d'argent, des anneaux et des outils anciens.

Qu'est-ce que la culture moscovite ancienne trouvée à Zvenigorod ?

La culture moscovite ancienne trouvée à Zvenigorod comprend des éléments architecturaux, des artefacts funéraires, des objets religieux et des vestiges de villages princiers anciens. Les découvertes archéologiques comprennent des pièces d'argent, des anneaux, des outils anciens et des éléments de la culture matérielle de l'époque. Ces trouvailles offrent un aperçu de la vie et de la culture des habitants de la région à l'époque de la culture moscovite ancienne.

Qu'est-ce que la culture moscovite ?

La culture moscovite fait référence à la culture et à la civilisation qui se sont développées dans la région de Moscou et de ses environs au cours de l'histoire. Elle englobe les traditions, les coutumes, l'architecture, l'art, la religion et d'autres aspects de la vie quotidienne des habitants de la région. La culture moscovite a joué un rôle important dans la formation de l'identité russe et a laissé une empreinte significative dans l'histoire et la culture de la Russie.

(b) Conversation 2



Antonin Larvor est un étudiant de la promotion 2025 qui a réalisé un compte-rendu de TP sur la conception d'une Interface Homme-Machine de traitement d'image sous Matlab. Il a également travaillé sur une transformation linéaire d'une image.

(a) Exemple 1 : Texte généré



(b) Exemple 1 : Source



(c) Exemple 2

FIGURE 15 – Recherches dans la base de données locale



To be filled in by the host organization tutor and returned by August 19th to ecole-entreprise@seatech.fr

Host organization: University of Turku, Finland

Student: Antonin Larvor

Subject of internship: Retrieval-augmented generation (RAG)

Assessment : Skill development											
<p>The evaluation grids is consistent with the study skills framework. The 6 skills are developed over the three years of study. Please evaluate the student according to the different criteria.</p> <p style="text-align: center;">Reference for cotation</p> <table border="1" style="margin-left: auto; margin-right: auto; border-collapse: collapse;"> <tr> <td style="padding: 2px;">(A) Excellent</td> <td style="padding: 2px;">Beyond expectations (beyond students of the same level)</td> </tr> <tr> <td style="padding: 2px;">(B) Suitable</td> <td style="padding: 2px;">Meets expectations (within the norm for students at the same level)</td> </tr> <tr> <td style="padding: 2px;">(C) Insufficient</td> <td style="padding: 2px;">Does not meet expectations (below students at the same level)</td> </tr> <tr> <td style="padding: 2px;">(N) Not applicable</td> <td style="padding: 2px;">Criterion not assessable in the context of the internship</td> </tr> </table>				(A) Excellent	Beyond expectations (beyond students of the same level)	(B) Suitable	Meets expectations (within the norm for students at the same level)	(C) Insufficient	Does not meet expectations (below students at the same level)	(N) Not applicable	Criterion not assessable in the context of the internship
(A) Excellent	Beyond expectations (beyond students of the same level)										
(B) Suitable	Meets expectations (within the norm for students at the same level)										
(C) Insufficient	Does not meet expectations (below students at the same level)										
(N) Not applicable	Criterion not assessable in the context of the internship										
Skill	Criterion	Assessment	Comment								
Come up with engineering solutions	Analyse need	A									
	Answer need	A									
	Use appropriate tools	B									
	Document choices and sources	B									
Implement solution	Analyse existing solution	A									
	Propose a new solution	B									
	Use the principles of continuous improvement	N									
	Write a scientific document and technical support	B									
Develop R&D steps	Carry out continuous technological surveillance	A									
	Formulate hypotheses	A									
	Propose an experimental approach, a protocol or a model	A									
	Adopt an innovative	A									

Assessment Form 2023-24

	approach		
--	----------	--	--

Manage engineering projects	Get involved in or lead one or more stages of a project	N	
	Take into account the overall management of organizations or the operating, economic or legal	B	
	Use project management and collaborative tools	N	
	Identify or mobilize resources appropriate	B	
Coach a team	Integrate and collaborate	B	
	Team working multidisciplinary and/or international	A	
	Ensure leadership animation	N	
	Communicate (written and oral) in adapted manner	B	
Act like a responsible professional	Take into account ethical and societal issues	B	
	Take responsibility for his actions and decision	A+	Remarkably independent!
	Take a critical look at the meaning of the activity carried out	A	
	Being open to learning new things	A+	Remarkably capable and fast learner!
Overall assessment of the intern and recommendations for occupying an engineering position (it is recommended to discuss this evolution with the intern).			
<p>If you have detected an intern's personal qualities such as his/her ability overcome difficulties, accept criticism, express points of view in an argumentative manner, or any other aspect, you can note them for his/her benefit</p> <p>Antonin is an excellent intern and I am sorry to see him go. Of especial note is Antonin's ability to very fast learn new concepts and skills, and his remarkable independence in doing so. I also appreciated Antonin's enthusiastic and serious attitude towards work – he was always present and always did what was asked. As a suggestion for further self-development, I would suggest putting effort in spoken English comprehension by e.g. listening more to varied spoken English in free time, especially if planning an international career.</p>			

13.8.2024, Prof. Filip Ginter, Department of Computing, University of Turku
Date, stamp of the host organization, name and signature of the tutor:



**TURUN
YLIOPISTO
UNIVERSITY
OF TURKU**

Tämä dokumentti on allekirjoitettu sähköisesti Turun yliopiston UTUsign-järjestelmällä

This document has been electronically signed with UTUsign system of the University of Turku

Päiväys / Date: 14.08.2024 14:46:42 (UTC +0300)

Filip Ginter

professori

Turun yliopisto

Organisaation varmentama (UTU-käyttäjätunnus)
Certified by organization (UTU user account)

Organisaation varmentama

Fiche d'évaluation étudiant

Ce document est à insérer dans le rapport. Le rapport est à fournir pour le 19 aout dernier délai.

Organisme d'accueil :

Sujet du stage :

Evaluation qualitative du stage	I	S	B	E	N	Commentaires
Qualité de l'entreprise				X		
L'entreprise offre-t-elle un contexte propice à une carrière d'ingénieur ?				X		
L'entreprise connaît-elle l'école (accueil de stagiaire, embauche, relations autres...) ?					X	
L'entreprise a-t-elle mis à votre disposition les moyens nécessaires pour réaliser votre mission (documents, éléments d'information, matériels) ?				X		
Qualité de la missions						J'ai tout appris sur place même si c'était dans le domaine des data sciences
Vos missions étaient-elles en rapport avec votre formation ?		X				
Les missions effectuées étaient-elles bien celles définies au départ ?				X		
Qualité de l'encadrement				X		
Votre tuteur organisme d'accueil a-t-il pris le temps de vous présenter le fonctionnement de la structure et l'équipe ?				X		
Votre tuteur organisme d'accueil vous a-t-il aidé et conseillé quand cela était nécessaire ?				X		
Votre enseignant référent vous a-t-il aidé et conseillé lorsque cela était nécessaire ?				X		

Explication des cotations	
I	(I) Insuffisant
S	(S) Suffisant
B	(B) Bien
E	(E) Excellent
N	(N) Non applicable

NB : Dans le but d'alléger la lecture du document, le genre masculin est utilisé sans discrimination pour le genre masculin et féminin.

Fiche d'évaluation étudiant

Autoévaluation : développement des compétences et trajectoire professionnelle

En prenant un peu de recul sur votre activité durant le stage pensez-vous avoir travaillé / développé certaines des compétences du référentiel de la formation Seatech ? Lesquelles ? Pourquoi et comment ? D'autres compétences ?

Compétence	Critère	Commentaire
Concevoir des solutions d'ingénierie	Analyser le besoin	Le besoin était de concevoir une application de RAG utilisable localement. Cette nouvelle technologie présente un grand intérêt pour les entreprises privées qui ont des données confidentielles.
	Répondre au besoin	J'ai conçu l'application de RAG demandé par mon tuteur
	Utiliser les outils appropriés	Python, HTML, CSS, JavaScript, Linux...
	Documenter ses choix et ses sources	Gros travail de documentation, j'ai notamment découvert les papiers scientifiques.
Mettre en œuvre des solutions	Analyser et améliorer une solution existante	Découverte de toutes les technologies déjà existante, création de la mienne pour répondre parfaitement au besoin
	Proposer une solution nouvelle	Création de l'application
	Utiliser les principes de l'amélioration continue	
	Rédiger un document scientifique et technique d'appui	J'ai essayé au maximum de d'expliquer les notions fondamentales utilisées pour mon système RAG et j'ai essayé d'y mettre les formes
Développer une démarche R&D	Réaliser une veille technologique / un état de l'art	Fait
	Formuler des hypothèses	Je discutais en permanence des améliorations possibles avec mon tuteur
	Proposer une démarche expérimentale, un protocole ou un modèle	
	Adopter une démarche d'innovation	

Fiche d'évaluation étudiant

Piloter des projets d'ingénierie	S'insérer dans ou conduire une ou plusieurs étapes d'un projet	J'ai mené le projet dans son entièreté
	Prendre en compte la gestion globale des organisations ou les règles de fonctionnement, économiques ou juridiques	
	Utiliser les outils de gestion de projet et outils collaboratifs	Gestion des versions avec Github
	Identifier ou mobiliser les ressources appropriées	
Encadrer une équipe	S'insérer et collaborer	Je me suis très bien intégré à l'équipe
	Assurer une responsabilité d'animation	
	Travailler en équipe pluridisciplinaire et/ou internationale	Compte rendu hebdomadaire minimum à mon tuteur
	Communiquer (écrit et oral) de manière adaptée	Très bien
Agir en professionnel responsable	Prendre en compte les enjeux éthiques et sociaux (RSE, DD, RGPD, ...)	
	Assumer la responsabilité de ses actes et décision	
	Porter un regard critique sur le sens de l'activité conduite	
	Être dans une dynamique d'apprentissage	En continu.
Trajectoire professionnelle		
A la suite de votre stage, avez-vous confirmé ou affiné votre projet professionnel d'être ingénieur (métier plus précis, secteur, contexte ou type d'entreprise, ...) ? Si oui, quelles actions pensez-vous devoir entreprendre pour y arriver (renforcer certaines connaissances, développer certaines compétences, lesquelles)?		
<p>J'ai absolument adoré le domaine du NLP et, plus largement, de l'IA. Je suis désormais convaincu que je veux travailler dans ce domaine en pleine explosion tant il me passionne. J'ai adoré la recherche et j'ai dans un coin de ma tête l'idée du doctorat. J'aimeraï beaucoup travailler dans une entreprise proposant des technologies d'IA à la pointe. Les entreprises avec des dynamiques assez jeunes ou les startup m'intéressent beaucoup et j'aimeraï regarder de ce côté pour mon stage de fin d'étude.</p>		
Date et signature de l'étudiant : 19/08/2024 Antonin LARVOR		

Fiche d'évaluation étudiant