Haute Ecole

« ICHEC – ECAM – ISFSC »

Enseignement supérieur de type long de niveau universitaire

Paramétrisation et intégration d'un chatbot au sein d'Europ Assistance

Mémoire présenté par :

Arnaud TILMANT

Pour l'obtention du diplôme de

Master Ingénieur commercial

Année académique 2024-2025

Promoteur :

Thierry VAN DEN BERGHE

Boulevard Brand Whitlock 6 - 1150 Bruxelles

REMERCIEMENTS

Je tiens à remercier mon promoteur, Monsieur Thierry Van den Berghe, enseignant à l'ICHEC, qui m'a conseillé et suivi tout au long de l'élaboration de ce mémoire.

Je remercie tous les collègues avec qui j'ai eu l'occasion de travailler au sein d'Europ Assistance, en particulier Monsieur Heykel Boulares et les membres de l'équipe CCA.

Enfin je remercie mes parents pour leur soutient durant la rédaction de ce mémoire.

ENGAGEMENT ANTI-PLAGIAT

« Je soussigné, TILMANT Arnaud, 2025 déclare par la présente que le travail ci-joint respecte les règles de référencement des sources reprises dans le règlement des études en signé lors de mon inscription à l'ICHEC (respect de la norme APA concernant le référencement dans le texte, la bibliographie, etc.) ; que ce travail est l'aboutissement d'une démarche entièrement personnelle ; qu'il ne contient pas de contenus produits par une intelligence artificielle sans y faire explicitement référence. Par ma signature, je certifie sur l'honneur avoir pris connaissance des documents précités et que le travail présenté est original et exempt de tout emprunt à un tiers non-cité correctement. »

18/05/2025, Tilmant Arnaud, 231029

TEXTE DE CONFORMITÉ AVEC L'IA

Je soussigné, Tilmant Arnaud, 231029, déclare sur l'honneur les éléments suivants concernant l'utilisation des intelligences artificielles (IA) dans mon travail :

Type d'assistance		Case à cocher
Aucune assistance	J'ai rédigé l'intégralité de mon travail sans avoir eu recours à un outil d'IA générative.	
Assistance avant la rédaction	J'ai utilisé l'IA comme un outil (ou moteur) de recherche afin d'explorer une thématique et de repérer des sources et contenus pertinents.	X
Assistance à l'élaboration d'un texte	J'ai créé un contenu que j'ai ensuite soumis à une IA, qui m'a aidé à formuler et à développer mon texte en me fournissant des suggestions.	
	J'ai généré du contenu à l'aide d'une IA, que j'ai ensuite retravaillé et intégré à mon travail.	X
	Certains parties ou passages de mon travail/mémoire ont été entièrement été générés par une IA, sans contribution originale de ma part.	
Assistance pour la révision du texte	J'ai utilisé un outil d'IA générative pour corriger l'orthographe, la grammaire et la syntaxe de mon texte.	X
	J'ai utilisé l'IA pour reformuler ou réécrire des parties de mon texte.	X
Assistance à la traduction	J'ai utilisé l'IA à des fins de traduction pour un texte que je n'ai pas inclus dans mon travail.	
	J'ai également sollicité l'IA pour traduire un texte que j'ai intégré dans mon mémoire.	
Assistance à la réalisation de visuels	J'ai utilisé une IA afin d'élaborer des visuel, graphiques ou images.	
Autres usages		

Je m'engage à respecter ces déclarations et à fournir toute information supplémentaire requise concernant l'utilisation des IA dans mon travail, à savoir : J'ai mis en annexe les questions posées à l'IA et je suis en mesure de restituer les questions posées et les réponses obtenues de l'IA. Je peux également expliquer quel le type d'assistance j'ai utilisé et dans quel but.

Fait à Bruxelles, le 18/05/2025 Signature : Arnaud Tilmant, 231029

TABLE DES MATIÈRES

TABLE DES FIGURES

TABLE DES TABLEAUX

INTRODUCTION

Dans un monde où l'intelligence artificielle se répand rapidement et modifie les manières de travailler, Europ Assistance y a vu l'opportunité d'améliorer sa qualité de service afin de répondre aux mieux aux exigences élevées du secteur de l'assistance.

En cas de sinistre, ce sont les chargés d'assistance qui interviennent en première ligne. La manière dont les appels sont pris en charge par cette équipe est donc particulièrement importante pour l'image de l'entreprise. Cependant, la complexité grandissante des contrats, le turnover élevé et la nécessité de renforcer cette équipe avec des étudiants lors des périodes d'été (pics d'appels) peut entraîner des erreurs ou des délais dans le traitement des dossiers.

La mise en place d'un chatbot pour les chargés d'assistance doit permettre à cette équipe un accès plus facile aux informations, en particulier pour la vérification des couvertures et l'identification des procédures d'assistance. Des réponses contextuelles lors des appels peut également permettre de diminuer le temps de formation des nouveaux employés et des étudiants.

Ce mémoire présente tout d'abord l'entreprise et certains concepts importants, en particulier les concepts de l'intelligence artificielle utilisés dans les chatbots. Vient ensuite une description des éléments clés réalisés durant le projet. Enfin, un chapitre est consacré au bilan du projet et aux perspectives d'évolution de ce projet pilote.

La technologie au centre de la solution mise en œuvre est le « Retrieval Augmented Generation » ou RAG qui permet à un chatbot de récupérer les données pertinentes et de générer sur cette base une réponse contextualisée.

1 CONTEXTUALISATION DU PROJET

1.1 PRÉSENTATION DE L'ENTREPRISE

1.1.1 Historique

Fondée en France en 1963, Europ Assistance (EA) est pionnière dans le domaine de l'assistance, offrant une assistance aux personnes autant dans la vie au quotidien que lors de leurs voyages. En 1964, EA Belgique (EAB) devient la première filiale d'EA Groupe.

Rapidement, elle augmente sa couverture, se limitant dans un premier temps à l'Europe pour ensuite l'étendre à l'Amérique du Nord en 1967, puis au Brésil et au Kenya en 1977 et pour finir en Asie. En parallèle, elle ouvre d'autres filiales en Europe entre 1970 et 1980, puis aux Etats-Unis, en Afrique du Sud et en Chine entre 1980 et 1990.

En 2025, EA continue de renforcer sa présence mondiale. L'entreprise est désormais présente dans plus de 200 pays et territoires, s'appuyant sur un réseau de 750 000 partenaires agréés et 57 centres d'assistance. Avec ses 12 000 employés, elle dispose de bureaux dans 39 pays, assurant une couverture mondiale pour plus de 95 millions de clients dont plus de 2 millions en Belgique (Europ Assistance 4, 2025).

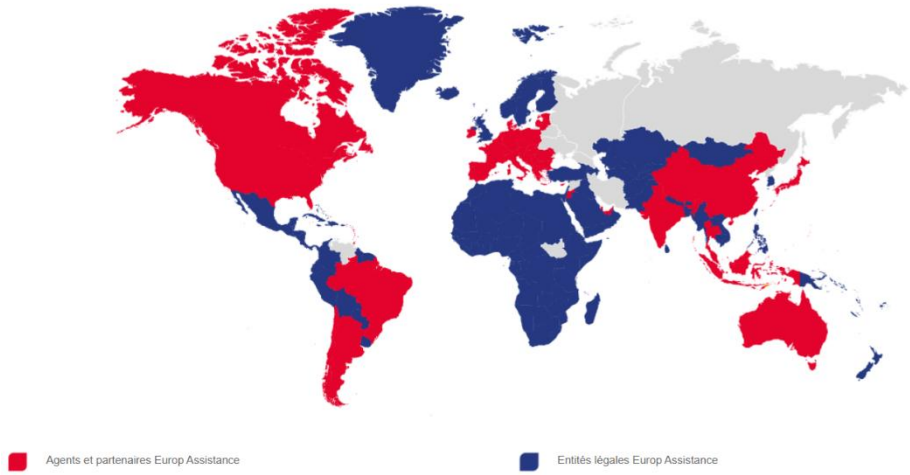


Figure 1 : EA couverture (Europ Assistance 3, 2025)

1.1.2 Vision, mission et valeurs de l'entreprise

Europ Assistance définit sa mission comme « From distress to relief anytime anywhere », soit « Du stress au soulagement, partout et à tout moment ». Celle-ci n'a pas changé depuis sa création. Son but est de simplifier la vie de ses clients, en les aidant à surmonter des situations difficiles et stressantes, tout en leur offrant confort et sécurité au quotidien.

Sa vision, « To be the most reliable care company in the world », soit d'être la compagnie d'assistance la plus fiable au monde.

Afin d'y parvenir, chaque employé d'Europ Assistance doit incarner ses valeurs qui sont représentées par « You Live, WE Care ». Le « We » représentant les employés et le « CARE » étant l'acronyme pour Caring (Bienveillant), Available (Disponible), Reliable (Fiable) et Easy to work with (Collaboratif) (Europ Assistance 3, 2025).

1.1.3 Activités et produits

Bien qu'Europ Assistance soit une compagnie d'assistance, elle propose également des produits d'assurance liés aux voyages.

L'assurance intervient pour indemniser un assuré après qu'un sinistre se soit produit, tandis que l'assistance vise à fournir un service pour aider une personne à surmonter une situation difficile.

EA propose trois grands domaines de services : l'assistance technique, l'assistance médicale et l'assistance à domicile. L'assistance technique est la plus répandue et couvre tout ce qui concerne le véhicule du client, incluant le dépannage, le remorquage, le rapatriement, ainsi que la mise à disposition d'un taxi, d'un chauffeur ou d'un véhicule de remplacement, que ce soit en Belgique ou à l'étranger. L'assistance médicale englobe des services liés à la santé du client, tels que le rapatriement médical, la prise en charge des frais médicaux, ainsi que des missions de secours à l'étranger. L'assistance à domicile couvre des interventions comme la sécurisation, la plomberie, l'électricité, et même l'hébergement temporaire en Belgique.

En ce qui concerne l'assurance, elle couvre les annulations, les interruptions de voyage, mais également les vols et pertes de bagages.

EAB se distingue des autres filiales du groupe en proposant des contrats combinant tous ses produits, tandis que le reste du groupe se limite à un contrat par service ou produit d'assurance.

En plus de proposer ses produits d'assistance (produits B2C ou Business to Customer), EAB a également des partenaires pour lesquels elle fournit les services d'assistance (produits B2B ou Business to Business). Elle opère par exemple pour Belfius assistance (technique, médicale et home), Deloitte assistance, Nissan, Porsche, Ferrari (technique), Baloise assistance (Home). Ces partenaires représentent 39 % des produits d'EAB (Europ Assistance 2, 2024).

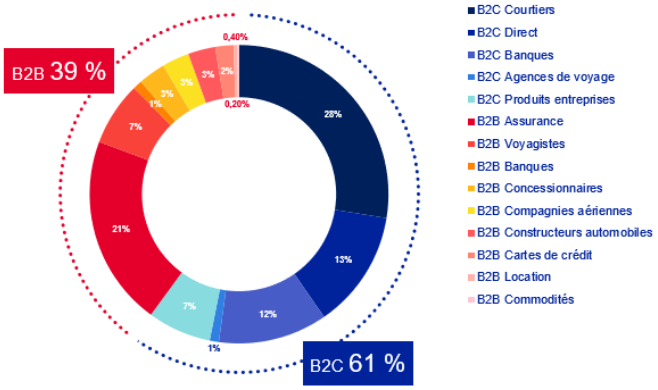


Figure 2 : EAB clients (Europ Assistance 2, 2024)

1.1.4 Concurrence

Le marché belge est en croissance, augmentant de 39 % sur les 5 dernières années pour atteindre un chiffre d'affaires de 368,7 millions d'euros en 2023 (Assuralia 2, 2024).

Le marché est assez compétitif. EAB, avec un chiffre d'affaires de 88 millions en 2023 et 19 % de parts du marché, se place en deuxième position, juste derrière AXA (Assuralia 1, 2024).

EAB se positionne comme marque premium pour la qualité de son service. Ses principaux concurrents sont VAB, AXA, Touring, AG et Allianz. VAB est la compagnie qui se rapproche le plus d'EAB en proposant une assistance technique et médicale à l'internationale, tandis que Touring se concentre sur l'assistance technique et l'assurance voyage en Europe. AXA, AG et Allianz offrent des services d'assistance et d'assurance à l'échelle mondiale (Europ Assistance 2, 2024).

Figure 3 : EAB Positionnement (Europ Assistance 2, 2024)

1.1.5 Moyens et indice de performance

EAB propose un service 24/7, avec plus de 500 000 appels par an, soit plus de 1350 appels par jours (Europ Assistance 2, 2024).

Le principal indicateur de performance (KPI) d'Europ Assistance Belgique est la satisfaction du client. Elle est mesurée par un « Net Promoter Score » (NPS). A partir de 60, un NPS est considéré comme excellent et celui d'Europ Assistance calculé en interne est de 62. Une évaluation externe réalisée par l'entreprise eKomi lui assigne un sceau de qualité Argent avec une note de 4.5/5 (eKomi, 2025). Malgré ces résultats très positifs, EAB s'efforce d'améliorer encore ce score car le client est au centre de ses priorités (Europ Assistance 2, 2024).

1.1.6 Structure et environnement de stage

Europ Assistance a subi une restructuration de ses filiales. Désormais, elle est structurée en plusieurs unités géographiques, supervisées par la Holding. En Europe, on retrouve l'unité « NE » qui comprend l'Allemagne, l'Autriche, la Suisse, la Belgique et le Luxembourg, l'unité « CE » qui contient uniquement la France et l'unité « SE » qui se compose de l'Espagne, de l'Italie et du Portugal. Chaque unité est composée de plusieurs départements : OPS & IT, Marketing et Business Development (Europ Assistance, 2024).

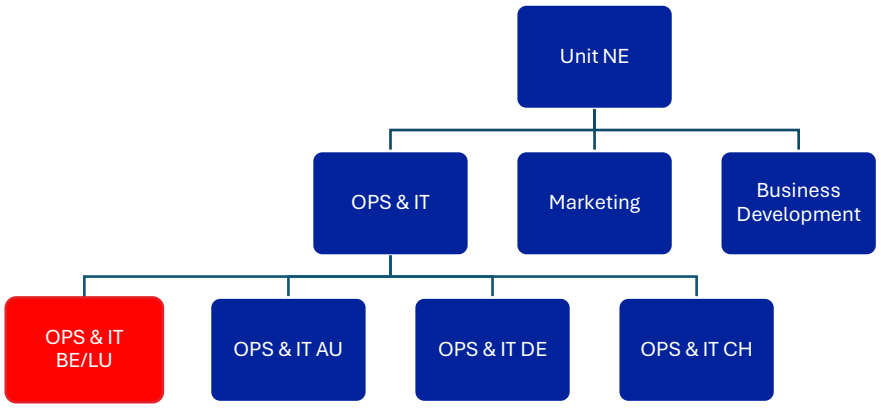


Figure 4 : Organigramme Unit NE (Europ Assistance 1, 2025)

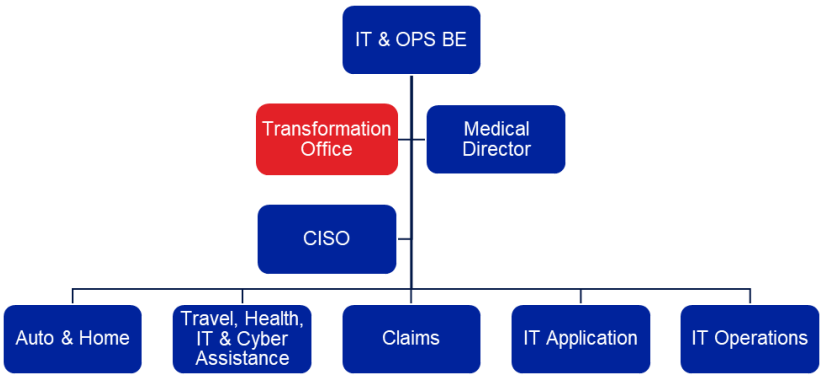


Figure 5 : Organigramme département OPS & IT (Europ Assistance 1, 2025)

Le département OPS & IT est composé de plusieurs services. Pour la partie « OPS », il est composé principalement du service d'assistance, divisé lui-même en « business lines » (« Auto & Home » et « Travel, Health, IT & Cyber assistance ») et du service Claims qui s'occupe des remboursements. La partie IT se compose de « IT Application », qui, comme son nom l'indique, gère les applications dont le site web, les outils de gestion mais aussi les intégrations entre systèmes et de « IT Operations », qui s'occupe de l'ensemble des infrastructures nécessaires au bon fonctionnement de l'entreprise (Europ Assistance 1, 2025).

En support, on retrouve la direction médicale, le CISO, qui supervise les équipes « Risk » et « Legal » et le service « Transformation ». Ce dernier chapeaute le service AI, Digital & Data dont l'équipe CCA (« Competence Center Automation »), en charge du projet, dépend. Il contient également le service « Network & Procurement », responsable de gérer les relations avec les prestataires de services tels que le dépannage, la location de véhicules, les agences de voyage et les hôpitaux, mais aussi le contenu de l'application Athena dans laquelle se trouvent les conditions générales et les procédures (équipe Product) (Europ Assistance 1, 2025).

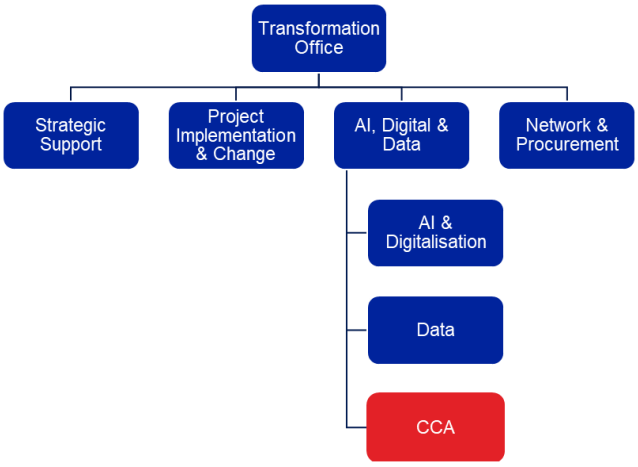


Figure 6 : Organigramme Transformation office (Europ Assistance 1, 2025)

L'équipe CCA rapporte directement au responsable AI, Digital et Data. Hiérarchiquement, il est composé d'une personne en charge du business support et de stagiaires. L'équipe est complétée matriciellement par un business analyst et un développeur rattachés hiérarchiquement à IT Application (Europ Assistance 1, 2025)

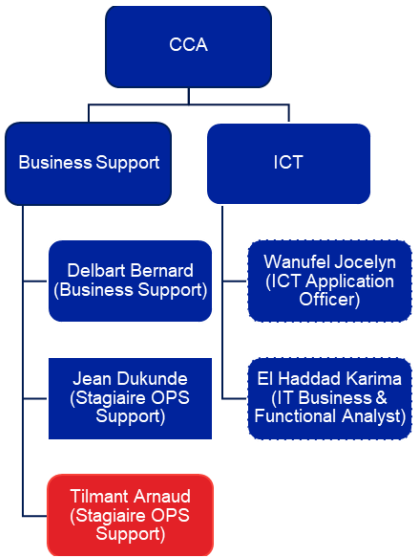


Figure 7 : Organigramme équipe CCA (Europ Assistance 1, 2025)

1.2 ENJEU DU PROJET

Comme indiqué en introduction, Europ Assistance donne beaucoup d'importance à la satisfaction client et dès lors à la qualité du service offert. C'est pourquoi deux des enjeux sont le temps de réaction et la qualité de la réponse apportée. Cette qualité dépend entre autres de la capacité à vérifier quelles sont les couvertures du client et donc le service auquel il a droit.

Les produits proposés par Europ Assistance ou ses partenaires sont de plus en plus complexes pour répondre à la diversité des besoins clients. Le nombre de produits, de couvertures, d'options est en croissance constante. On estime que cela représente aujourd'hui plus de 1000 combinaisons différentes, charge à l'agent d'assistance d'identifier celle qui s'applique au client qu'il a en ligne.

La formation d'un agent se déroule en plusieurs temps. Durant 4 à 6 semaines, il suit des modules de formation et apprend « on the job » en étant suivi de près par une équipe d'encadreurs. Après une évaluation positive, il devient semi-autonome mais peut encore s'adresser à un « learning specialist » en cas de questions. Après une période d'environ deux mois supplémentaires, on évalue si l'agent est suffisamment autonome, ce qui l'autorise à travailler partiellement de chez lui. Les « learning specialists » estiment cependant qu'il faut 6 mois avant d'être réellement autonome.

Cette durée d'apprentissage est problématique si l'on tient compte du taux élevé de turn-over ou du besoin de faire appel ponctuellement à des étudiants.

Concernant le turn-over, 50 % des agents d'assistance ont moins de 2 ans d'ancienneté et 16 % moins de 6 mois, ce qui signifie qu'ils ne sont pas encore totalement autonomes.

Durant l'été, ce sont environ 50 étudiants qui viennent renforcer les équipes. Il n'est évidemment pas possible de les former complètement. Ils suivent donc une formation de 2 à 3 jours avant de rentrer en fonction. On limite leurs interventions à l'assistance technique.

Deux outils principaux sont utilisés par les chargés d'assistance. L'application STAR pour gérer les dossiers d'assistance et l'application Athena pour retrouver la documentation : conditions générales, procédures (plus de 5000 pages HTML) et FAQ.

Même si l'application STAR est vieillissante, c'est surtout la navigation dans Athena qui pose problème : arborescence complexe, mauvaise ergonomie, multitude d'acronymes pour faire référence aux contrats et aux prestations. En outre, le vocabulaire spécifique, à connotation juridique, des conditions générales pose parfois des problèmes d'interprétation. Quant aux procédures, elles ne sont pas traduites systématiquement en français et en néerlandais.

L'enjeu du projet est dès lors de mettre à disposition des chargés d'assistance, qu'ils soient permanents ou temporaires, une information claire et précise, que ce soit lorsqu'ils doivent vérifier une couverture ou identifier la procédure d'assistance à suivre.

2 ETAT DE L'ART

Dans cette section, nous allons parcourir différents concepts, technologies et méthodes utilisés dans le projet.

2.1 INTELLIGENCE ARTIFICIELLE

2.1.1 Définition

« The use or study of computer systems or machines that have some of the qualities that the human brain has, such as the ability to interpret and produce language in a way that seems human, recognize or create images, solve problems, and learn from data supplied to them » (Cambridge Dictionary 1, 2025)

L'intelligence artificielle est un domaine mathématique qui crée des systèmes capables d'imiter des fonctions cérébrales humaines, telles que la compréhension de texte, la prise de décisions et la résolution de problèmes. Elle repose sur des systèmes mathématiques comme le machine learning, qui lui permet d'apprendre automatiquement à partir de données et de traiter de nouvelles situations basées sur ses données d'entraînement (Kavlakoglu et al., 2024).

2.1.2 NLP

Les applications de l'intelligence artificielle sont vastes. Le fonctionnement d'un chatbot consiste en des interactions entre les ordinateurs et le langage humain. Dans ce cadre, c'est le traitement du langage naturel (NLP : Natural Language Processing) qui est utilisé pour traiter, comprendre et générer du texte.

Le NLP permet aux machines de traiter, comprendre et générer du texte via différentes capacités telles que l'analyse syntaxique, l'extraction de texte, la tokenisation (cf infra section 2.1.4.2.1), le « stemming » (réduction des mots à leur forme de base), la « lemmatisation » (« stemming » avec prise en compte du contexte).

La compréhension du langage naturel (NLU : Natural Language Understanding) est une sous-discipline du NLP qui se concentre sur la compréhension du sens et du contexte du langage. Elle permet non seulement une analyse syntaxique et lexicale du texte, mais également une compréhension sémantique, permettant ainsi de saisir le sens, les intentions et les sentiments de celui-ci. Le NLU comprend la reconnaissance des entités nommées et l'analyse des sentiments.

La génération du langage naturel (NLG : Natural Language Generation) est une autre sous-discipline du NLP, qui se concentre sur la création de textes compréhensibles et cohérents. Le NLG permet de produire des réponses textuelles, des résumés, des descriptions et d'autres formes de contenus écrits (Kavlakoglu et al. 2, 2020).

2.1.3 LLM

Les modèles de langage de grande taille (LLM) sont des modèles entrainés sur d'immenses quantités de données avec des techniques de machine learning et d'auto-apprentissage pour comprendre et générer du texte en langage naturel.

Les réseaux de neurones récurrents (RNN), utilisés dans le passé, fonctionnent en séquence en apprenant des neurones précédents. Les architectures LLM actuelles se

basent sur une architecture « Transformer ». Celle-ci utilise des mécanismes d'attention et de « self-attention » qui lui permettent d'apprendre en parallèle et améliorent fortement la compréhension sémantique des données (par rapport au contexte), tant en termes de qualité que de rapidité (Ferrer, 2024).

Le processus d'entrainement d'un LLM se compose de trois phases : la collecte des données, l'entrainement du modèle et le fine-tuning. Lors de la collecte des données, le LLM est exposé à de grandes quantités de données provenant d'une large variété de sources. Ensuite vient la phase d'entrainement, durant laquelle le modèle commence à construire une compréhension du langage grâce à un apprentissage non supervisé. Enfin, la phase de fine-tuning, consiste à entrainer le modèle sur des données plus spécifiques, lui permettant d'affiner ses connaissances et d'augmenter ses performances dans certains domaines particuliers (Microsoft 2, 2025).

2.1.3.1 Types de modèles LLM

Il existe différents modèles de LLM utilisant l'architecture « Transformer » (cf infra section 2.1.4), mais tous n'exploitent pas cette architecture de la même manière.

La première catégorie comprend les modèles basés uniquement sur l'encodeur, tels que BERT (Bidirectional Encoder Representations from Transformer). Ces modèles, qui se concentrent sur l'encodeur, sont principalement utilisés pour la classification de texte, la recherche de réponses dans un texte en fonction d'une question ou la reconnaissance d'entités nommées (NER).

La deuxième catégorie regroupe les modèles unidirectionnels utilisant uniquement le décodeur, comme GPT (Generative Pre-trained Transformer). Ces modèles sont conçus pour générer du texte en se basant uniquement sur les mots précédents.

Enfin, il existe des modèles qui utilisent à la fois l'encodeur et le décodeur, comme BART (Bidirectional and Auto-Regressive Transformers). Ces modèles sont particulièrement utiles pour des tâches nécessitant une compréhension contextuelle approfondie, telles que la complétion de texte à trous, la correction de texte, la génération automatique de résumés ou la traduction, où il est essentiel de comprendre l'ensemble du document.

Dans tous ces modèles, l'encodeur et/ou le décodeur sont d'abord entraînés sur des données. Lors de leur utilisation, ils reçoivent en entrée la requête contenant la question, les informations et le prompt (Esmaielbeiki, 2023).

2.1.3.2 Modèles LLM de type GPT

Les modèles LLM de type GPT (Generative Pre-Trained Transformer) sont caractérisés par trois éléments essentiels : les paramètres d'entrainement, la longueur maximale des tokens et les données d'entrainement.

	GPT-1	GPT-2	GPT-3/ GPT 3,5	GPT-4
Training Parameters	117 million	1.5 billion	175 billion	Unknown
Dataset	BooksCorpus	WebText	CommonCrawl	Public and Private available dataset
Release Date	June 2018	February 2019	GPT-3 was released on June 2020, GPT.3.5 March 2022	March 2023
Maximum Token Length	1024	1024	4096	8192-32,768
NLP tasks	Yes	Yes	Yes	Yes
Image Generation	No	No	No	Yes
Academic and Professional Performance Benchmark	No	No	No	Human Level performance on Bar, Medical, and SAT exam

Figure 8 : Comparaison des versions GPT (Bengesi et al., 2024)

Les paramètres d'entraînement d'un réseau de neurones représentent le nombre de poids et de biais qui y sont répartis. Ces paramètres sont essentiels pour permettre au modèle de traiter efficacement les données. Lors de l'entraînement, ces paramètres sont ajustés afin d'améliorer la capacité du modèle à prédire, comprendre et générer du texte. Plus le nombre de paramètres est élevé, plus le modèle est capable de représenter et de comprendre des relations complexes. Par exemple, GPT-3 dispose de 175 milliards de paramètres, tandis qu'on estime à 1800 milliards, soit dix fois plus, ceux de GPT-4 (ChatGPT info, 2025).

La longueur maximale des tokens fait référence à la quantité de texte que le modèle peut traiter en une seule fois. Les tokens peuvent représenter des mots, des sous-mots ou des caractères. Avec GPT-4, il est capable de gérer des documents dont la taille varie entre 8192 et 32768 tokens, ce qui lui permet de traiter des textes relativement longs en une seule requête. Cette capacité est cruciale pour des tâches où un contexte étendu est nécessaire (Bengesi et al., 2024).

Les modèles GPT sont préentraînés sur de vastes corpus de textes provenant de diverses sources, y compris des livres, des articles scientifiques, des sites web et d'autres documents textuels disponibles sur Internet. Ce préentraînement permet au modèle de comprendre une large gamme de sujets (Bengesi et al., 2024).

2.1.3.3 Limites

Les modèles de langage de grande taille (LLM) présentent plusieurs limites notables, notamment en ce qui concerne la qualité des données d'entraînement et les hallucinations. Des données insuffisantes ou biaisées peuvent entraîner des interprétations erronées et des résultats inexactes. Par exemple, si les données d'entraînement contiennent des biais, le modèle peut reproduire ces biais dans ses réponses. De plus, des données de mauvaise qualité peuvent restreindre la capacité du modèle à généraliser correctement à de nouvelles informations, ce qui nuit à sa fiabilité.

Les hallucinations constituent un autre problème majeur des LLM. Elles se manifestent par des résultats incorrects qui s'écartent de la réalité. On peut distinguer trois catégories principales d'hallucinations : les erreurs factuelles, les contenus fabriqués et les sorties absurdes. Les erreurs factuelles incluent des inexactitudes historiques ou scientifiques, tandis que les contenus fabriqués sont des réponses fictives générées en l'absence d'informations. Les sorties absurdes, quant à elles, sont grammaticalement correctes mais dépourvues de sens. Ces hallucinations peuvent être causées par des données d'entraînement biaisées, un surajustement du modèle, une architecture défectueuse ou

des paramètres de génération mal configurés (par exemple, la température) (Farnschläder, 2025).

Bien que les LLM puissent être fine-tunés pour améliorer leurs performances sur des tâches spécifiques, ils ne peuvent pas être mis à jour de manière continue avec de nouvelles informations sans un réentraînement complet. Le fine-tuning permet d'ajouter des informations supplémentaires sur certains sujets, mais il ne remplace pas la nécessité de réentraîner le modèle à partir de nouvelles données pour intégrer pleinement les nouvelles connaissances. Cette limitation peut poser des défis pour maintenir les modèles à jour avec les informations les plus récentes.

Enfin, il est souvent difficile de connaître les sources des informations générées par les LLM. Les modèles ne citent pas explicitement leurs sources, ce qui peut poser des problèmes de vérifiabilité et de confiance. Cette opacité rend difficile la validation des informations fournies par le modèle, ce qui est particulièrement problématique dans des contextes où la précision et la fiabilité des données sont cruciales (Jumelle, 2024) (Microsoft 1, 2025).

2.1.4 Transformer

Dans la première partie de cette section, nous décrivons le fonctionnement général du transformer. Une deuxième partie (cf supra section 2.1.4.2) reprend plus en détail les concepts utilisés dans la description.

2.1.4.1 Description

L'architecture du transformer lui permet d'apprendre la pertinence et le contexte de tous les mots dans une phrase. Pour le faire, son architecture est divisée en deux parties distinctes, l'encodeur et le décodeur. Les composants des deux parties travaillent ensemble et partagent un certain nombre de similitudes.

Avant de passer le texte dans le modèle, il faut tokeniser les mots. Cette tokenisation va permettre de traiter des chiffres au lieu de mots (Karpathy, 2023)

Une fois les données introduites dans le modèle, soit dans l'encodeur, soit dans le décodeur, celles-ci sont « embedded ». Chaque token représentant un mot, sera transformé en vecteur qui permet de le représenter dans l'espace et de comparer les tokens entre eux.

Un « positional encoding » est ajouté au vecteur du token. Cet ajout permet de préserver l'information sur l'emplacement du mot dans la phrase, ce qui est important pour comprendre le sens des données.

Du côté de l'encodeur, les données passent ensuite par différentes couches contenant chacune :

- un « multi-head attention » : combinaison de mécanisme de « self-attention » permettant de percevoir différents aspects du langage, et ainsi de capturer les relations complexes entre eux ;
- un « add & norm » : permettant de récupérer de rajouter l'information d'entrées et de normaliser le tout ;
- un « feed forward » : qui renvoie un vecteur de logits contenant des valeurs réelles qui représentent le score de probabilité qu'un token soit sélectionné;

- un nouveau add & norm.

La sortie de l'encodeur est une représentation enrichie des mots qui est envoyée au décodeur.

Du côté du décodeur, les données passent également dans plusieurs couches contenant les mêmes éléments que l'encodeur à la particularité qu'un « multi-head attention » et un « add & norm » sont rajoutés afin de concaténer les données reçues par l'encodeur et les données d'entrées du décodeur (son résultat de sortie précédent). De plus contrairement à l'encodeur qui compare chaque mot à tous les autres mots des données d'entrées lors des phases de « self-attention », le décodeur ne voit que les mots précédents.

Enfin, le système génère un vecteur de probabilité qui somme 1 via la fonction « softmax », ce qui permet de sélectionner le mot ayant la plus haute probabilité et de générer la suite de la réponse (Ferrer, 2024).

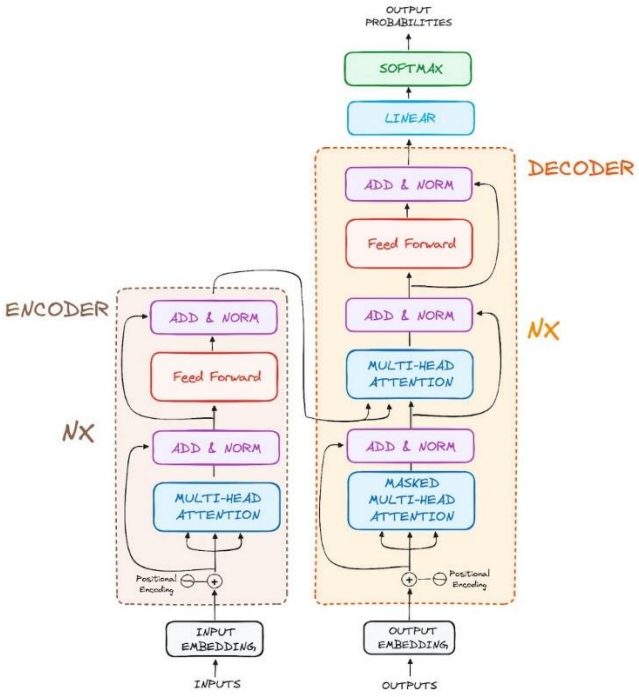


Figure 9: Architecture Transformer (Ferrer, 2024)

2.1.4.2 Composants et traitements du Tranformer

2.1.4.2.1 Tokenisation

En début de traitement, chaque mot ou partie de mot du texte en entrée est converti en nombre. Chaque nombre représente une position dans un dictionnaire comprenant tous les mots utilisés par le modèle. Cela permet de transformer le texte (string) en nombres entiers (int), et de faciliter la suite du traitement (Karpathy, 2023).

2.1.4.2.2 Embedding

L'« embedding » a pour but de capturer le sens sémantique des tokens et de les convertir sous forme de vecteur (Oracle 2, 2025).

2.1.4.2.3 Positional Encoding

Le « positional encoding » est le fait de rajouter à chaque token l'information de la position du mot dans la phrase. La préservation de cette information permet de mieux comprendre le sens du mot en fonction de sa position dans une phrase (Vaswani et al., 2023).

2.1.4.2.4 Self-attention

Le mécanisme de « self-attention » permet d'enrichir chaque mot/token individuel avec son contexte sur base de ses relations avec les autres mots. Il utilise la même fonction que le mécanisme d'« attention » mais pas les mêmes données.

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

Où :

- Q : représente ce qu'un mot cherche à comprendre sur lui-même ;
- K : représente une clé reprenant ce que chaque mot offre comme information ;
- V : représente le contenu réel à utiliser ;
- d_k = dimension de k ;
- T = la transposée (Vaswani et al., 2023).

Lors de ce mécanisme, les vecteurs Q et K sont comparés entre eux pour produire un score de similarité (sous forme de matrice). Le score est ensuite divisé par $\sqrt{d_k}$ pour éviter que les scores deviennent trop grands. Le résultat est ensuite normalisé à l'aide d'une fonction « softmax » afin d'obtenir des poids (leur somme vaut 1). Pour finir, le nouveau vecteur de poids est multiplié avec les vecteurs V correspondant aux valeurs réelles. Le résultat final est un vecteur contenant la combinaison pondérée des informations des autres tokens, devenant ainsi une représentation enrichie du mot initial (Ferrer, 2024).

2.1.4.2.5 Multi-Head Attention

La « multi-head attention » est la concaténation de plusieurs mécanismes d'« attention ». Chaque « attention head » se concentre sur un aspect différent des données, puis les différents résultats sont concaténés, ce qui permet d'améliorer le résultat final (Vaswani et al., 2023).

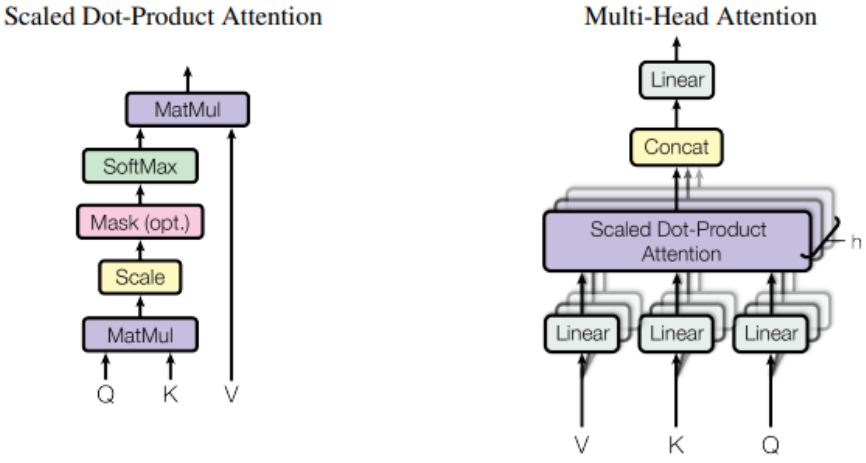


Figure 10: Architecture Multi-Head Attention (Vaswani et al., 2023)

$\mathrm{MultiHead(Q,K,V)} = \mathrm{Concat}(head_1, \dots, head_h)W^O$
Où $head_i = Attention\ (QW_i^Q, KW_i^K, VW_i^V)$ (Vaswani et al., 2023).

2.1.4.2.6 Add & Norm

A chaque itération (couche) du processus d'attention ou de « feed forward », l'« add » permet d'ajouter au résultat de sortie les données de départ. Ce mécanisme crée une connexion résiduelle qui aide à préserver les informations initiales et facilite l'apprentissage des gradients pendant l'entrainement.

Le mécanisme « norm », qui est appliqué après l'« add » standardise les résultats en ajustant et en redimensionnant les valeurs pour avoir une moyenne nulle et une variance unitaire. Ceci stabilise et accélère l'entrainement en réduisant les covariances internes (Ferrer, 2024).[1]

2.1.4.2.7 Feed Forward

Le « feed forward » introduit des non-linéarités dans le modèle, ce qui lui permet de capturer des relations complexes dans les données.

$$FFN(x) = \max(0, xW_1 + b_1)\,W_2 + b_2$$

Les transformations du « feed forward » sont traitées de manière distincte et uniforme à chaque élément du vecteur.

Lors de la première transformation, l'entrée est multipliée par une matrice de poids W_1 qui permet d'augmenter temporairement la taille des vecteurs, et donc sa capacité. Un biais b_1 est également ajouté. Ensuite, la fonction d'activation « ReLu » ($f(x) = \max(0, x)$) remplace toutes les valeurs négatives par zéro. Enfin, le résultat de cette fonction est multiplié par une seconde matrice de poids W_2 pour ramener la sortie à la dimension initiale, à laquelle un nouveau biais b_2 est également ajouté (Vaswani et al., 2023).

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Google Brain, Google Research, Gomez, A. N., University of Toronto, Kaiser, Ł., & Polosukhin, I. (2023). *Attention is all you need*. 31st Conference on Neural Information Processing Systems (NIPS 2017). https://arxiv.org/pdf/1706.03762.pdf

2.1.4.2.8 Encodeur

Le rôle de l'encodeur est de capturer les relations entre les mots de la séquence d'entrée permettant de comprendre le contexte global. Lors du mécanisme de « self-attention » dans l'encodeur, celui-ci voit tous les mots de son contexte. La sortie de l'encodeur est fournie au décodeur (Ferrer, 2024).

2.1.4.2.9 Décodeur

Le décodeur a pour objectif de générer une réponse. Pour ce faire, il traite à la fois l'output de l'encodeur et son propre output. La grande différence avec l'encodeur est, que lors du mécanisme de « self-attention », seuls les mots précédents sont pris en compte afin d'enrichir son contexte (Ferrer, 2024).

2.1.4.2.10 Softmax

$$softmax(y_i) = \frac{e^{y_i}}{\sum_{j=1}^{n} e^{y_j}}$$

Où y_i : valeur individuelle d'un vecteur de logits.

La fonction « softmax » sert à transformer un vecteur de « logits », c'est-à-dire un vecteur de valeurs réelles, positives ou négatives, qui représente les scores non normalisés produits par un réseau neuronal, en un vecteur de probabilité qui somme 1. C'est sur base du résultat normalisé de la fonction « softmax » que le prochain mot est choisi. Généralement, le mot avec la plus haute probabilité sera choisi comme le prochain mot (Vaswani et al., 2023).

2.1.4.2.11 Température

Le paramètre de température a un rôle important dans la diversité des réponses en prenant place dans la fonction « softmax » comme dénominateur des exposants des valeurs réelles du vecteur de logits. Celui-ci va influencer la distribution de probabilité. La valeur du paramètre peut varier entre]0 ;2]. Lorsque sa valeur vaut 1, le paramètre n'a aucun impact sur la fonction « softmax ». En revanche, lorsque le paramètre a une valeur située dans l'intervalle]0;1[, la courbe de distribution des probabilités va subir une transformation qui va l'étirer. Tandis que lorsque la valeur du paramètre est fixée dans l'intervalle]1;2], la courbe de de distribution des probabilités va subir une transformation d'aplatissement. En pratique, la température est souvent limitée entre 0 et 1, car au-delà, la courbe est trop aplatie pour que les probabilités du prochain mot soient réellement différenciées. A ce moment-là, la créativité est telle qu'elle crée uniquement des hallucinations.

$$softmax(y_i) = \frac{\frac{e^{y_i}}{T}}{\sum_{j=1}^{n} \frac{e^{y_j}}{T}}$$

Où T : la température (Sharma, 2024)

Concrètement, une température basse favorise les mots les plus probables, idéal pour des réponses précises et concises. Tandis qu'une température élevée donne plus de chances aux mots moins probables d'être choisis, idéal si l'on recherche de l'originalité ou bien une interaction plus engageante et humaine. Le paramétrage de la « température » dépend fortement des besoins selon le contexte (plus créatif ou plus

factuel). Il est donc conseillé d'expérimenter avec différents réglages pour trouver celui qui convient le mieux (Murel et al., 2024).

2.2 CHATBOT

2.2.1 Définition

Un chatbot est un « Programme informatique basé sur l'intelligence artificielle, capable de répondre en temps réel aux questions d'un internaute, faisant ainsi office de conseiller virtuel : agent conversationnel. » (Larousse, 2025)

La technologie des chatbots connaît une évolution rapide et significative. Les premiers types de chatbots étaient capables de tenir une conversation avec leurs utilisateurs en suivant un script prédéfini. Aujourd'hui, les chatbots basés sur l'intelligence artificielle possèdent la capacité de comprendre et de générer des conversations de manière autonome, sans qu'elles soient prédéfinies à l'avance.

Les chatbots sont utilisés dans de nombreux domaines, notamment dans le support client, et se révèlent être des outils extrêmement utiles. Ils offrent de nombreux avantages, tels que leur disponibilité 24 heures sur 24, 7 jours sur 7, leur capacité à gérer un grand nombre de requêtes simultanément et leur aptitude à fournir des réponses rapides et précises (IBM 3, 2025).

Dans ce chapitre, nous présenterons différentes technologies de chatbots ainsi que leurs composants afin d'identifier les meilleures solutions capables de répondre aux besoins d'Europ Assistance. Nous examinerons rapidement les chatbots basés sur des règles, pour ensuite nous attarder sur ceux utilisant l'intelligence artificielle et, plus précisément, ceux utilisant le traitement du langage naturel (NLP). Nous analyserons également les critères de performance, de sécurité et d'intégration pour déterminer les options les plus adaptées.

2.2.2 Types de chatbots

Il existe plusieurs types de chatbots avec des fonctionnalités différentes. On peut les classer en deux grandes catégories : les chatbots traditionnels et les chatbots basés sur l'intelligence artificielle.

2.2.2.1 Traditionnels

Les chatbots de menus ou de boutons permettent aux utilisateurs d'interagir en sélectionnant des options dans un menu scénarisé. Bien qu'efficaces pour traiter des questions répétitives, ces chatbots peinent avec des demandes plus complexes, car ils se limitent à des réponses prédéfinies. Cela peut allonger le processus pour les utilisateurs, qui doivent naviguer à travers plusieurs options pour trouver ce qu'ils cherchent. De plus, si un besoin spécifique n'est pas couvert par les options disponibles, le chatbot devient inefficace.

Le chatbot basé sur des règles utilise la logique conditionnelle « si/alors » pour créer des flux d'automatisation de conversation. Bien qu'il soit facile à entraîner et efficace pour des questions prédéfinies grâce à une détection basique de mots-clés, il échoue face à des requêtes complexes. Ces chatbots limités ne peuvent pas gérer les questions non anticipées, ce qui peut entraîner une expérience frustrante pour l'utilisateur qui doit être transféré vers un agent d'assistance (Teaganne, 2025).

2.2.2.2 Basé sur l'Intelligence artificielle

L'utilisation de l'intelligence artificielle va offrir une compréhension plus nuancée des questions des utilisateurs, quelle que soit leur formulation, saisir les informations contextuelles pertinentes et rendre les échanges plus fluides (Teaganne, 2025).

2.2.2.2.1 Chatbot « Retrieval-based »

Les chatbots « retrieval-based » fonctionnent en s'appuyant sur des paires de questions-réponses (Q-R) prédéfinies stockées dans une base de données.

Lorsqu'une question est posée, ces chatbots l'interprètent en utilisant la technologie NLP et plus précisément (NLU), qui leur permet de trouver des questions similaires en fonction du sens. Une fois la question interprétée, ils identifient la question la plus pertinente et renvoient la réponse associée.

Ces systèmes agissent comme des FAQ améliorées, proposant des questions basées sur la recherche et fournissant les réponses appropriées, sans recourir à la technologie de génération de langage naturel (NLG), car ils ne créent pas de nouvelles réponses (Teaganne, 2025).

2.2.2.2.2 Generative-based

Contrairement aux chatbots « retrieval-based » qui récupèrent des réponses prédéfinies, les chatbots « generative-based » utilisent des données textuelles pour générer de nouvelles réponses non prédéfinies.

Pour cela, ils utilisent la technologie NLP et plus précisément les sous-technologies NLU et NLG. Grâce à la NLU, ils comprennent le sens de la question posée et, grâce à la NLG, et plus précisément aux grands modèles de langage (LLM), ils génèrent une nouvelle réponse, ne se basant pas sur des réponses prédéfinies, mais sur des documents sources.

Les données utilisées par ces chatbots sont des textes sans questions-réponses associées. Ils peuvent être spécialisés dans un domaine spécifique (domaine fermé) ou couvrir une large gamme de sujets (domaine ouvert). La qualité des données est primordiale, car si les données contiennent des erreurs, le résultat produit en contiendra également (Gite et al., 2024).

2.2.2.2.3 Hybride

Les chatbots hybrides, appelés également « Retrieval-Augmented-Generation » (RAG) combinent les approches « retrieval-based » et « generative-based » pour offrir des réponses plus complètes et adaptées aux besoins des utilisateurs. Ces chatbots utilisent deux types de mécanismes pour répondre aux requêtes : se baser sur des questions/réponses prédéfinies et générer des réponses inédites basées sur une source.

Cette combinaison est particulièrement efficace pour surmonter certaines limites des modèles de langage de grande taille (LLM), telles que la mise à jour des données et la difficulté à citer des sources. Les chatbots hybrides peuvent utiliser des données prédéfinies pour des questions prévisibles et répétitives, tout en ayant la capacité de générer des réponses contextuelles et personnalisées lorsque cela est nécessaire (Gite et al., 2024).

C'est ce type de chatbot que nous décrivons plus en détail dans la section 2.3.

2.3 RAG (RETRIEVAL-AUGMENTED-GENERATION)

2.3.1 Architecture RAG

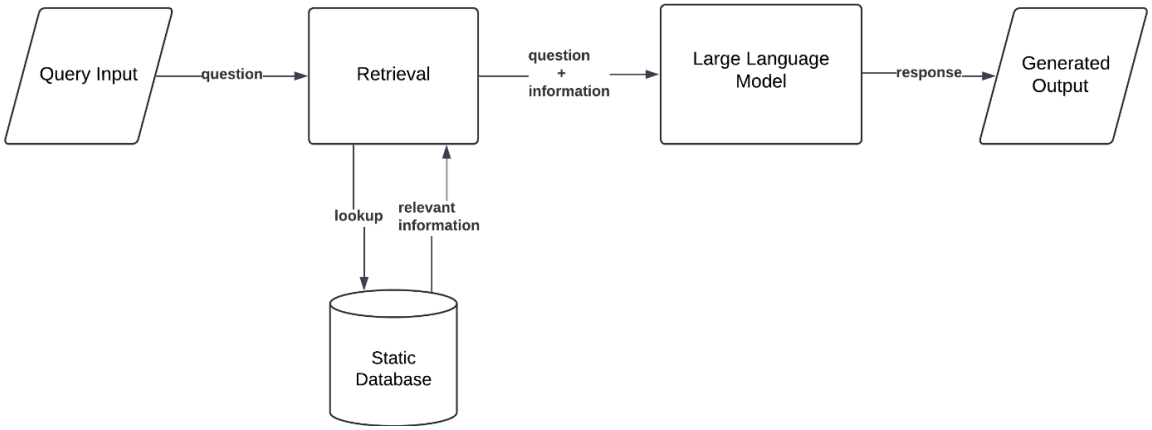


Figure 11 : Architecture RAG (Kelly, 2025)

Le RAG a une architecture particulière combinant un modèle LLM pour la compréhension et la génération de texte à un modèle de récupération d'informations lié à sa propre base de données.

Le RAG présente un double avantage par rapport aux modèles classiques de LLM. Tout d'abord, il dispose de sa propre base de données, ce qui lui permet de mettre à jour ses informations facilement, sans nécessiter un réentraînement du LLM. Ensuite, il a la capacité de sélectionner uniquement les documents pertinents ou même des parties spécifiques de ceux-ci. Cette fonctionnalité permet au LLM de s'appuyer uniquement sur les documents fournis, utilisant ainsi beaucoup moins de tokens que s'il devait traiter l'intégralité des données à chaque fois. De plus, cela facilite la traçabilité des informations (identification des sources). Ce double avantage contribue à améliorer les temps de réponse et à optimiser les ressources lors de son utilisation.

La première étape est le chargement des documents sources. Le premier traitement est la découpe du document en segments appelés chunks. Le but est d'optimiser la pertinence du contenu lors de la recherche d'informations. Puis, lors de l'« embedding », étape primordiale, les chunks sont transformés en vecteurs.

Lorsqu'un utilisateur soumet une requête, cette requête est également vectorisée par un mécanisme similaire. Le retrieval se base sur un calcul de similarité entre le vecteur de la requête et les vecteurs des chunks stockés en base de données.

La requête initiale, les chunks liés aux vecteurs sélectionnés et le prompt sont soumis au modèle LLM (partie décodeur uniquement) afin de générer une réponse dans un langage compréhensible pour un être humain.

Lors de son fonctionnement, le RAG se compose de trois étapes principales : le data processing (dont l'indexation), le Retrieval et la Generation (Belcic, 2024).

2.3.2 Fonctionnement du RAG

2.3.2.1 Data Processing

La première étape est la préparation des données.

2.3.2.1.1 Extraction de texte

Afin que le chatbot puisse traiter les données, il faut que celles-ci soient sous un bon format. Pour cela, il faut dans un premier temps extraire les données des documents.

L'extraction manuelle est chronophage et est souvent source d'erreur. Plusieurs technologies existent pour faciliter le travail.

2.3.2.1.1.1 Reconnaissance Optique de caractère (OCR)

L'OCR est une technologie basée sur le « machine learning » qui permet d'extraire automatiquement des données à partir d'images de texte. Le logiciel OCR identifie les lettres dans les images, forme des mots et des phrases, permettant ainsi l'accès au contenu original en éliminant les tâches de saisie manuelle répétitives.

Son fonctionnement se divise en plusieurs étapes. La première étape est l'acquisition d'images : les pages du document sont alors copiées, puis le moteur OCR convertit le document numérique en version bicolore. Le programme identifie ensuite les parties sombres comme des caractères qui doivent être reconnus, tandis que les zones claires sont identifiées comme un arrière-plan.

La deuxième étape suivante est le prétraitement : l'image numérique est nettoyée pour éliminer les pixels superflus.

L'étape suivante est la reconnaissance de texte, les parties sombres sont traitées pour identifier les lettres, chiffres ou symboles. Cette étape consiste généralement à cibler un caractère, un mot ou un bloc de texte à la fois. Les caractères sont ensuite identifiés à l'aide de l'un de ces deux algorithmes : la reconnaissance de motifs (qui se base sur des données d'entrainement) ou la reconnaissance de caractéristiques (qui se base sur des caractéristiques lorsqu'il travaille avec une police avec laquelle il n'a pas été entrainé).

Les programmes OCR les plus complets analysent également la structure des images contenues dans le document en divisant la page en éléments tels que des blocs de textes, des tableaux ou des images. Tandis que les lignes sont divisées en mots puis en caractères.

La dernière étape est le post-traitement : les informations recueillies sont stockées au format numérique (IBM 1, 2024).

2.3.2.1.1.2 Intelligence Documents Processing (IDP)

La technologie d'IDP applique des techniques avancées de « machine learning » pour organiser et structurer les données récupérées via l'OCR (Farley, 2024).

2.3.2.1.1.3 Basic Model

Le modèle « Basic Model » d'extraction de texte est optimisé pour l'extraction de texte à partir de documents avec des dispositions prévisibles. Il est particulièrement efficace pour les PDF simples et bien structurés qui suivent un design cohérent. C'est une solution rapide et légère qui n'est pas gourmande en ressources. Cependant, le modèle

fonctionne mal sur les documents numérisés ou mal structurés (Europ Assistance 2, 2025).

2.3.2.1.1.4 Azure Document Intelligence

L'« Azure Document Intelligence » est un produit de Microsoft qui utilise une combinaison des technologies de l'OCR et de l'IDP. Cette option est recommandée pour l'extraction de texte à partir de structures de documents simples et complexes. Elle excelle dans l'extraction d'informations à partir de formulaires, de tableaux et de dispositions à double colonne.

Cette méthode est coûteuse et plus lente comparée aux modèles plus simples. Elle dépend également des appels à des API externes, ce qui peut ajouter de la latence et des coûts de service (Farley, 2024).

2.3.2.1.2 Tokenisation

(Cf supra section 2.1.4.2.1)

2.3.2.1.3 Chunking

Selon le dictionnaire de Cambridge, le chunking est une manière de traiter ou de mémoriser des informations en les séparant en petits groupes ou morceaux (Cambridge Dictionary 2, 2025).

Lors du chunking, il est primordial qu'un chunk se suffise en lui-même en étant compréhensible indépendamment des chunks qui l'entourent. Si un chunk est trop petit, il peut perdre des informations contextuelles, tandis que s'il est trop grand, des bruits (causés par des informations non pertinentes ou trop larges) peuvent causer de l'imprécision (Gutowska, 2020).

Il existe plusieurs méthodes de chunking.

2.3.2.1.3.1 Syntaxique

Le chunking syntaxique utilise une méthode structurée pour diviser le texte en fonction de la mise en forme du document. Les titres de sections sont souvent utilisés comme séparateurs naturels. Cependant, pour les documents sans informations encodées, comme les PDF, les tailles de police peuvent servir de marqueurs pour diviser le texte.

Cette approche est facile à mettre en œuvre, mais elle suppose une hiérarchie informationnelle, ce qui peut entraîner des omissions de marqueurs syntaxiques. Le chunking syntaxique est particulièrement utile lorsque la structure du document est prévisible et bien formatée pour le partitionnement (Europ Assistance 2, 2025).

2.3.2.1.3.2 Taille fixe

Le chunking à taille fixe consiste à diviser les données d'entrée en segments de taille égale. Cette méthode sépare souvent le texte en tailles prédéfinies basées sur le nombre de tokens.

Elle est simple à mettre en œuvre et à comprendre, ce qui la rend adaptée aux situations où les données d'entrée ont une taille constante et prévisible.

Les avantages de cette méthode incluent sa facilité de mise en œuvre et son efficacité en termes de charge de calcul. Cependant, elle est rigide et peut entraîner la perte

d'informations contextuelles importantes. Un chevauchement peut être mis en place afin de s'assurer de ne pas perdre du contexte sémantique entre les chunks (Oracle 1, 2025).

2.3.2.1.3.3 Hybride

Le chunking hybride associe les méthodes syntaxiques et de taille fixe. Cette approche permet de segmenter le texte en fonction de sa structure tout en imposant une limite de taille, au cas où une section ou un paragraphe serait trop long.

Dans un premier temps, il découpe le texte en paragraphes. Dans un second temps, il compte le nombre de tokens dans un paragraphe. Si celui-ci est inférieur à un certain nombre défini au préalable, il incrémente les paragraphes suivants un par un jusqu'à dépasser la limite. Une fois la limite atteinte, il met tous les paragraphes incrémentés jusque-là dans un seul chunk, puis il recommence avec le prochain paragraphe. Dans le cas où un paragraphe est plus grand que la limite définie, il le découpe en plusieurs chunks pour respecter la limite.

Les avantages du chunking hybride incluent une taille de chunks optimale. Cependant, cette méthode nécessite plus de ressources et pourrait tout de même couper un paragraphe en deux alors que les deux parties sont dépendantes (Europ Assistance 2, 2025).

2.3.2.1.3.4 Sémantique

Le chunking sémantique sépare le texte en fonction du contexte et regroupe les composants significatifs à l'aide de techniques algorithmiques. Cette méthode est particulièrement importante lorsque la signification du contenu est essentielle pour formuler une réponse, comme dans le cas des documents juridiques.

Elle commence par diviser le texte par phrase. Pour chaque phrase, le calcul de l'embedding est fait. Puis un produit matriciel est fait entre les 2 premières phrases. Le produit matriciel vérifie si la corrélation entre les deux phrases atteint un certain taux de similarité. Tant que c'est le cas, les phrases sont mises dans le même chunk. En revanche, si ce n'est pas le cas, les phrases sont séparées dans des chunks différents. Cette étape est itérative avec toutes les phrases du texte. Cette méthode accepte donc les chunks de taille plus importante et peut être composée de différents paragraphes qui se suivent.

Les avantages du chunking sémantique incluent une plus grande précision et pertinence des données extraites, ainsi que le maintien de la cohérence de la base de connaissances. Cependant, sa mise en œuvre est complexe et nécessite des algorithmes sophistiqués, ce qui le rend plus exigeant en termes de charge de traitement (Europ Assistance 2, 2025).

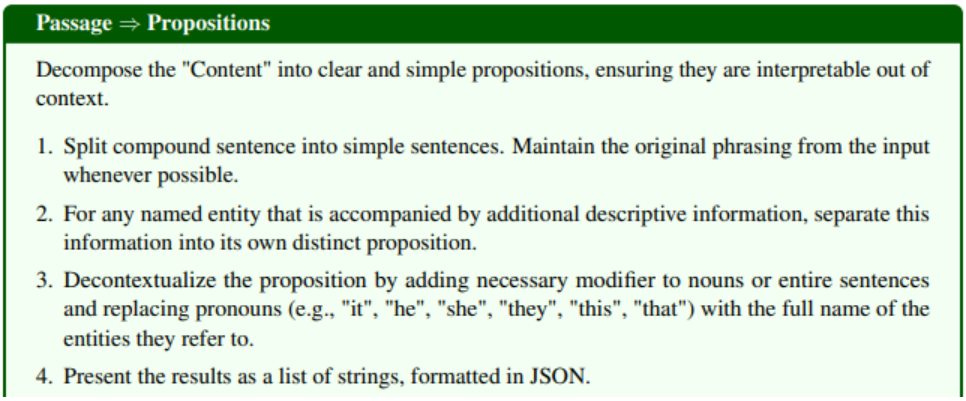
2.3.2.1.3.5 Agentique

Le chunking agentique utilise des modèles de langage de grande taille (LLM) pour générer des segments de textes cohérents et contextuellement appropriés.

Cette approche transforme le texte en « n » propositions grâce à un LLM qui suit trois règles. La première est de garder le plus possible les phrases originales. La deuxième est de créer une nouvelle proposition pour chaque nom dans une phrase accompagnée d'informations descriptives supplémentaires. La dernière est de décontextualiser chaque proposition en remplaçant tous les pronoms avec le nom complet des entités qu'il

représente. A la suite de ces traitements, chaque proposition est une phrase simple et précise. Les propositions sont ensuite regroupées en fonction de leur similarité et un titre est attribué à chaque regroupement via un LLM. Dans cette méthode, la notion de position dans le texte disparait et plusieurs propositions se trouvant à des endroits complètement différents du document peuvent être regroupées.

Les avantages du chunking agentique incluent l'autonomie, permettant aux LLM de prendre des décisions sans intervention humaine, et la cohérence, garantissant que les segments générés sont significatifs et faciles à comprendre. Toutefois, cette méthode sollicite davantage le LLM, ce qui entraîne des coûts plus élevés et peut également provoquer des hallucinations, où le modèle génère du texte basé sur ses propres biais internes plutôt que sur des données réelles (Chen et al., 2024).



Passage ⇒ Propositions

Decompose the "Content" into clear and simple propositions, ensuring they are interpretable out of context.

1. Split compound sentence into simple sentences. Maintain the original phrasing from the input whenever possible.
2. For any named entity that is accompanied by additional descriptive information, separate this information into its own distinct proposition.
3. Decontextualize the proposition by adding necessary modifier to nouns or entire sentences and replacing pronouns (e.g., "it", "he", "she", "they", "this", "that") with the full name of the entities they refer to.
4. Present the results as a list of strings, formatted in JSON.

Figure 12 : Chunking Agentic Model (Chen et al., 2024)

2.3.2.1.4 Vectorisation

Un vecteur est un objet mathématique défini par une origine, une norme (une longueur absolue) et une orientation (une direction et un sens). Un vecteur A est la combinaison de ses composantes vectorielles (vecteurs dirigés selon l'axe de leur composante). De ce fait, chaque vecteur est constitué d'une suite de nombres représentant chaque composante vectorielle et donc son emplacement dans chacune des dimensions (UCLouvain, 2024).

2.3.2.1.4.1 Embedding

L'embedding est le traitement qui permet de représenter des données sous forme de vecteurs.

Cette technique basée sur la technologie NLP permet de capturer la sémantique des données, se basant sur la signification et le contexte des mots et non sur le sens littéral des mots eux-mêmes.

L'embedding peut être appliqué à une large variété de données comme des mots, du texte (phrases, paragraphes ou documents), mais également des images ou des fichiers audios (Barnard, 2023).

A noter que le nombre de dimensions des vecteurs créés pour représenter les relations sémantiques peut atteindre un millier ou plus, dépendant de la complexité des données.

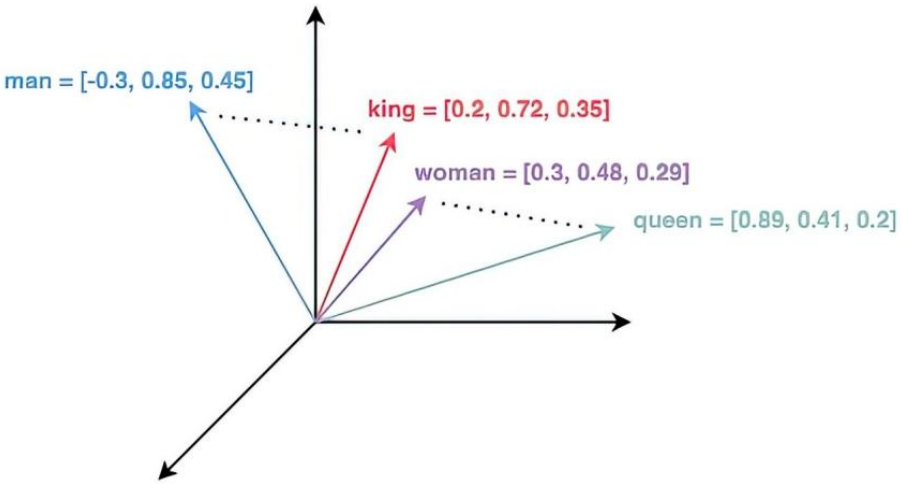


Figure 13: Embedding process (NVIDIA, s.d.)

ADA2, utilisé dans le cadre du projet, est un algorithme d'embedding appartenant à Microsoft Azure utilisant la technologie « Transformer ». Il peut créer des vecteurs à 1536 dimensions et permet ainsi de capter des relations complexes entre les données (Mrbullwinkle, 2025).

2.3.2.1.4.2 TF-IDF

La méthode TF-IDF (fréquence des termes et fréquence inverse des documents) est une méthode de vectorisation alternative basée sur une approche lexicale.

« TF » et « IDF » sont deux métriques.

La fréquence des termes (TF) fait référence à la fréquence d'utilisation d'un terme dans un corpus de texte. La fréquence inverse des documents (IDF) analyse combien de documents dans un corpus contiennent ce terme. Ces deux métriques sont calculées pour chacun des mots. A partir de ces métriques, l'objectif est de déterminer l'importance d'un mot.

Formule : $W_{x,y} = tf_{x,y} * \log\left(\frac{N}{df_x}\right)$

Où :

- W : valeur du TF-IDF ;
- x : un mot ;
- y : un document ;
- $tf_{x,y}$: fréquence de x dans y ;
- df_x : nombre de documents contenant x ;
- N : Nombre total de documents.

Un mot ayant une TF et une IDF importantes (TF-IDF faible) signifie que le mot n'est pas particulièrement utile pour comprendre de quoi parle le document. En revanche, un mot avec une forte TF mais une faible IDF (TF-IDF élevé) signifie que le mot est plus important pour comprendre le contenu du document (Staff, 2025).

En pratique, le TF-IDF est calculé pour chaque mot ou sous-mot d'un texte ou chunk. Ensuite, un vecteur est créé pour chaque chunk. Ce vecteur est composé des valeurs TF-IDF de chaque mot.

Ceci est également effectué sur la requête, ce qui permettra de comparer les chunks et la requête sur la base de leur similarité pour sélectionner les chunks les plus pertinents.

2.3.2.1.5 Indexation (NLU)

Les vecteurs et le chunk source sont indexés de manière à pouvoir optimiser leur récupération dans la phase de retrieval (Oracle 6, 2025).

La manière d'indexer les chunks aura une grande importance dans les mécanismes de recherche décrits dans la section 2.3.2.2.2.4.

2.3.2.2 Retrieval (NLU-NLP)

L'étape « retrieval » est au centre de la solution. Elle consiste à récupérer les chunks pertinents pour répondre à la requête du client.

Afin de savoir quels chunks sont pertinents par rapport à la requête, celle-ci subit également une vectorisation. Ensuite, le système récupère les chunks pertinents en comparant le vecteur de la requête avec les vecteurs des chunks sur la base d'un niveau de similarité.

Tous les chunks inférieurs au niveau de similarité souhaité ne sont pas pris en compte. Il est également possible de limiter le nombre de résultats (nombre maximum de résultats).

La plupart des méthodes calcule la similarité sur base d'une comparaison des vecteurs.

A noter que la méthode Jaccard ci-dessous fait exception. Elle est décrite à titre d'information (Negre, 2013).

2.3.2.2.1 Calcul de similarité

2.3.2.2.1.1 Jaccard

L'indice Jaccard calcule une similarité lexicale qui n'utilise pas de vecteurs mais directement le texte pour comparer la similarité entre des éléments.

Jaccard : $\frac{|set(V) \cap set(W)|}{|set(V) \cup set(W)|} \rightarrow \frac{|Intersection|}{|Union|}$.

L'intersection entre 2 phrases représentée par le symbole (∩) correspond à tous les mots se trouvant dans la phrase 1 et la phrase 2. Tandis que l'union représentée par le symbole (U) correspond aux mots se trouvant dans la phrase 1 ou dans la phrase 2. La division de l'intersection sur l'union nous donne une valeur entre 0 et 1 correspondant à la similarité entre les 2 éléments comparés.

Cette méthode se base uniquement sur le sens littéral et non sur les relations plus complexes comme le sens sémantique (Negre, 2013).

2.3.2.2.1.2 Similarité cosinus

Avec cette méthode, la similarité entre 2 vecteurs est calculée en calculant le cosinus de l'angle séparant les 2 vecteurs. Sa formule est :

$$Sim(A,B) = \cos(\theta) = \frac{A.B}{\|A\| \, \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}.$$

Le cosinus de l'angle entre deux vecteurs est calculé en divisant le produit scalaire des deux vecteurs par le produit de leurs normes (Negre, 2013)

La particularité de la similarité cosinus est qu'elle se base uniquement sur la direction et le sens des vecteurs. Ceci ne pose pas de problème étant donné que tous les vecteurs sont normés au moment de la vectorisation (Tankoua Yojuen, 2025). Lors de la comparaison entre deux vecteurs, plus le cosinus de l'angle qui les sépare est petit, plus l'angle est grand et donc leur similarité est faible (Negre, 2013).

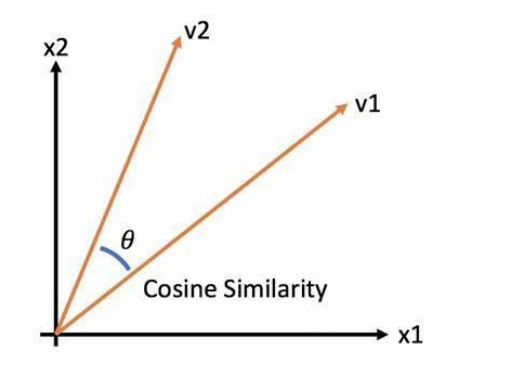


Figure 14: Cosinus similarité (Data Camp, s.d.)

2.3.2.2.1.3 Similarité euclidienne

Initialement, cette méthode ne calcule pas la similarité entre 2 vecteurs, mais leur distance. Cependant, il est possible de calculer une similarité une fois la distance calculée.

Das un premier temps, on calcule la distance avec la distance euclidienne.

Formule de la distance euclidienne : $d(A,B) = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2}$

Où n est le nombre de dimensions des vecteurs, et A_i et B_i sont les composantes des vecteurs A et B (Negre, 2013).

On transforme ensuite la distance en taux de similarité, valeur comprise entre 0 et 1. Plus la distance est grande, plus la valeur résultante de la transformation se rapproche de 0. À l'inverse, plus la distance est petite, plus la valeur transformée se rapproche de 1. Ainsi, il devient plus aisé de comparer les vecteurs (Tankoua Yojuen, 2025).

2.3.2.2.2 Limitation des sources utilisées pour la recherche

Les calculs de similarité sont utilisés dans les modèles de recherche. Une manière d'optimiser les recherches et la limitation du nombre de documents et/ou de chunks à traiter.

2.3.2.2.2.1 Exacte

Le vecteur requête est comparé à chaque vecteur existant pour calculer leur similarité. Seuls les k vecteurs les plus proches sont sélectionnés via la technique « k-nearest neighbors » (kNN). A noter que cette méthode a un coût important en termes de temps.

Elle est intéressante lorsque la priorité est mise sur la qualité des résultats et que le temps de réponse n'est pas critique ou que le volume de vecteurs à traiter est faible (Oracle 3, 2025).

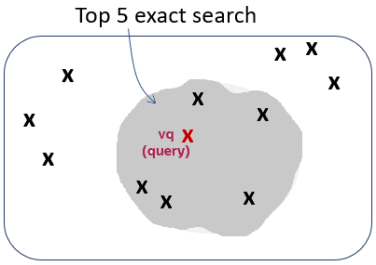


Figure 15 : Exact search (Oracle 5, 2025)

2.3.2.2.2.2 Approximative

La recherche de similarité approximative utilise des index de vecteurs. Lors de l'indexation, on regroupe certains vecteurs en classes. Lors de la recherche, on identifie la/les classe(s) pertinente(s). Seuls les vecteurs de ce(s) classe(s) sont sélectionnés.

Cette technique est plus adéquate lorsque la vitesse est importante et qu'il y a un grand nombre de vecteurs à comparer, tandis que la qualité nécessaire est moindre (oracle, 2025).

Figure 16 : Approximative search (Oracle 5, 2025)

2.3.2.2.2.3 Multi-vecteurs

Cette méthode consiste à effectuer une première recherche servant à limiter le nombre de documents et à sélectionner ensuite les bons chunks (Oracle 4, 2025).

2.3.2.2.2.4 Recherche hybride

La recherche hybride combine la recherche sémantique basée sur les vecteurs et la recherche par mots-clés grâce à un index spécifique combinant les deux types de paramètres.

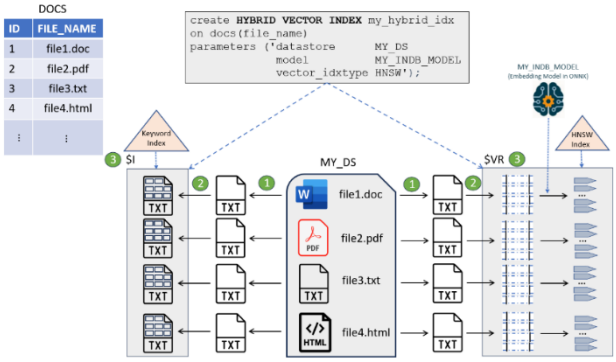


Figure 17 : Hybrid search index (Oracle 6, 2025)

Comme montré dans le schéma ci-dessous, le système est capable d'adresser les différents types de requêtes puis de combiner les résultats via des opérations du type « union », « intersection », … (Oracle 6, 2025)

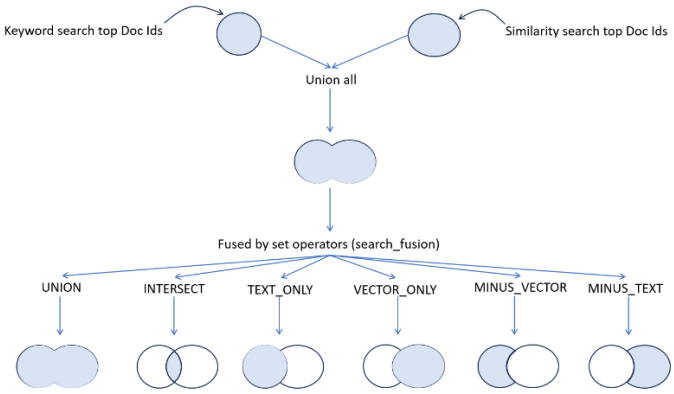


Figure 18 : Hybrid search method (Oracle 6, 2025)

2.3.2.3 Generation

Lors de la génération, le modèle RAG utilise la technologie NLG via un LLM pour générer une réponse basée uniquement sur les chunks sélectionnés lors de la phase précédente « retrieval ». Outre la question, le système utilise également les informations pour structurer la réponse. Le paramètre « température » permet également d'influencer le niveau de créativité lors de la génération de la réponse (Belcic, 2024)

2.3.2.3.1 Prompt engineering

Le prompt engineering désigne les techniques et méthodes visant à optimiser les formulations d'instructions pour les outils d'intelligence artificielle générative. La formulation utilisée dans le prompt et les informations fournies ont un impact important sur la manière dont la réponse est renvoyée (Ionos, 2023).

Concernant les informations fournies, la Holding a sélectionné, dans le cadre du projet, 5 éléments clés : le contexte, l'objectif, les instructions, la structure de la réponse et la langue de la réponse.

Concernant la structure du prompt, on trouve une série de « best practices » sur le site d'OpenAI, entre autres :

- utiliser des délimiteurs entre les éléments clés, par exemple ### ou “““ ;
- décomposer les instructions complexes en une série d'instructions plus simples afin de réduire les risques d'erreur ;
- imposer au système de travailler par étape pour éviter qu'il ne prenne des raccourcis ;
- donner des exemples concrets pour guider le modèle (OpenAI Help Center, s.d.).

2.3.2.4 Types de RAG

Le modèle présenté précédemment est le modèle RAG de base. Cependant, il existe différentes variantes.

2.3.2.4.1 Simple RAG with Memory

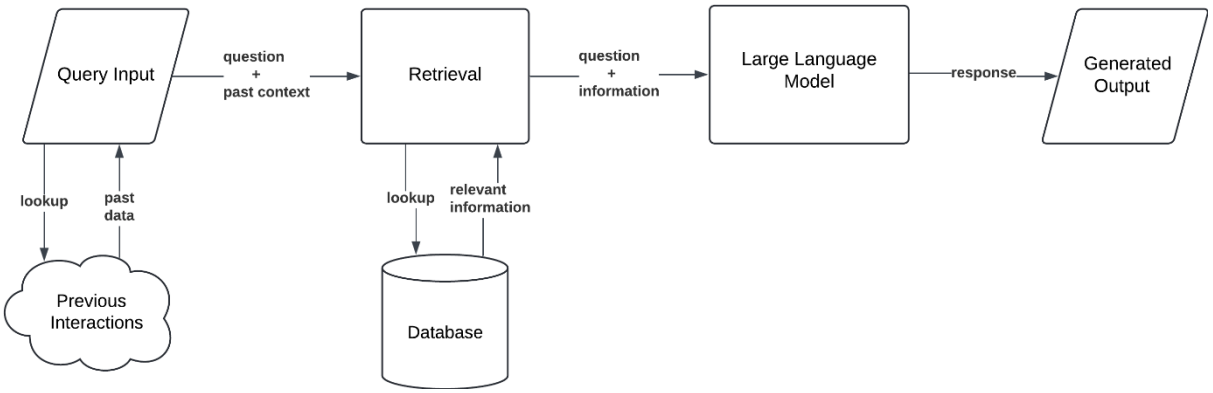


Figure 19: Architecture Memory RAG (Kelly, 2025)

La particularité de ce modèle est l'introduction d'un composant de stockage lui permettant de conserver des informations provenant d'interactions précédentes de la conversation en cours. Lorsqu'il reçoit une nouvelle requête, il l'enrichit avec les interactions précédentes, ce qui lui permet de fournir des réponses plus pertinentes (Kelly, 2025).

2.3.2.4.2 Branched RAG

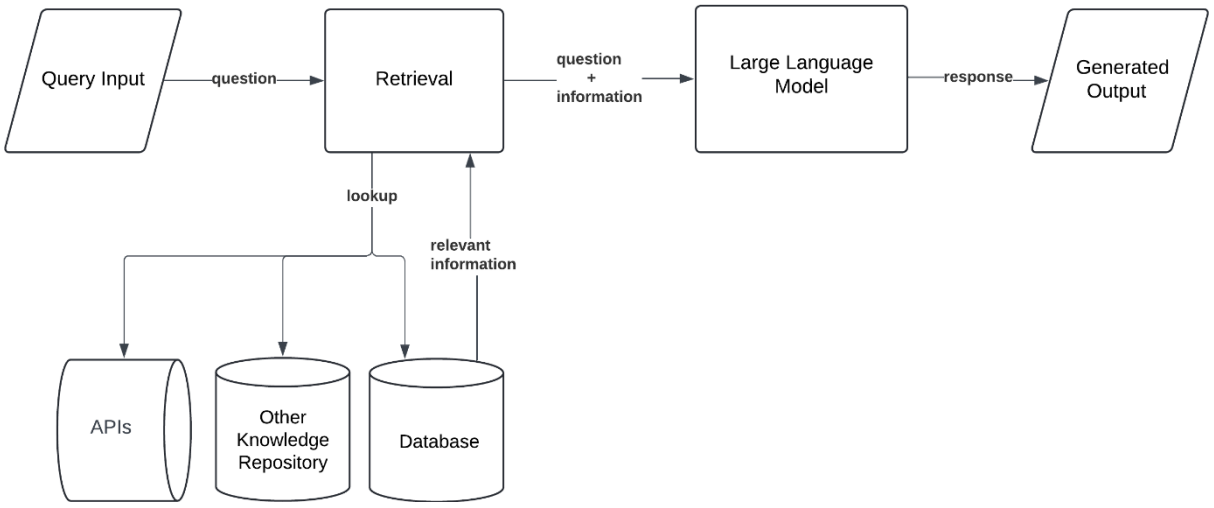


Figure 20: Architecture Branched RAG (Kelly, 2025)

Avant d'effectuer la recherche elle-même, le modèle « Branched RAG » sélectionne les sources de données pertinentes pour la requête, ce qui permet de limiter le nombre de vecteurs à traiter.

Cette méthode est idéale lorsqu'il y a beaucoup de sources différentes de documents ou qu'il faut s'assurer de ne pas sélectionner des informations dans des documents non pertinents (Kelly, 2025).

2.3.2.4.3 HyDe (Hypothetical Document Embedding)

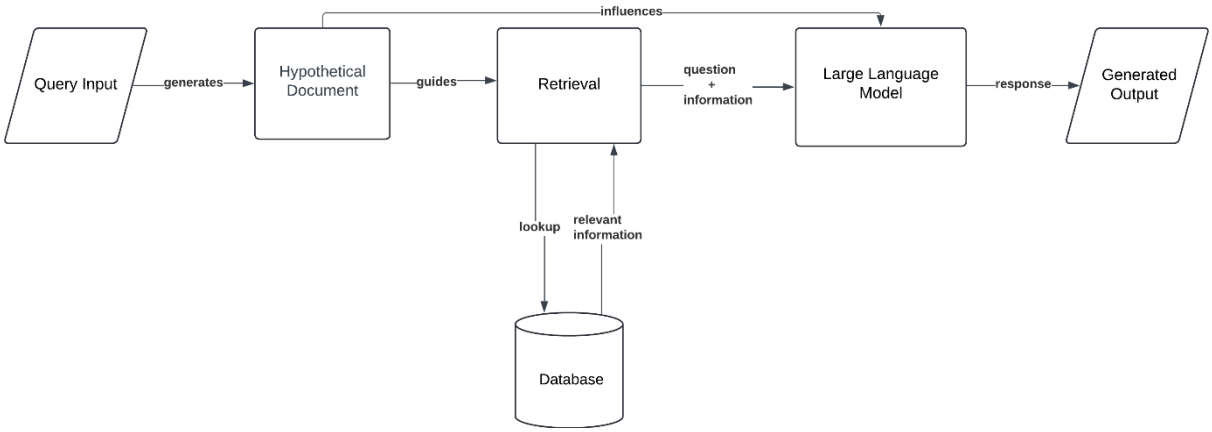


Figure 21: Architecture HyDe RAG (Kelly, 2025)

L'« Hypothetical Document Embedding » RAG (HyDe), quant à lui, génère une représentation hypothétique de ce à quoi pourrait ressembler une réponse idéale. C'est donc à partir de cette réponse hypothétique et non plus à partir de la requête initiale que la phase de « retrieval » va se baser pour récupérer les chunks pertinents. Lors de la génération, la réponse est basée sur les chunks sélectionnés tout en étant influencée par l'hypothèse générée au début.

Cette méthode est plus adaptée lorsque l'on souhaite avoir de la flexibilité et de la créativité ou que la requête est vague. C'est par exemple intéressant dans un contexte « Recherche & Développement » (Kelly, 2025).

2.3.2.4.4 Corrective (CRAG)

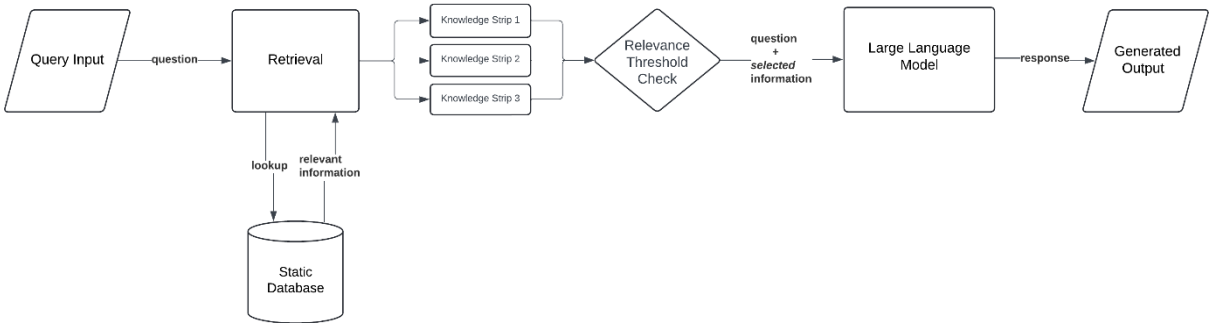


Figure 22: Architecture CRAG (Kelly, 2025)

Le RAG correctif (CRAG) introduit un mécanisme de « self-reflection » qui lui permet de vérifier la qualité des chunks sélectionnés avant de passer à la génération. Pour cela, il découpe les chunks sélectionnés en plus petits morceaux. Chacun de ces morceaux est ensuite analysé afin de vérifier son niveau de qualité. Seuls les morceaux de qualité suffisante sont utilisés pour générer la réponse.

Dans certains cas, si la qualité est suffisante, un nouveau « retrieval » peut être lancé.

Cette méthode est adaptée aux applications nécessitant une précision et une qualité de réponse importante, par exemple lorsque de petites erreurs peuvent avoir de grosses conséquences (Kelly, 2025).

2.3.2.4.5 Self-RAG

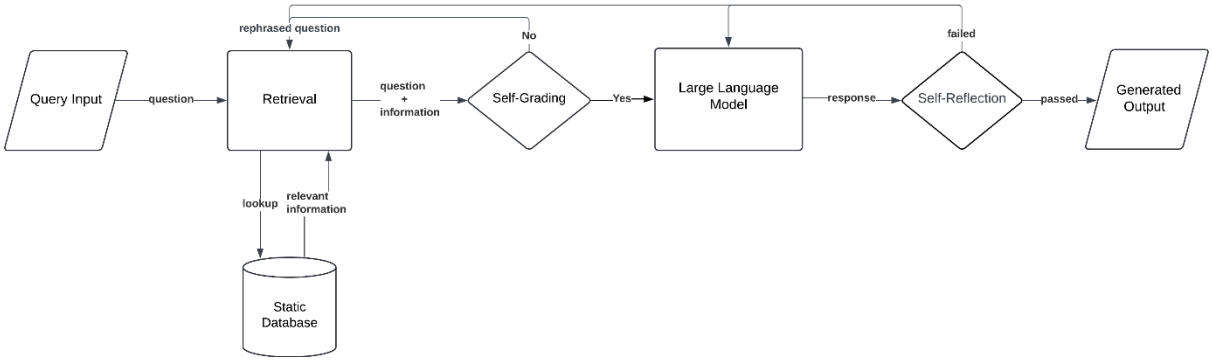


Figure 23: Architecture Self-RAG (Kelly, 2025)

Le « self-RAG » introduit un mécanisme de « self-retrieval » qui lui permet d'affiner la récupération des chunks pendant la génération de sa réponse lorsqu'il détecte qu'une information est manquante.

Ce modèle est très efficace pour la recherche exploratoire, où le modèle a besoin de récupérer de nouvelles informations lors de la génération ou lorsqu'un texte fait référence à un autre texte (Kelly, 2025).

2.3.2.4.6 Agentic RAG

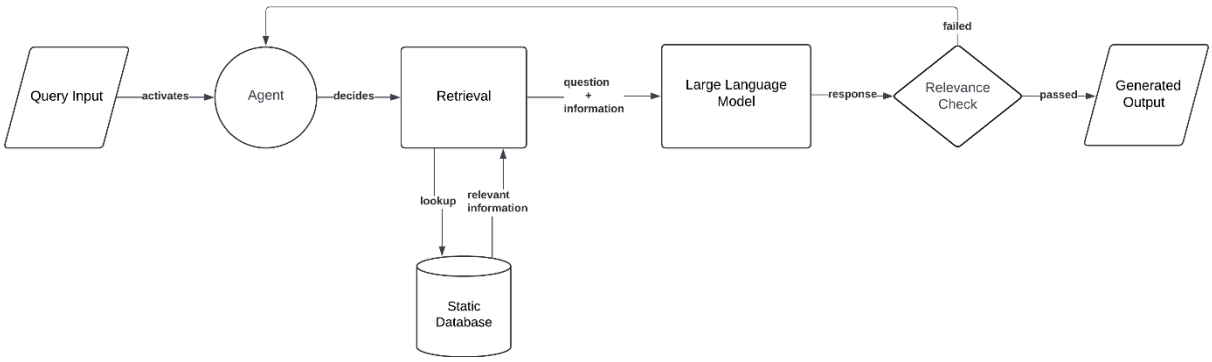


Figure 24: Architecture Agentic RAG (Kelly, 2025)

L'« Agentic RAG » agit comme un « agent » pour gérer des tâches complexes de manière autonomes. Le système peut assigner des « documents agents » à chaque document individuel. Un méta-agent orchestre les opérations et est responsable de générer la réponse sur la base de l'ensemble des outputs.

Ce modèle est idéal pour l'agrégation de données provenant de sources multiples et qui nécessitent des décisions intermédiaires en raison de la complexité de la question (Kelly, 2025).

2.4 MÉTHODOLOGIE PROJET

Il existe deux grands types d'approche pour gérer un projet : l'approche « Waterfall », prédictive, et l'approche « Agile ». Beaucoup d'entreprises implémentent en pratique des approches hybrides visant à utiliser le meilleur des deux mondes.

Je vais également aborder dans cette section quelques concepts importants qui interviennent dans la gestion de projet.

2.4.1 Eléments généraux

Tout projet, quelle que soit sa méthode de gestion, se découpe en 4 grands blocs :

Conception ➔ Planification ➔ Exécution ➔ Terminaison

Parallèlement à ces blocs, des activités de maîtrise (suivi du scope, du planning, des coûts, ainsi que de la qualité et des risques) permettent de suivre le bon déroulement du projet (Nollevaux, 2024, p. 66).

C'est principalement dans l'organisation de ces activités que les méthodes diffèrent.

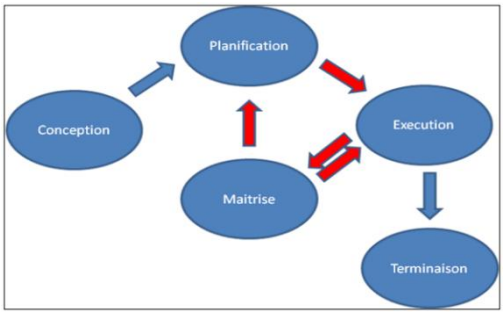


Figure 25 : éléments généraux d'un projet (Nollevaux, 2024, p. 69)

2.4.2 Waterfall (prédictive)

L'approche méthodologique traditionnelle, souvent désignée sous le terme « Waterfall » ou « en cascade », est une méthode de gestion de projet qui se caractérise par une progression linéaire et séquentielle des activités du projet.

Lors de la phase de conception, la charte du projet est rédigée. Cette charte contient les objectifs du projet, les exigences à un niveau élevé, les dates jalons, le budget, le registre des parties prenantes, les opportunités et les risques, les ressources principales du projet.

La phase de planification consiste à élaborer un plan de projet détaillé au niveau des activités et livrables, du calendrier, des coûts, de la qualité et des risques.

La phase d'exécution couvre la mise en œuvre des activités prévues et la production des livrables conformément au plan de projet.

Le plan de projet permet de suivre l'avancement du projet tout au long de sa réalisation et de détecter les écarts éventuels, en particulier concernant le scope, le planning et le budget mais également la qualité et les risques. Cela permet de prendre les mesures correctives si nécessaire.

La phase de clôture marque l'achèvement du projet, avec la remise des livrables et une évaluation finale qui doit permettre de tirer des leçons pour des projets futurs.

L'avantage de l'approche « Waterfall » est que les exigences sont définies tôt dans le projet. Cela permet de définir une baseline assez précise et facilite les activités de suivi. Cependant, cette méthode présente un manque de flexibilité face aux changements d'exigences, puisque les utilisateurs n'interviennent qu'en début de projet (analyse) et en

fin de projet (test). Ils ne voient pas les résultats intermédiaires (Nollevaux, 2024, p. 215) (Microsoft 1, 2025).

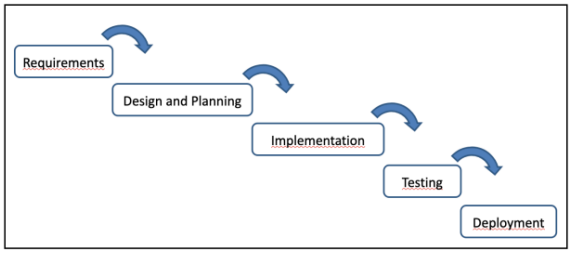


Figure 26 : étapes Watefall (Nollevaux, 2024, p. 215)

2.4.3 Agile

La méthode Agile permet, quant à elle, une grande flexibilité par rapport aux besoins des utilisateurs. Les activités se déroulent en cycles courts (2 à 4 semaines) appelés itérations ou sprints durant lesquels le business, les analystes et les développeurs travaillent en étroite collaboration.

Chaque itération commence par une réunion de planification. Le backlog est parcouru et l'équipe décide des exigences prioritaires à prendre en compte dans le sprint à venir. L'équipe de développement travaille ensuite sur ces exigences pendant la période du sprint. Le sprint aboutit à la livraison d'un incrément du produit qui doit être fonctionnel. Le produit est présenté aux parties prenantes afin d'obtenir des retours immédiats, permettant ainsi d'ajuster les exigences et les priorités pour la prochaine itération. Cette boucle de rétroaction continue permet de s'assurer que le produit final répond aux besoins réels des utilisateurs.

L'approche agile repose sur plusieurs principes clés, dont la collaboration étroite avec le client, la flexibilité face aux changements et la livraison fréquente de petites améliorations du produit. Les équipes agiles sont généralement petites, pluridisciplinaires et autonomes, ce qui favorise une communication efficace et une prise de décision rapide. Les réunions quotidiennes, appelées « stand-ups », permettent de suivre l'avancement du projet, de résoudre les problèmes rapidement et de maintenir l'équipe alignée sur les objectifs.

Les avantages de cette approche sont une meilleure capacité à répondre aux changements, une livraison plus rapide de la valeur au client, et une amélioration continue du produit grâce aux retours fréquents des utilisateurs. Cependant, elle présente aussi des défis, tels que la nécessité d'une forte collaboration et communication au sein de l'équipe, et une gestion rigoureuse des priorités pour éviter la dérive des objectifs (Nollevaux, 2024, p. 228) (Microsoft 1, 2025).

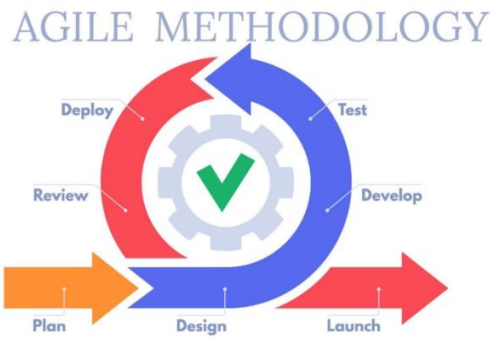


Figure 27 : Agile steps (Davies, 2025)

2.4.4 Hybride

Les méthodes hybrides en gestion de projet combinent les éléments des approches prédictives et agiles pour tirer parti des avantages de chacune. Elles permettent une planification structurée tout en offrant la flexibilité nécessaire pour s'adapter aux changements en cours de projet. Dans une méthode hybride, certaines phases du projet peuvent suivre une approche prédictive, avec une planification détaillée et des échéanciers fixes, tandis que d'autres phases utilisent des cycles itératifs et incrémentaux typiques des méthodes agiles. Par exemple, la phase de conception et de planification initiale peut être réalisée de manière prédictive, en définissant clairement les objectifs, les exigences et les ressources nécessaires. Ensuite, la phase de développement peut adopter une approche agile, avec des itérations courtes et des livraisons fréquentes de produits fonctionnels. Cela permet de recueillir des retours réguliers des utilisateurs et d'ajuster les priorités et les exigences en fonction de l'évolution de leurs besoins.

La méthode « Water-Scrum-Fall » est une combinaison des approches « Waterfall » et Scrum. Elle commence par une phase de planification traditionnelle (Water), où les objectifs, les exigences et les ressources nécessaires sont définis de manière détaillée. Ensuite, la phase de développement utilise la méthode Scrum, avec des itérations courtes et des livraisons fréquentes de produits fonctionnels. À la fin de chaque itération, les produits sont présentés aux parties prenantes pour obtenir des retours immédiats et ajuster les priorités et les exigences pour l'itération suivante. Enfin, la phase de clôture revient à une approche traditionnelle (Fall), avec une évaluation finale et la remise des livrables aux utilisateurs. Cette méthode permet de bénéficier de la rigueur de la planification « Waterfall » tout en offrant la flexibilité et l'adaptabilité de « Scrum » pendant le développement.

Les avantages des approches hybrides sont une meilleure gestion des risques grâce à une planification initiale détaillée, tout en permettant une adaptation rapide aux changements grâce aux cycles itératifs. Elles favorisent également une communication efficace entre les parties prenantes, en combinant des points de contrôle formels avec des réunions régulières et des revues de sprint. Cependant, elles nécessitent une gestion rigoureuse pour équilibrer les deux approches et éviter les conflits entre les phases prédictives et agiles (Nollevaux, 2024, p. 246) (Microsoft 1, 2025).

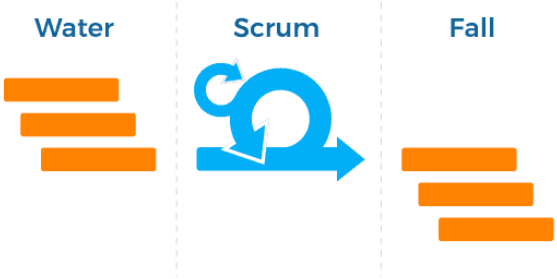


Figure 28: Water-Scrum-Fall steps (Plutora, 2019)

2.4.5 UAT

La phase de test « User Acceptance Testing » (UAT), est cruciale dans les approches traditionnelles de gestion de projet. Elle intervient généralement à la fin du cycle de développement, après que toutes les autres phases de test ont été complétées. L'objectif principal de la phase UAT est de s'assurer que le produit développé répond aux exigences des utilisateurs finaux avant sa mise en production.

La phase UAT consiste à faire tester le produit par les utilisateurs finaux dans un environnement qui simule les conditions réelles d'utilisation. Les utilisateurs exécutent des tests basés sur des cas d'utilisation réels pour vérifier que le produit fonctionne et qu'il répond à leurs besoins. Les UAT permettent de détecter les problèmes qui n'ont pas été identifiés lors des phases de test précédentes, notamment les tests fonctionnels, les tests d'intégration et les tests système.

L'analyse des résultats permet d'évaluer les problèmes identifiés pour déterminer leur gravité et leur impact sur l'utilisation du produit, et de les corriger avant la mise en production. La validation finale par les utilisateurs consiste à confirmer que les corrections apportées répondent à leurs attentes et que le produit est prêt pour la mise en production.

La phase UAT est essentielle pour garantir que le produit développé répond aux attentes des utilisateurs et est prêt à être utilisé dans des conditions réelles. Elle permet de réduire les risques de défaillance après la mise en production et d'assurer une satisfaction maximale des utilisateurs. En impliquant les utilisateurs finaux dans le processus de test, la phase UAT favorise également une meilleure communication et collaboration entre l'équipe de développement et les utilisateurs, ce qui contribue à la réussite globale du projet (Nollevaux, 2024, p. 228) (Microsoft 1, 2025).

2.4.6 TDD

Le « Test Driven Development » (TDD) est une méthode de développement qui met l'accent sur l'écriture des tests avant de développer. Cette approche est particulièrement utilisée dans les environnements agiles, où elle vise à améliorer la qualité du produit et à s'assurer que celui-ci répond bien aux exigences définies. Le TDD est utile pour maintenir un rythme de développement rapide tout en garantissant la qualité du produit, et il favorise une meilleure collaboration entre les développeurs et les parties prenantes. Au fur et à mesure des itérations, les tests sont écrits, le développement effectué et les tests réalisés (Nollevaux, 2024, p. 232) (Microsoft 1, 2025).

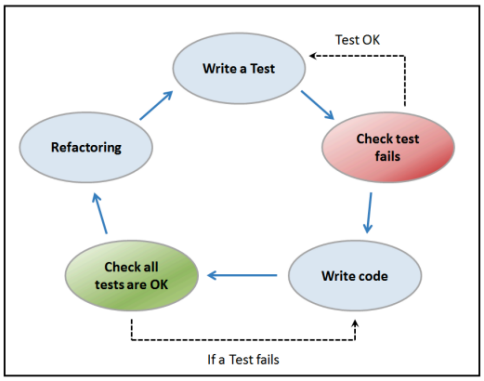


Figure 29 : Test Driven Development process (Nollevaux, 2024, p. 232)

2.4.7 Problem Solving Solution

La méthode de résolution de problèmes « Problem Solving Solution » (PSS) est une approche structurée pour identifier, analyser et résoudre les problèmes. Elle est souvent utilisée pour améliorer les processus et les produits.

La PSS comprend plusieurs étapes clés :

- l'identification du problème : où l'équipe définit clairement le problème à résoudre ;
- l'analyse du problème : où l'équipe collecte des données, identifie les causes potentielles et utilise des outils comme le diagramme de cause à effet pour visualiser les relations entre les causes ;
- la génération de solutions : où l'équipe propose des solutions potentielles et évalue leur faisabilité ;
- la mise en œuvre des solutions : où l'équipe met en œuvre les solutions choisies et surveille leur efficacité ;
- l'évaluation des résultats : où l'équipe évalue les résultats des solutions mises en œuvre et ajuste les actions si nécessaire.

En suivant ces étapes de manière systématique, cette méthode aide les équipes à résoudre les problèmes de manière efficace et à améliorer continuellement les processus et les produits.

Le diagramme « causes à effet », décrit dans la section suivante, est un outil permettant d'appliquer cette méthode (ASQ, s.d.) (Microsoft 1, 2025).

2.4.8 Diagrammes causes à effet

Le diagramme de cause à effet, aussi appelé diagramme d'Ishikawa ou diagramme en arêtes de poisson, est un outil de gestion de la qualité qui permet d'identifier et d'explorer les causes potentielles d'un problème. Il est souvent utilisé dans les approches prédictives pour identifier les facteurs qui contribuent à un problème spécifique et pour trouver des solutions efficaces.

Le diagramme de cause à effet se compose :

- d'une ligne horizontale : représentant le problème à résoudre ;
- des branches principales : représentant les catégories de causes potentielles ;
- des sous-branches : détaillant les causes spécifiques.

En identifiant et en organisant les causes potentielles, le diagramme aide les équipes à comprendre les relations entre les différentes causes et à déterminer les actions correctives nécessaires (Nollevaux, 2024, p. 189) (Microsoft 1, 2025).

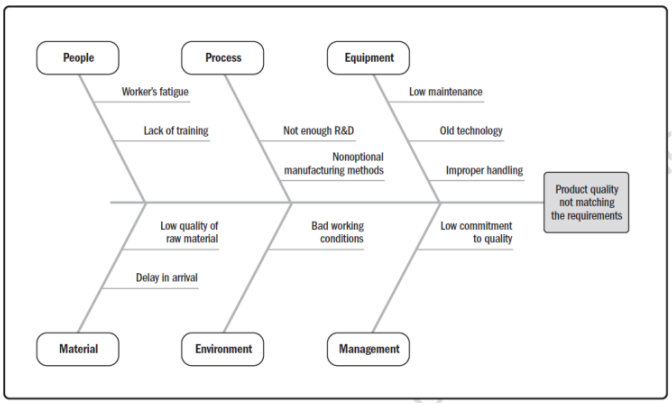


Figure 30: Diagramme Cause-effet (Nollevaux, 2024, p. 189)

2.5 INTÉGRATION

2.5.1 API

« Une interface de programmation d'application est un ensemble de règles et de spécifications qui permet à des applications logicielles de communiquer entre elles. » (Jac, 2025)

Une API permet donc à une application d'envoyer des instructions à une application tierce et de récupérer une réponse.

Dans le cadre du projet, nous avons utilisé le protocole REST qui permet d'appliquer différentes méthodes indiquant le type d'opérations.

L'API envoie une requête composée de plusieurs éléments :

- un « endpoint » : une URL dédiée qui donne accès à une ressource spécifique ;
- une méthode : indiquant le type d'opérations que le client veut effectuer sur une ressource donnée (GET, PUT, POST, DELETE) ;
- des paramètres : pour fournir des instructions spécifiques à l'API ;
- un en-tête : pour fournir des détails supplémentaires sur la requête ;
- un corps : contenant les données nécessaires pour effectuer les opérations.

Lorsque le serveur reçoit une requête, il la traite après avoir effectué une authentification. Une fois traitée, il renvoie la réponse sous forme de requête comprenant les informations nécessaires ainsi qu'un code statut HTTP.

Les API permettent l'intégration de systèmes dans d'autres systèmes, l'automatisation de tâches répétitives et chronophages, et la mise en place de mesures de sécurité pour protéger les données sensibles qui y transitent (Mbiya, s.d.).

2.6 MONITORING

2.6.1 KPI

« a way of measuring a company's progress towards the goals it is trying to achieve » (Cambridge Dictionary 3, 2025)

Un KPI (Key Performance Indicator ou Indice clé de performance) est donc une mesure quantifiable pour évaluer les progrès d'une entreprise vers ses objectifs pour des activités commerciales spécifiques. Suivre ces signaux permet d'alerter l'entreprise sur sa santé (Cambridge Dictionary 3, 2025).

2.6.2 NPS

Le « Net Promoter Score » (NPS) est un indicateur qui mesure la satisfaction et la fidélité du client en posant la question clé : « Quelle est la probabilité que vous recommandiez la marque/produit/service à un proche ? » Basé sur leurs réponses, les clients sont regroupés en trois catégories : Promoteurs, Passifs et Détracteurs. Les promoteurs sont ceux qui donnent une note entre 9 et 10, les passifs entre 7 et 8, et les détracteurs entre 0 et 6. Le NPS est la différence entre le pourcentage de promoteurs et le pourcentage de détracteurs.

L'interprétation d'un NPS n'est pas forcément identique dans les différentes industries ou régions. Cependant, on interprète généralement les résultats selon 4 niveaux :

- en-dessous de 0 : l'entreprise a de nombreux problèmes à résoudre ;
- entre 0 et 30 : la situation est bonne, mais il y a de la place pour des améliorations ;
- au-dessus de 30 : l'entreprise se porte bien et a beaucoup plus de clients satisfaits que de clients mécontents ;
- supérieur à 70 : les clients adorent l'entreprise, ce qui génère beaucoup de bouche-à-oreille positif grâce à leurs recommandations.

Il faut cependant prendre ces résultats avec précaution, car des études montrent que ce sont généralement les personnes avec les avis extrêmes qui participent le plus souvent à ce type d'enquête, ce qui peut biaiser les résultats (Métais, 2025).

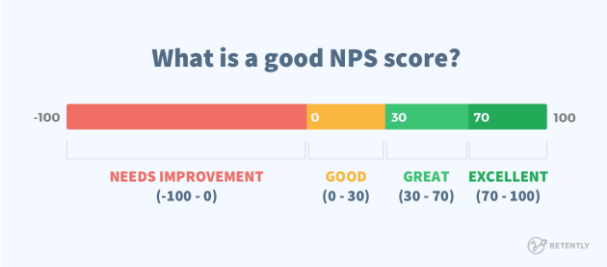


Figure 31: NPS (Grigore, 2025)

2.6.3 CES

Le CES (Customer Effort Score) est aussi un indicateur de satisfaction client. Il mesure l'effort requis par les clients pour accomplir une action spécifique. En fonction de leurs réponses, les clients sont classés en trois catégories : effort faible (1 à 2), effort moyen (3 à 5) et effort élevé (5 à 7). Le CES est calculé en soustrayant le pourcentage de faible effort du pourcentage d'effort élevé (Métais, 2025).

2.6.4 SLA

Un SLA (Service-level Agreement), est un accord entre un fournisseur de service et un client qui définit des standards de qualité de service à respecter (Staff 2, 2025).

2.7 RGPD

Le Règlement Général sur la Protection des Données (RGPD) est une législation européenne qui régule de manière stricte l'utilisation des données personnelles en matière de confidentialité et de sécurité. Il s'applique depuis 2018 à toutes les organisations qui traitent des données personnelles de citoyens ou résidents de l'UE. Ce règlement vise à renforcer la protection des données personnelles à une époque où les violations de données sont fréquentes.

Le RGPD repose sur sept principes qui guident le traitement des données personnelles :

- légalité, équité et transparence : exige que le traitement soit effectué de manière juste et claire pour les personnes concernées ;
- limitation des finalités : les données doivent être collectées pour des objectifs légitimes spécifiés ;
- minimisation des données : impose de ne collecter que les informations nécessaires ;
- exactitude : les données sont mises à jour ;
- limitation de la conservation : limite la durée de stockage des données ;
- intégrité et confidentialité : garantit la sécurité des données ;
- responsabilité : oblige les responsables du traitement à prouver leur conformité au RGPD.

En cas de non-respect du RGPD, les sanctions peuvent être très sévères. Les amendes peuvent atteindre jusqu'à 20 millions d'euros ou 4 % du chiffre d'affaires mondial de l'entreprise. De plus, les personnes concernées ont le droit de demander des compensations pour les dommages subis. Les entreprises doivent également notifier les violations de données dans un délai de 72 heures, sous peine de sanctions supplémentaires.

Afin de s'assurer de respecter le RGPD, un document nommé « DPIA » (Data Protection Impact Assessment) doit être rédigé. Ce document doit être écrit très minutieusement, car une violation de cette réglementation pourrait entraîner des conséquences financières très graves pour l'entreprise (Wolford, 2024) (Microsoft 1, 2025).

2.8 AI Act

L'« AI Act » a pour objectif d'encadrer le développement et l'utilisation de l'intelligence afin de s'assurer qu'il respecte les valeurs de l'Union Européenne (UE) en termes de respect des droits fondamentaux. Les règles mises en place dépendent des cas

d’utilisations classées selon des niveaux de risques (minimal, limité, élevé ou inacceptable).

Les cas d’utilisations dont le risque est évalué comme inacceptable sont bannis de l’UE. Il s’agit par exemple de la manipulation cognitive ou comportementale, des systèmes d’attribution d’un score social, de la catégorisation biométrique des personnes et de l’identification biométrique en temps réel. Des exceptions sont possibles pour les forces de l’ordre dans le cadre de leurs missions spécifiques, mais seulement sous certaines conditions strictes.

Les cas d’utilisation à risque élevé sont fortement encadrés et prévoient des mécanismes pour que les citoyens puissent porter plainte auprès des autorités le cas échéant. Cela concerne principalement des utilisations liées à la sécurité et aux systèmes manipulant des données sensibles.

Les systèmes génératifs appartiennent à la catégorie dont le risque est limité pour autant qu’un principe de transparence soit appliqué. Cela signifie que l’on doit indiquer clairement que les informations sont générées par une IA. Une autre exigence est de veiller à ce qu’aucun contenu illégal ne puisse être généré. Tout incident grave doit être remonté à une commission (European Parliament, 2025).

3 DESCRIPTION DU PROJET ET APPROCHE MÉTHODOLOGIQUE

3.1 DÉFINITION DU PROJET

3.1.1 Périmètre

Le périmètre du projet est de mettre à disposition des agents d'assistance un chatbot intégré dans leur outil de gestion de dossier (STAR) pour faciliter la récupération des données nécessaires au niveau des couvertures et des procédures d'assistance.

3.1.2 Objectifs

Sur base des enjeux décrits dans la section 1.2, des objectifs ont été définis.

L'implémentation du chatbot doit permettre de diminuer le temps de recherche d'information et dès lors le temps de traitement des appels, en augmentant ainsi la productivité des chargés d'assistance.

- Atteindre des temps de recherche à 60 secondes pour les agents fixes et 90 secondes pour les agents temporaires.
- Respecter les SLA des clients B2C et B2B.

De la même manière, la mise à disposition d'une interface de type chatbot doit permettre de diminuer le temps de formation nécessaire, tant pour les employés que pour le personnel saisonnier, et de libérer du temps pour les experts (learning specialists).

- De diminuer d'un mois en moyenne le temps avant qu'un agent soit totalement autonome.

A l'heure actuelle, les informations sont souvent difficiles à comprendre, que ce soit en raison de la structure des documents, du vocabulaire utilisé (termes juridiques) ou encore de l'absence des informations dans la langue du chargé d'assistance. Le chatbot doit permettre de présenter les informations de manière concise, dans la langue de l'agent, et éviter ainsi les problèmes d'interprétations.

- Augmentation de la qualité du service en apportant les réponses correctes et indirectement la satisfaction client.
- Maintenir le NPS constant toute l'année (supprimer les diminutions durant les périodes de pics).
- Maintenir ou diminuer le taux et le montant de leakage à chiffre d'affaire constant (0,16 % et 142302,43 € en 2024) (Europ Assistance 1, 2024)

En résumé, ce projet d'intégration du chatbot vise à améliorer l'efficacité, la qualité du service et la satisfaction client.

3.1.3 Contraintes

Concernant les ressources humaines et budgétaires, nous avons eu plusieurs contraintes importantes :

- le développeur en charge de l'intégration du chatbot dans l'application STAR disponible uniquement de manière ponctuelle en « best effort » ;

- aucun SLA défini avec la Holding alors que la solution est basée sur une interface fournie par elle et qu'elle seule peut apporter les corrections/adaptations nécessaires ;
- disponibilité limitée de l'équipe projet constituée de 2 ressources, 1 autre stagiaire (45 mandays) et moi-même (90 mandays), avec l'équipe CCA en support, sur une période définie (septembre 2024 à avril 2025) ;
- pas de budget d'investissement pour le projet ;
- garantir le respect du GDPR, et donc l'établissement d'un DPIA.

3.1.4 Risques

Les risques identifiés sont classés en fonction de leur impact et de leur probabilité selon le tableau ci-dessous.

Tableau 1: Table de classement des risques (Nollevaux, 2025, p.203

		Impact			
		1-Faible	2-Moyen	3-Elevé	4-Critique
Probabilité	4-Très probable				
	3-Probable				
	2-Possible				
	1- Improbable				

Tableau 2 : Risques du projet

Description du risque	Domaine impacté	Vulnérabilité / Danger	Niveau d'impact	Probabilité	Mitigation
Dépendance de la Holding sans SLA défini.	Planning	Non-respect du planning / Délais.	Elevé	Très probable	Entamer un processus d'« escalation » si la déviation par rapport au planning convenu est trop importante.
Dépendance de la Holding, outil de base fourni par la Holding, identique pour toutes les filiales.	Intégration Projet	Pas d'autonomie pour décider des évolutions.	Faible : la Holding a pour objectif de fournir un outil de qualité et suivre les évolutions du marché.	Très probable	Risque accepté. Pas de mitigation.
Résistance au changement dû à une expérience négative par le passé.	Répercussion Métier	Attitudes négatives envers le projet.	Faible : car limitée à une des parties prenantes (Marketing) et pas aux utilisateurs finaux.	Possible	Atténuation : organisation d'une réunion pour tirer les leçons du passé et éviter de les reproduire.
Résistance au changement suite à un manque de compréhension des mécanismes mis en œuvre.	Répercussion Métier	Perte de confiance et non utilisation de la solution par les acteurs métiers.	Elevé : Non réalisation du Business Case: pas d'amélioration de la productivité et de la qualité de service.	Possible	Communication, formation, proactivité dans le traitement des cas problématiques. Mise en place feedback digitale in tgré au Chatbot. Enquête de satisfaction régulière et des suivi des tendances.
Mauvaise qualité des documents sources.	Répercussion Métier	Qualité des réponses du Chatbot.	Elevé : les résultats étant directement basés sur la qualité des documents.	Très probable	Limitation du scope du projet en ne traitant que les CG propres à CG. Introduction d'un phasage des travaux en mettant une priorité sur les FAQ et CG de pure à EA.
Déresponsabilisation des agents d'assisitance.	Répercussion Métier	Risque d'utiliser les résultats du Chatbot sans analyse critique. Perte progressive du niveau de maîtrise des conditions générales et procédures d'assistance.	Elevé : la qualité de service et l'image de l'entreprise sont directement liées à la qualité des informations fournies par les agents en contact avec la clientèle.	Improbable	Le Chatbot doit être vu comme une aide à la recherche d'information et ne remplace pas la formation des agents. Ce risque concerne essentiellement les nouvelles recrues ou le personnel intérimaire qui sont par ailleurs encadrés par des experts.
Projet non encore mis en service avant la fin de mission du Chef de Projet.	Opérationnel	Non mise en service du projet	Elevé : non réalisation du Business Case.	Probable	L'équipe CCA assure la continuité du projet.
Transfert de connaissance insuffisante durant le projet.	Opérationnel	Abandon de la solution en cas d'incapacité pour le Help Desk et Equipe IT de faire un diagnostic rapide, de corriger ou faire évoluer la solution.	Elevé : non réalisation du Business Case	Possible	La solution est documentée. L'outil de base est fourni et maîtrisé par la Holding. L'équipe informatique locale a été implisuée et maîtrise l'intégration du Chatbot avec les autres outils de l'entreprise.

3.1.5 Expérience passée

EAB, plus précisément le département Marketing avait déjà tenté de créer un chatbot, via une entreprise externe, pour pouvoir poser des questions sur les conditions générales des contrats. Cet essai fut un échec et cela a rajouté de la résistance au projet de leur part. Des réunions ont été créées par la suite pour impliquer d'avantage Marketing afin de leur montrer les avancées technologiques et de diminuer leur niveau de résistance face à cette nouvelle tentative.

3.1.6 Parties prenantes

Le tableau ci-dessous décrit les principales parties prenantes du projet, leur intérêt pour l'initiative et leur niveau d'influence.

Tableau 3 : Liste des parties prenantes du projet

Rôle dans le projet	Fonction dans l'Entreprise	Description du rôle	Intérêt	Influence
Sponsor	Manager AI, Digital & Data Transformation	Commanditaire du projet au niveau EAB	Élevé	Élevée
Project Team	Project Manager	Chef de projet du projet	Élevé	Moyenne
Project Team	Développeur	Intégration du chatbot sur l'outil de gestion STAR	Faible	Élevée
Senior Users	Manager des agents d'assistance	Manager des utilisateurs finaux	Moyen	Élevée
Senior Users	Marketing	S'occupe de l'image d'EAB et de l'écriture des contrats et CG	Faible	Moyenne
Senior Users	Claims	S'occupe du remboursement des clients.	Faible	Faible
Senior Users	Risk & Compliances	Gère les risques liés aux projets au niveau légal. (GDPR, ...)	Élevé	Élevée
Users	Expert & Learning Specialist	Responsable de formation des agents	Moyen	Moyenne
Users	Chargé d'assistance	Utilisateurs finaux du chatbot.	Moyen	Faible
External Supplier	Holding EA	Holding : fournisseur de l'interface du chatbot au niveau de la Holding.	Élevé	Élevée

Supplier	Team Product	Gère le contenu d'Athena	Faible	Faible

3.2 APPROCHE : CHOIX ET JUSTIFICATION

C'est l'approche méthodologique « Water-Scrum-Fall » qui a été choisie pour la gestion du projet, permettant de tirer avantage à la fois de l'approche Waterfall et Agile.

La structure du projet a été définie en suivant l'approche prédictive, ce qui a permis de qu'aucune activité importante ne soit oubliée. On retrouve ces activités dans le planning repris dans la section suivante. Ceci a été possible étant donné que les besoins à haut niveau et le périmètre global étaient connus dès le tout début du projet.

L'étape de planification globale a ainsi permis de bien définir les périodes du projet durant lesquelles des interactions avec la Holding étaient nécessaires, ainsi que celles durant lesquelles les développeurs devaient être disponibles pour l'intégration du chatbot dans l'outil de gestion.

Pour les activités liées au paramétrage du chatbot, c'est le mode Agile qui a été choisi. Ce choix se justifiait par les évolutions rapides de la technologie (livraisons régulières de nouvelles versions de l'outil par la Holding), la méconnaissance des efforts nécessaires pour nettoyer les documents sources, la nécessité de travailler par itération pour l'optimisation du paramétrage (amélioration durant l'itération n des résultats de l'itération n-1).

En complément, des méthodes « Test Driven Development » (TDD) et « Problem Solving Solution » (PSS) ont été proposées pour optimiser les paramètres et les modèles du chatbot. Le TDD permet en effet de s'assurer que chaque modification apportée aux paramètres et au prompt du chatbot est rigoureusement testée avant d'être intégrée et ne provoque aucune régression. Le PSS, quant à lui, aide à identifier, analyser et résoudre les problèmes de manière structurée, en utilisant des outils comme le diagramme de cause à effet qui permettent de visualiser les relations entre les causes et les effets et de faciliter l'identification des actions correctives nécessaires.

3.3 PLANNING

La phase de « découverte de l'entreprise » indiquée dans le planning ne fait pas réellement partie du projet mais m'a permis de découvrir le fonctionnement général de l'entreprise et de mieux comprendre les attentes du sponsor concernant la mise en œuvre d'un chatbot.

ID	Task Name	Duration	Start	Finish
1	**Découverte de l'entreprise**	**7,5 days**	**Mon 09-09-24**	**Wed 18-09-24**
8	**Initier le projet**	**7 days**	**Wed 18-09-2**	**Fri 27-09-24**
12	Rédaction du DPIA	59 days	Mon 30-09-2	Fri 20-12-24
13	**Optimisation des modèles**	**119,5 days**	**Mon 30-09-24**	**Fri 14-03-25**
26	Chatbot integretion	15 days	Mon 10-03-2	Fri 28-03-25
32	Documentation	5 days	Mon 31-03-2	Fri 04-04-25
35	Testing utilisateur	13 days	Mon 07-04-2	Wed 23-04-2
41	Training	6 days	Wed 23-04-2	Wed 30-04-2
45	mise en production	10 days	Thu 01-05-25	Wed 14-05-2
46	After Care	15 days	Thu 15-05-25	Wed 04-06-2
47	Close Project	16 days	Thu 05-06-25	Thu 26-06-25

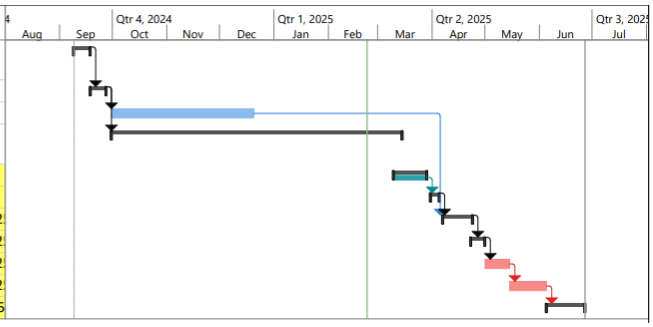


Figure 32 : Planning haut niveau

La phase d'initiation reprise dans le planning prévoit les activités liées à la récolte des besoins, la définition du scope, l'identification des parties prenantes et des risques, ainsi que la fixation des objectifs de qualité et des KPI.

La phase d'exécution a été découpée en plusieurs blocs : la rédaction du DPIA, l'optimisation des modèles, l'intégration du chatbot, la documentation, le testing, le training, la mise en production et l'after care.

Les activités intitulées « optimisation des modèles » ont été organisées en mode agile. Elles couvrent l'identification et la récolte des données sources (FAQ, CG, procédures Athena), l'optimisation des paramètres et des modèles (similarité, température, prompt et modèle LLM).

L'intégration du chatbot dans l'outil de travail existant STAR prévoit les analyses fonctionnelles et techniques, le développement et les tests d'intégration.

Les phases de testing utilisateur (UAT) et la formation des agents sont particulièrement importantes pour s'assurer du bon fonctionnement et de l'acceptation de la solution.

Après la mise en production, la phase d'after care a pour objectif de corriger les erreurs qui pourraient subsister et de mettre en place un support opérationnel.

Le planning initial détaillé est repris en Annexe 1 : Planning.

Malheureusement, ce planning initial a dû être revu en cours de projet, principalement en raison du manque de réactivité de la Holding et du manque de disponibilité des développeurs pour l'intégration du chatbot dans STAR.

A ce jour (mai 2025), la solution est, d'un point de vue EAB, prête pour les UAT, sous réserve de la livraison par la Holding d'une clé d'autorisation. De ce fait, aucune date de mise en production n'a pu être établie.

4 ACTIVITÉS CLÉS

4.1 DÉCOUVERTE DE L'ENTREPRISE

L'objet du projet est essentiellement de fournir une aide aux agents d'assistance en charge d'apporter de l'assistance aux clients, tant pour la business line « Auto & Home » que pour la business line « Travel & Medical ». C'est pourquoi, avant de débuter le projet, j'ai eu l'occasion d'observer et d'interroger des personnes affectées à ces activités. J'ai en particulier pu écouter des appels d'assistance et la manière dont ces appels sont traités.

Les chargés d'assistance exécutent différents types d'activités : la prise en charge des appels entrants (demande d'assistance), la gestion des activités administratives de suivi de dossier (par exemple, le rappel du client pour obtenir des informations complémentaires, ...) et les activités « ticketing » visant par exemple à organiser les activités de rapatriements lorsque nécessaire. Ces activités sont appelées respectivement « flux chaud », « flux tiède » et « flux froid ». Les flux tièdes et chauds sont enregistrés via des « passations ».

J'ai également rencontré des représentants d'autres équipes et services qui apportent du support aux chargés d'assistance, en particulier :

- « Product » : rassemble les experts dans les procédures d'assistance et est responsable de la mise à jour du système Athena ;
- « Network » : s'occupe de la gestion des partenaires ;
- « Claims » : s'occupe des réclamations et des remboursements.

Ces rencontres m'ont permis de mieux comprendre les procédures que les agents d'assistance doivent suivre, d'identifier les domaines susceptibles d'être améliorés ou automatisés et d'identifier les parties prenantes du projet.

4.2 PROCESSUS D'ASSISTANCE

Sur base des premières rencontres, j'ai modélisé le processus d'assistance high level.

Le processus d'assistance à un client se déroule en plusieurs étapes.

Il est déclenché par un appel du client. L'agent d'assistance identifie le client et vérifie si un dossier existe déjà. Les appels concernant un dossier existant consistent en général en une demande d'information sur le statut du dossier ou à la fourniture d'une information manquante.

Lors de la création d'un nouveau dossier, l'agent enregistre les informations du sinistre et vérifie si le client est couvert.

Si le client n'est pas couvert, celui-ci se voit proposer une mise en contact avec un prestataire (à ses frais) et le dossier est clos. Si le client est couvert, l'agent va récupérer les procédures applicables au dossier. Il informe le client des prochaines étapes et met fin à l'appel. Si certaines actions ne peuvent pas être faites immédiatement, l'agent les enregistre via des « passations ». Il rédige également un CRC (Compte Rendu de Communication) qui résume les événements survenus lors de l'appel ou du traitement du dossier afin de faciliter les traitements ultérieurs.

A noter que la complexité des procédures dépend du type d'assistance et que les procédures d'assistance techniques sont majoritairement automatisées (l'envoi d'un dépanneur, la réservation d'un taxi ou la demande d'une voiture de remplacement).

Cependant, les procédures concernant l'assistance médicale et l'assistance voyage sont souvent complexes.

Aujourd'hui, la vérification de la couverture et la recherche des procédures à suivre se fait manuellement, via des recherches dans la bibliothèque Athena.

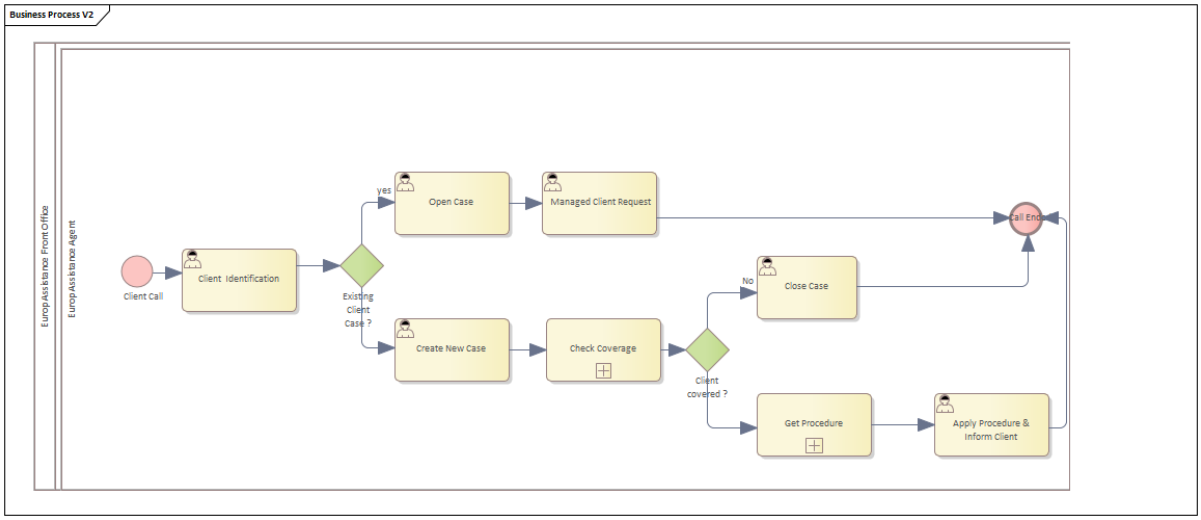


Figure 33: Processus assistance

4.3 COLLECTE DES BESOINS

L'objectif du projet est de créer un chatbot permettant de faciliter deux étapes importantes du processus (vérification des couvertures et identification des procédures) et d'accélérer ainsi l'autonomie des nouveaux engagés et la qualité du service.

Pour collecter les besoins, j'ai mené des interviews plus approfondies au sein des services OPS (Opérations) et du département marketing, partie prenante importante, identifiée en cours de projet. (Voir Annexe 2 : Collecte des besoins)

4.3.1 Agent d'assistance

L'expérience étant un facteur important dans la capacité à vérifier les couvertures et à identifier les procédures à suivre, la première interview a été réalisée avec un agent expérimenté et un agent relativement nouveau.

Leur outil de travail principal étant l'application STAR (gestion des dossiers d'assistance), les deux agents souhaitent que le chatbot puisse être intégré dans celle-ci. Ils doivent au minimum pouvoir lancer le chatbot à partir de leur dossier.

Le contrôle des couvertures et l'identification des procédures se font dans l'application Athena, présentée dans la section 1.2. Cette application fait l'objet de nombreuses remarques et critiques.

- Procédures multilingues
    - Les procédures d'assistance stockées dans Athena ne sont pas entièrement disponibles dans chacune des langues (FR et NL). Une même

procédure contient des paragraphes en FR et d'autres en NL, ce qui pose des problèmes pour les agents qui ne sont pas bilingues. Une mauvaise compréhension du contenu peut avoir de lourdes conséquences sur le traitement des dossiers.

- Recherche d'information
    - Lorsqu'un agent identifie un type de sinistre dans son application de gestion, le système lui présente, via un lien Athena, les conditions de couverture et les procédures liées au type de sinistre sélectionné. Néanmoins, ceci nécessite une bonne maîtrise des types de sinistres et donc un temps d'apprentissage assez important.
    - L'agent peut également effectuer ses recherches directement dans Athena. Les éléments sont structurés selon une arborescence complexe et utilisent un jargon nécessitant encore une fois un temps d'apprentissage important.
- FAQ
    - Les agents peuvent accéder à une FAQ dans Athena, mais celle-ci n'est pas à jour et n'est accessible ni à partir de STAR ni à partir des procédures.

Basé sur ces informations, le nouveau système devra permettre à un agent de poser ses questions et de recevoir les réponses dans sa langue, sans devoir connaître la manière dont les informations sont structurées dans Athena et sans connaître les abréviations utilisées dans cette structure.

4.3.2 Learning Specialist (LS)

Les learning specialists supervisent l'apprentissage des nouveaux agents et leur apportent de l'assistance lorsqu'ils sont bloqués dans le traitement d'un dossier.

Selon le LS interrogé, les nouveaux agents posent trop de questions jugées « inutiles » selon leurs critères. Il s'agit de questions liées à des éléments expliqués dans les procédures d'Athena ou de questions liées aux procédures de base, non comprises par l'agent. Ces questions ont un impact direct sur leur charge de travail, les empêchant d'être disponibles pour des questions relatives à des dossiers plus complexes.

Les pics de questions « inutiles » surviennent surtout pendant les périodes de vacances, lorsque le nombre d'agents est quasiment doublé par l'arrivée d'étudiants.

Le LS souligne également que les FAQ ne sont pas suffisamment mises à jour.

Outre les besoins identifiés par les agents, le chatbot pourrait être enrichi avec les documents de formation afin d'aider en particulier les étudiants qui ne suivent qu'une formation accélérée.

4.3.3 Product

L'équipe Product gère Athena et son contenu. Il s'agit d'une équipe relativement petite qui ne dispose dès lors pas des ressources nécessaires pour traduire l'intégralité des documents.

La personne interviewée explique également que les liens entre les types de sinistres dans STAR et les pages HTML d'Athena sont gérés manuellement, entrainant une charge de travail colossale.

Le chatbot permettrait de traduire le contenu d'Athena dans plusieurs langues en temps réel, sans avoir à le faire manuellement, et faciliterait l'accès direct au paragraphe relatif au « type de sinistre », éliminant ainsi la nécessité de lier manuellement tous les « types de sinistres » à leur page HTML.

4.3.4 CISO

J'ai interviewé la personne responsable de la protection des données.

Nous avons discuté des points importants à surveiller, tels que le type de données clients utilisé, le lieu de stockage des données, le droit d'accès et de suppression des informations, ainsi que l'utilisation des données des clients. Un DPIA doit être rédigé dans le cadre de la mise en œuvre du chatbot.

4.3.5 Département Marketing

Au cours du projet, j'ai appris que le département Marketing avait lancé un projet de chatbot un an auparavant en se faisant assister d'une entreprise externe.

J'ai dès lors voulu les rencontrer pour comprendre pourquoi ce projet n'avait pas abouti et éviter de tomber dans les mêmes pièges (lessons learned).

Il s'agissait d'un chatbot accessible par les clients eux-mêmes.

Problèmes rencontrés :

- type de solution pas assez efficace : modèle LLM optimisé par renforcement (peu efficace) sur base d'un seul document ;
- problème de mise en page des conditions générales ;
    - Ce point est également problématique pour le projet actuel et est présenté en détail dans la section 4.4.2.

En outre, Marketing a également, lors de cet entretien, formulé ses préoccupations spécifiques par rapport au projet, à savoir le respect des taux de qualité. Un non-respect des taux établis par marketing peut en effet soit nuire à l'image d'Europ Assistance soit entraîner des coûts importants en cas de sinistres couverts à tort. Ils souhaitent également que tous les historiques de conversation soient stockés afin de permettre l'identification des responsabilités en cas d'erreur.

4.3.6 Résumé des besoins collectés

En conclusion, les besoins clés du projet de chatbot sont les suivants :

- l'utilisation des conditions générales des contrats, des procédures et des FAQ, tout en les enrichissant continuellement ;
- le respect des réglementations légales, notamment le GDPR et la complétion d'un DPIA ;
- le chatbot doit offrir une grande précision, fournir des réponses dans plusieurs langues et stocker les historiques dans les dossiers clients pour vérifier l'origine des erreurs éventuelles, qu'elles proviennent des agents ou du chatbot ;

- l'intégration du chatbot dans STAR pour éviter que les agents ne perdent du temps à ouvrir une autre application et pour personnaliser le chatbot avec les informations pertinentes sur le contexte des clients.

4.4 ANALYSE DES DONNÉES

Comme expliqué dans la section 2.2.2.2, le chatbot utilise des sources de données pour générer les réponses aux utilisateurs. Dans ce cadre, il est important d'analyser les sources de données pertinentes pour les objectifs du projet. Il faut en particulier que les données répondent à des formats attendus et soient de qualité suffisante. Si les données sont défectueuses, les réponses le seront également.

4.4.1 FAQ

Comme indiqué dans la section 4.3, tant les agents d'assistance que les « learning specialists » ont indiqué que les FAQ (Frequent Asked Questions) étaient incomplètes et n'étaient pas suffisamment mises à jour. La première action à prendre a donc été de demander une mise à jour des FAQ par les « learning specialists ». Par ailleurs, un formulaire a été remis aux agents pour qu'ils puissent suggérer de nouvelles FAQ. Les FAQ ont été organisées sous forme de tableau Excel pour faciliter leur intégration dans le chatbot. A noter que la mise à jour des FAQ sera facilitée dans le nouveau processus en permettant aux agents de liker ou non les réponses proposées par le chatbot.

4.4.2 CG

La vérification des couvertures se fait essentiellement via l'analyse des conditions générales du contrat.

Si les conditions générales sont bien disponibles tant en français qu'en néerlandais pour l'ensemble des contrats, Europ Assistance rencontre des problèmes d'uniformité de formats (PDF vs Word vs scan) et de mise en page (1 colonne, 2 colonnes, tableaux, …). Certains de ces documents dépendent de compagnies partenaires, d'où la difficulté d'imposer un format et une présentation unique.

Pour cette raison, le périmètre a été limité dans une première phase aux CG propres à Europ Assistance et a exclu celles des partenaires.

Cependant, même parmi les CG d'EA, il subsiste des différences ou problèmes de formats nécessitant un nettoyage avant soumission au chatbot. Ce nettoyage a été réalisé grâce à un script VBA que j'ai écrit (Voir Annexe 3 : Code VBA pour la mise en page). Il permet d'une part d'éliminer les en-têtes et pieds de page et d'autre part de remettre tout le texte en une seule colonne.

A ce stade, un problème subsiste au niveau du nettoyage du sommaire, mais n'est pas bloquant pour la mise en œuvre du chatbot.

A noter également que la qualité globale du chatbot peut être diminuée en raison de l'usage important de références dans le texte des CG. La version actuelle du chatbot ne permet pas de récupérer les chunks liés à ces références. Une piste de solution pour pallier ce problème est décrite à la section 5.1.3.5.

4.4.3 Athena

Les procédures d'assistance se trouvent dans Athena sous forme de pages HTML (1 page par thème/sous-thème).

Deux problèmes sont rencontrés :

- les modèles d'extraction n'acceptent pas les pages HTML et les documents doivent être convertis en PDF ;
- une recherche de procédure par un agent doit lui permettre d'obtenir toutes les informations pertinentes pour son contrat, ce qui implique que toutes les procédures pertinentes doivent se retrouver dans un même document (PDF unique).

Aucune solution automatisée n'ayant été trouvée pour générer les documents dans le bon format à court terme, il a été décidé de construire manuellement ces documents pour quelques types de contrat et de permettre de démarrer la phase de test.

A moyen terme, EAB envisage de stocker les procédures dans une base de données qui permettrait de générer directement les documents au format PDF.

A noter que le chatbot permet de traiter automatiquement le multilinguisme, ce qui dispense EAB de traduire l'ensemble de ses procédures.

Comme pour les conditions générales, les procédures liées aux produits des partenaires sont disponibles sous divers formats et ne suivent pas toujours les normes d'Europ Assistance, raison supplémentaire pour se concentrer uniquement sur les contrats EAB dans un premier temps.

On se rend compte que la mise en place d'un chatbot de qualité est directement liée à la qualité des données de base et que leur nettoyage est crucial.

EAB, comme beaucoup d'autres entreprises, a sous-estimé la charge et les coûts liés à cette mise en conformité, entrainant des limitations de périmètre (contrats EAB uniquement) et la mise en place de solutions de contournement.

4.5 MISE EN ŒUVRE DU RESPECT DES LÉGISLATIONS

Afin de s'assurer de la conformité au GDPR, Europ Assistance exige pour chaque projet de compléter une grille d'évaluation permettant de savoir s'il faut compléter uniquement un DPIA simplifié ou un DPIA complet.

Etant donné que le projet prévoit l'utilisation de données contextuelles du client et les soumet à une intelligence artificielle, j'ai dû, sur base de cette évaluation, compléter le DPIA complet.

Les sections principales de ce DPIA sont :

- une description des grandes lignes du projet ;
- la description des traitements utilisant des données personnelles et leur finalité et la justification de l'intérêt légitime à utiliser ces données. Ci-après, les points importants concernant ces traitements et ces données :
    - La finalité du traitement est d'apporter une aide efficace et correcte aux assurés dans un temps le plus court possible.

- Même si la solution utilise des données contextuelles de la situation du client, aucune donnée permettant d'identifier la personne n'est transmise automatiquement au chatbot. Il s'agit uniquement des conditions générales, des garanties souscrites, des dates de validité du contrat du pays d'assistance.
- Dans certains cas, d'autres informations pourraient être introduites manuellement dans le chatbot par les agents d'assistance. Pour limiter les risques, le LLM utilisé est de type fermé (interne à EA, OpenAI ne pouvant pas accéder à ces données). Par ailleurs, tous les systèmes d'Europ Assistance sont hébergés en Europe.

En ce qui concerne les droits des personnes concernées (information, accès, rectification, effacement, restriction du traitement, portabilité, objection, non-soumission à une décision automatisée, retrait du consentement), étant donné que les données utilisées sont les mêmes que celles initialement collectées et stockées dans les systèmes de gestion, les mesures existantes seront automatiquement appliquées aux données utilisées par le chatbot.

A noter que c'est l'agent qui prendra la décision finale sur la couverture ou non d'une personne. Car selon l'AI Act, l'IA ne peut pas prendre de décision seule, impactant les droits d'une personne. Du moins sans passer en risques élevés, et de voir la réglementation se durcir. De plus, il aura toujours accès aux documents sources utilisés par le chatbot pour donner sa réponse, respectant ainsi les règles de transparence concernant la classe des risques limités.

4.6 FIXATIONS DES KPI

Les objectifs et les KPI à suivre ont été définis en collaboration avec l'équipe Marketing et les responsables des différentes business lines au sein du département OPS (opérations). Cela permet de mesurer l'impact du chatbot et de vérifier que celui-ci est positif.

4.6.1 KPI de productivité

Les KPI de productivité actuels sont :

- le temps moyen de traitement de dossier (AHT) ;
- le temps de conversation ;
- le taux d'appel répondu dans les 20 secondes ;
- le taux d'appel répondu dans les 40 secondes ;
- le taux d'appels répondus ;
- le taux d'appels abandonnés ;
- le temps moyen pour prendre un appel ;
- le temps de formation des agents.

Ils serviront de référence pour l'évaluation de la nouvelle solution.

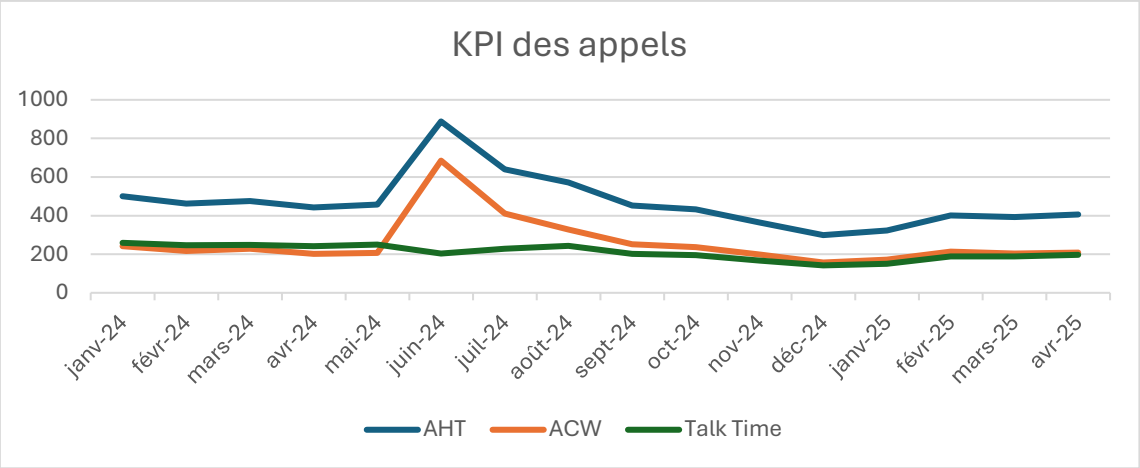


Figure 34: KPI des appels (Europ Assistance 5, 2025)

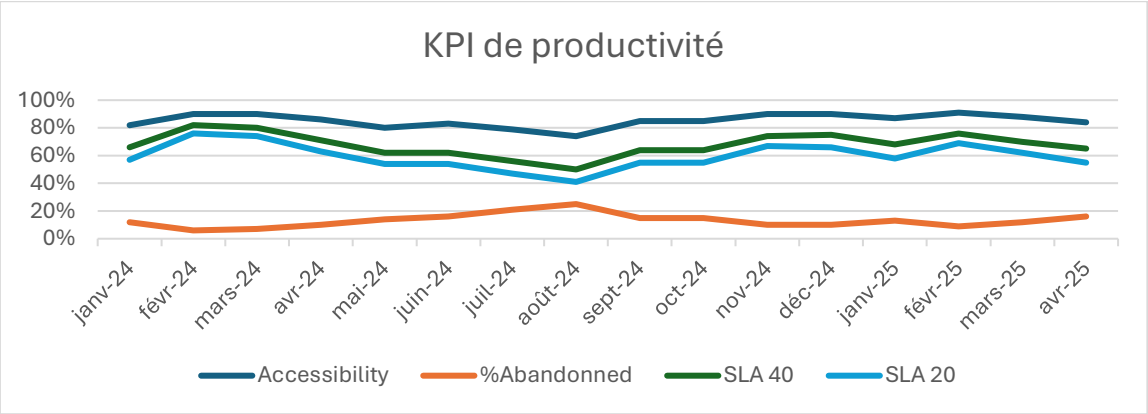


Figure 35: KPI de productivité (Europ Assistance 5, 2025)

4.6.2 KPI de qualité

La mesure des « leakages » (nombre et coût des dossiers acceptés par erreur) n'est pas suivie dans le reporting classique d'EAB, mais les chiffres sont disponibles au niveau de la comptabilité.

Sur base de ces chiffres, j'ai calculé le pourcentage de dossiers concernés et le coût moyen. Ces chiffres pourront servir de base pour suivre l'apport du chatbot.

Un autre indicateur qui pourrait être mis en place est le pourcentage d'appels nécessitant l'intervention d'un superviseur, mais qui n'existe pas actuellement.

4.6.3 KPI de satisfaction

Actuellement, EAB suit la satisfaction client via le NPS (Net Promoter Score).

Figure 36: Evolution NPS 2024 (Europ Assistance 5, 2025)

La figure ci-dessus montre l'évolution du NPS client sur l'année. On constate une forte dégradation de celui-ci lors des mois de juillet et d'août, expliquée par la période de grandes vacances.

Une enquête pourrait être réalisée après l'introduction du chatbot pour vérifier la satisfaction des clients et la satisfaction des agents pour ce nouvel outil de travail par rapport à Athena.

4.6.4 Résultats attendus

Un point important dans l'analyse des résultats est la prise en compte du caractère saisonnier des activités d'EA. En effet, le nombre d'appels augmente significativement durant les vacances. Etant donné en outre que les équipes sont renforcées durant ces périodes par des étudiants, personnel moins qualifié, le temps moyen des appels est plus important.

Outre la diminution globale des temps d'appel, un des résultats attendus du chatbot est d'accélérer la période d'apprentissage des étudiants et dès lors de limiter l'augmentation de la durée des appels durant les périodes critiques.

4.6.5 Objectifs chiffrés

Deux types d'objectifs ont été fixés :

- KPI de productivité et de qualité à atteindre avant mise en production (résultats déjà atteints) ;
    - minimum 95 % de réponses correctes,
    - maximum 4 % de réponses mitigées (ni complètement correctes ni totalement incorrectes),
    - maximum 1 % de réponses fausses,
    - temps de réponse inférieur ou égal à 5 secondes,
- Objectifs en production ;
    - Grâce au processus d'évaluation des réponses du chatbot, l'objectif est d'améliorer les KPI de manière continue et d'atteindre un taux de 99 % de réponses correctes et plus précisément de 99,77 % pour le contrôle des couvertures afin d'arriver aux mêmes chiffres qu'actuellement (leakage : 0,23 %, soit 527 dossiers et 0,16 % du chiffre d'affaires).

Ces objectifs sont ceux liés uniquement au chatbot avant intervention humaine. Le rôle de l'agent reste crucial pour ramener le taux d'erreur aussi proche de 0 que possible.

4.7 ROI

Je vais détailler dans cette section la manière dont les coûts et les bénéfices ont été estimés.

4.7.1 Coûts

Les coûts liés au développement du chatbot ont été considérés dans le calcul du ROI comme un montant unique d'investissement. Celui-ci est assez faible étant donné que :

- l'outil de base a été fourni gratuitement par la holding ;
- le projet a été mené essentiellement par des stagiaires.

Les autres coûts fixes pris en compte sont :

- l'intégration du chatbot dans l'outil STAR et les tests réalisés par l'ICT interne ;
- la préparation des documents (CG : automatique, Athena : manuel) ;
- l'upload des documents via les modèles d'embedding et chunking d'Azure ;
- une formation de 2 heures par agent.

Les coûts récurrents sont liés au nombre de questions posées par les agents (coût lié à l'usage du LLM) et au nombre estimé de tokens nécessaires pour traiter une question. (Voir Annexe 4.1 : ROI initial)

4.7.2 Bénéfices

Les bénéfices sont calculés uniquement sur base du temps gagné par appel.

J'ai d'abord calculé le coût par seconde d'appels d'un agent d'assistance en faisant l'hypothèse que cet agent travaille uniquement au call center sur une base de 35 heures par semaine.

J'ai ensuite estimé le nombre d'appels en tenant compte d'une augmentation de 15 % de ce nombre chaque année et d'une répartition 70/30 entre auto et les autres types de contrats.

Le nombre d'appels a ensuite été limité à 50 % étant donné que l'on ne va traiter que les contrats propres à EAB dans un premier temps.

Pour calculer le gain effectif apporté par le chatbot, j'ai réalisé une enquête auprès des agents qui a permis de mettre en évidence le temps consacré à la recherche d'information durant un appel (couverture et procédure) en tenant compte de la business line et de l'expérience de l'agent. J'ai pris les hypothèses que l'utilisation du chatbot pourra ramener à 60 secondes le temps total de recherche quel que soit la business line ou l'expérience et que leur efficacité augmentera de 20 % chaque mois afin d'atteindre 100 % correspondant au temps idéal par dossier.

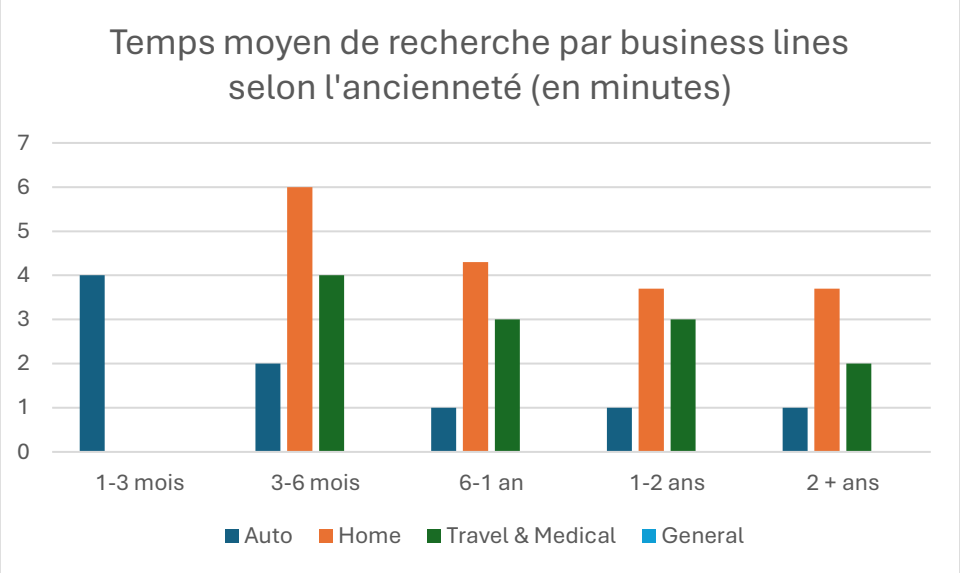


Figure 37 : Temps moyens de recherche d'information par business lines selon l'ancienneté (en minutes) (Europ Assistance 5, 2025) +(Europ Assistance 6, 2025)

Il apparait clairement que les gains potentiels sont plus importants sur les lignes de business « Home » et « Travel » que pour la ligne de business « Auto ».

4.7.3 Résultats

Le graphique ci-dessous illustre que le « break-even point » (le point où les gains générés par l'utilisation du chatbot égalent les coûts associés à celui-ci) est atteint au quatrième mois.

En ce qui concerne le ROI, il est estimé à 217 %, au bout d'un an sur base d'une mise en production le 01/09/2025.

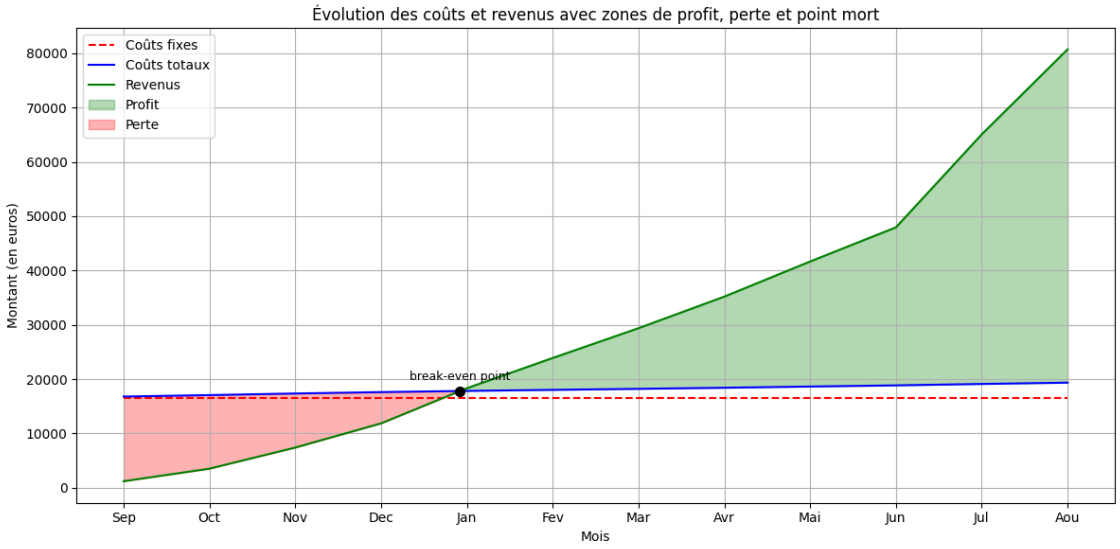


Figure 38: Break-even point (Annexe 4 : ROI Original)

Ce ROI est calculé uniquement sur l'investissement réalisé par EAB, sans tenir compte de la Holding. Quant au profit généré, il repose exclusivement sur le temps économisé lors des appels et ne prend pas en considération les impacts secondaires, tels qu'une accélération des formations, plus de temps pour les LS à se concentrer sur les dossiers à enjeux élevés, ainsi que de meilleurs SLA et NPS, ce qui pourrait permettre de renégocier des contrats grâce à des arguments de qualité améliorés ou d'attirer de nouveaux clients.

Concernant l'implémentation du chatbot, bien qu'il soit prêt durant la période estivale, il serait préférable d'attendre la fin de cette période pour commencer réellement à l'utiliser. Car cela modifierait la manière de travailler et pourrait perturber les agents en pleine période de forte demande. L'idéal aurait été que le projet ne prenne pas de retard et que l'intégration du chatbot ainsi que les formations soient réalisées en mai, offrant un temps d'adaptation jusqu'en juin pour être complètement opérationnels lors des vacances d'été, période où la main-d'œuvre est doublée avec des étudiants.

4.8 PARAMETRISATION DU CHATBOT

La technologie utilisée par la Holding pour développer son chatbot repose sur un modèle RAG simple avec mémoire. Comme mentionné dans la section 2.3.1, le RAG se distingue par sa capacité à récupérer des chunks pertinents en fonction d'une requête, puis à générer des réponses à l'aide d'un modèle LLM. Il garde en mémoire l'historique des conversations et donc un contexte plus précis, ce qui améliore la qualité des réponses.

La mise en place de ce type de solution nécessite une base de données dans laquelle sont stockés les chunks de documents.

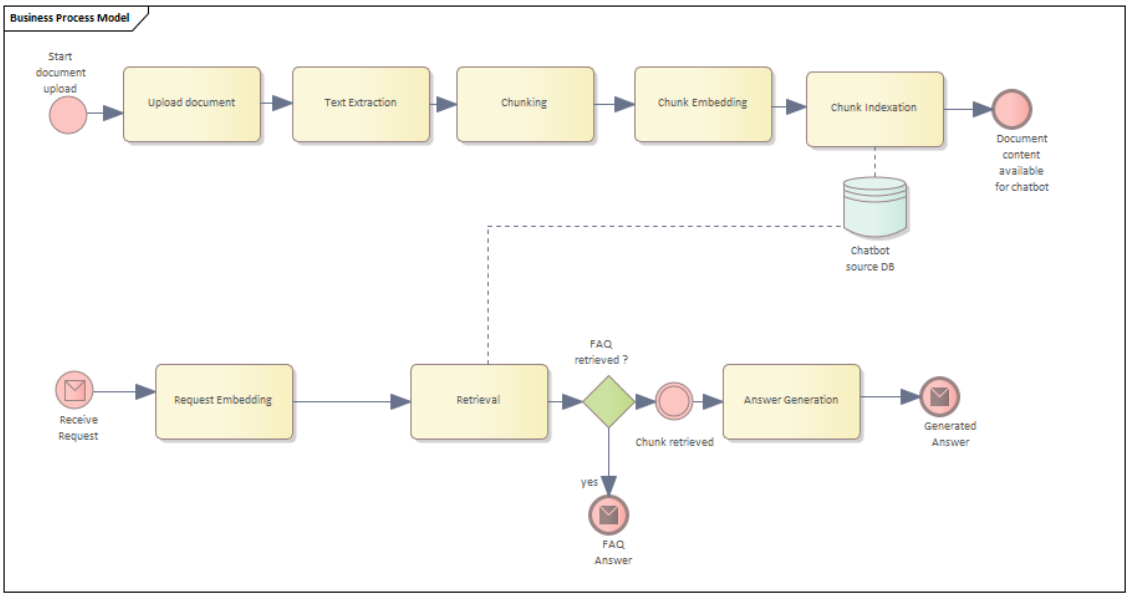


Figure 39 : Processus du RAG

4.8.1 Data Pre-Processing

La première étape consiste à identifier les données nécessaires pour permettre au chatbot de répondre aux questions des utilisateurs.

Ces données doivent ensuite être mises à disposition dans un format exploitable par les outils d'extraction de textes.

L'identification des sources et les étapes de nettoyage sont décrites dans la section 4.4.

Les modèles d'extraction proposés par la Holding reposent sur la technologie OCR (reconnaissance optique de caractères).

Le modèle « Custom » est optimisé pour l'extraction de texte dans des documents ayant une structure et une mise en page prévisibles. Il est particulièrement efficace pour les documents PDF simples et bien structurés. Il est rapide et nécessite peu de ressources.

Les résultats ne sont pas suffisants si on doit traiter des documents scannés ou des documents contenant des textes multicolonnes.

En combinant des technologies OCR et NLP, ce modèle permet de traiter des documents plus complexes (formulaires, tableaux, mise en page sur plusieurs colonnes) ou de moindre qualité (scan).

Cependant, il est plus lent et plus coûteux, car il nécessite l'utilisation d'une API externe dont les coûts dépendent de l'usage.

A noter cependant que, bien que plus efficace que le modèle custom, le modèle « Azure Document Intelligence » n'a pas pu traiter efficacement les CG à doubles colonnes d'EAB. C'est pourquoi, comme décrit dans la section 4.4.2, un nettoyage préalable a dû être réalisé.

4.8.2 Chunking

Trois modèles de chunking ont été fournis par la Holding.

4.8.2.1 Modèle hybride

Le modèle hybride combine une approche syntaxique et une taille fixe de tokens. Le texte est découpé selon sa structure syntaxique, mais si la limite de 1000 tokens est atteinte, le résultat est coupé en deux, ce qui évite les chunks trop volumineux. Cette découpe indépendante de la sémantique peut poser des problèmes lors du « retrieval ».

A noter que des techniques de chevauchement pourraient être mises en place (répétition d'une partie du texte) pour pallier ce problème, mais ne sont pas implémentées dans la version fournie.

Ce type de modèle ne gère pas bien les références entre paragraphes, cas de figure très présent dans les CG d'Europ Assistance.

4.8.2.2 Modèle sémantique

Le modèle sémantique permet de regrouper les phrases partageant une certaine similarité. Il découpe le texte en phrases, les vectorise et calcule la similarité d'une phrase avec la précédente.

Ce modèle présente cependant des limites de performance, en particulier dans le cadre des conditions générales d'EA. Lorsque le système doit traiter les listes d'exclusion de couverture.

Chaque type d'exclusion est considéré comme une phrase séparée et on perd la notion d'exclusion, alors qu'elle est très importante pour les cas d'usage du chatbot.

Ce modèle ne tient pas du tout compte de la structure du document et ne permet pas de répondre au problème de référence entre paragraphes déjà mentionné pour le modèle précédent.

Les ressources de calcul nécessaires pour faire tourner ce modèle sont plus élevées que pour le modèle hybride.

4.8.2.3 Modèle agentique

Le modèle agentique permet de découper chaque paragraphe en « n » propositions, remplaçant tous les pronoms ou références par les éléments référencés. Ces propositions sont ensuite regroupées selon leur similarité mais sans tenir compte de la structure du document.

Ici aussi, la non-prise en compte de la structure du document pose problème, car la similarité des textes liée au manque de contexte génère des réponses erronées.

4.8.2.4 Choix du modèle

C'est finalement le modèle hybride qui a été retenu. Il apporte les meilleurs résultats et est le moins coûteux en termes de ressources de calcul.

Idéalement, il faudrait, avant traitement, revoir les conditions générales et remplacer toutes les références par les textes eux-mêmes. Cette étape n'est cependant pas prévue actuellement.

Une autre alternative serait de ne pas utiliser le modèle RAG simple avec mémoire, mais un modèle RAG « self RAG ». Ce type de modèle n'est pas mis à disposition par la holding à ce jour.

4.8.3 Indexation

La Holding a adopté le modèle « ADA2 » de Microsoft Azure pour la vectorisation. Ce modèle permet de représenter les chunks sous forme de vecteurs de 1536 dimensions.

Les résultats sont ensuite indexés pour faciliter la récupération (retrieval) en permettant une comparaison simple de la similarité entre les vecteurs.

4.8.4 Retrieval

Le retrieval consiste à retrouver les chunks pertinents pour la requête. Dans la solution fournie par la holding, les chunks de l'ensemble des documents sont stockés dans une seule base de données. Pour augmenter la qualité des réponses, une première étape consiste à limiter les recherches aux seuls documents pertinents liés au contrat spécifique du client. Cette présélection se fait via un script qui permet d'affiner les sources.

Le modèle de recherche implémenté est un modèle de similarité cosinus et applique un modèle de sélection « exact ». Seul le top k des chunks les plus pertinents est alors sélectionné.

Dans l'interface fournie par la holding, un curseur permet de sélectionner un niveau de similarité compris entre 0 et 1. Ces deux limites correspondent en réalité à un pourcentage de similarité vectorielle de 70 à 100 %.

Pour le projet, le niveau de similarité 0,25 a été sélectionné, soit 77,5 %. En outre, seuls les 9 chunks les plus pertinents sont présentés.

4.8.5 Generation

La dernière étape du RAG consiste en la génération d'une réponse. Pour ce faire, trois éléments sont indispensables : un modèle génératif pour la création de texte, une température pour contrôler la créativité du modèle de génération et un prompt pour structurer la réponse de celui-ci.

4.8.5.1 Modèle génératif

Le modèle GPT a été choisi par la Holding car les produits Azure se trouvaient déjà dans l'écosystème d'Europ Assistance et que le contrat Microsoft garantissait déjà que les données ne pouvaient pas être partagées.

A noter que des tests ont été réalisés dans une seconde phase avec des modèles de Deepseek et de Gemini :

- Les résultats de Deepseek en termes de qualité et de rapidité se sont révélés moins bons que ceux du modèle GPT. Malgré des coûts inférieurs, la solution n'a pas été retenue.
- Les résultats des modèles les plus avancés de Gemini offrent de meilleures performances à un prix similaire. Ces avantages n'étaient cependant pas suffisants pour remettre en question les contrats Azure existants (Tankoua Yojuen, 2025).

C’est donc le modèle GPT d’OpenAI qui a été choisi par la Holding. Il s’agit d’un modèle de type LLM. Plusieurs versions ont été mises à notre disposition. Des tests ont été effectués afin de comparer les différentes versions en fonction de la qualité, de la rapidité et des coûts de génération tout en maintenant les conditions constantes (ceteris paribus) en termes de température et de prompt.

Les modèles les moins performants ont été écartés directement (versions GPT-3.5 et GPT-3.5 – 15K). Ensuite, les modèles les plus coûteux ont été mis de côté (versions GPT 4 et GPT-4-32K). (Voir Annexe 5 : Calcul de coûts)

Parmi les deux versions restantes (GPT 4o et GPT 4o-mini), la version GPT 4o-mini a été retenue car elle est plus rapide pour une qualité légèrement inférieure.

4.8.5.2 Température

Le paramètre de température influence directement la créativité des réponses. Plus ce paramètre est élevé, plus le risque de générer des erreurs, voire des hallucinations, est élevé.

Conscients de ces risques et du fait que les conditions générales utilisent des termes spécifiques, le niveau de température a été fixé à zéro. En effet, toute erreur pourrait avoir des conséquences négatives pour le client, notamment si son assistance lui est refusée alors qu’il y a droit, ou pour Europ Assistance, si le client reçoit une assistance alors qu’il n’y a pas droit, entraînant des coûts supplémentaires appelés "leakage".

4.8.5.3 Prompt

Le dernier élément, mais non le moindre, est le prompt. Ce dernier fournit des instructions claires au modèle de génération, lui indiquant comment structurer sa réponse et quelles étapes il doit suivre pour y parvenir.

Pour ce faire, le prompt est constitué de sous-éléments tels que le contexte, les objectifs, les tâches, les capacités, les restrictions et la structure de la réponse.

Le contexte indique au LLM dans quel environnement il agit. Les objectifs définissent les résultats finaux attendus. Les tâches englobent tout ce que le LLM doit accomplir avant de générer sa réponse. Les capacités viennent en soutien aux tâches qu’il doit réaliser. Les restrictions sont les éléments qu’il doit impérativement respecter. Enfin, la structure de la réponse indique quels éléments doivent être présentés et dans quel ordre.
(Voir Annexe 6 : Prompt)

4.8.6 Approche méthodologique utilisée pour la paramétrisation

Comme indiqué dans la section 3.2, la méthode Agile a été choisie pour organiser les activités de cette phase mais que des méthodes TDD et PSS ont été proposées en complément pour accélérer l’atteinte de résultats satisfaisants.

Concernant l’application de la méthode TDD, un set de question représentant les tests a été élaboré. Il a été enrichi au fur et à mesure d’une part pour résoudre des cas de plus en plus en plus complexes et d’autre part pour étendre le scope business (produits). Cela a permis de vérifier plus facilement si les valeurs des paramètres étaient optimales et si le prompt était suffisamment complet pour répondre correctement à une variété de requêtes. La méthode PSS quant à elle a permis d’analyser de manière efficace les

problèmes en cas de mauvais résultats (Voir Annexe 9 : « Root Cause Analysis » du chatbot).

La combinaison de développement itératif et du TDD a permis d'identifier rapidement les ajustements nécessaires à apporter aux paramètres et au prompt.

4.9 AMÉLIORATION DE L'INTERFACE

Au moment de mon arrivée, une première version de l'interface du chatbot, un RAG et un système de FAQ étaient déjà en place dans un environnement de test. Dans cet environnement, les réponses fournies par le chatbot pouvaient être likées, et ajoutées automatiquement aux FAQ.

Durant mon stage, j'ai introduit différentes demandes d'amélioration auprès de la holding.

- Développer un algorithme capable d'automatiser le test du chatbot en lui fournissant une liste de questions, pour lesquelles nous récupérerions immédiatement toutes les réponses.
    - Un algorithme a été développé mais mis à disposition quelques jours seulement dans le cadre du benchmark décrit au point suivant.
- Utiliser l'algorithme pour comparer différents modèles de GPT, différents niveaux de température, différents niveaux de similarité ou différents prompts afin d'optimiser le paramétrage du chatbot.
    - Seul le benchmark comparant différents modèles de GPT a été mis en place, et ce, assez tard dans le processus d'optimisation.
- Ajouter un bouton dislike dans l'interface pour permettre aux utilisateurs de signaler facilement les erreurs et la mise en place d'un processus pour traiter ces erreurs.
- Améliorer les suggestions de FAQ présentées à l'utilisateur en passant d'une recherche lexicale à une recherche sémantique.
    - Cette suggestion n'est pas encore implémentée.
- Améliorer la présentation des FAQ pertinentes dans le contexte d'un dossier : présenter les FAQ spécifiques mais les compléter systématiquement avec les FAQ générales.
    - En cours.

4.10 IMPLÉMENTATION

Les développements nécessaires pour l'intégration et l'initialisation du chatbot sont réalisés par les équipes de développeurs. Néanmoins, durant mon stage, j'ai réalisé des analyses et des prototypages afin de clarifier les spécifications et de faciliter leur travail.

4.10.1 Intégration du chatbot dans l'application STAR

4.10.1.1 Ergonomie

L'ergonomie de l'application est un point essentiel pour les agents. C'est pourquoi le choix de l'emplacement du bouton, dans les écrans STAR, permettant de lancer le chatbot est important. Suite à une discussion avec des agents d'assistance, il a été convenu de le positionner au sein de la section « Navigation », dans la sous-section «

Documentation Produit ». Le bouton sera placé juste en dessous du bouton permettant actuellement d'accéder à la documentation d'Athena.

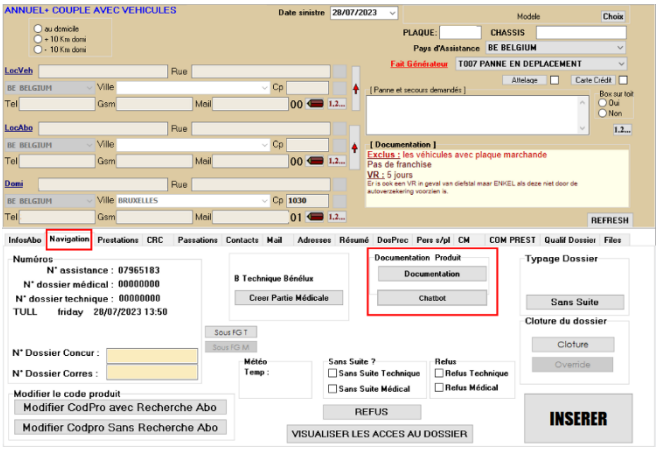


Figure 40: Emplacement du bouton dans STAR

A la fin de la conversation avec le chatbot, il a été décidé que les échanges soient stockés et accessibles à partir de l'application STAR, plus précisément à partir de la section CRC (Compte Rendu de Conversation) dans laquelle une nouvelle sous-section « Historique » sera créée. Cette mise à disposition de la conversation permettra en particulier d'aider à l'identification de l'origine d'une erreur lorsque nécessaire.

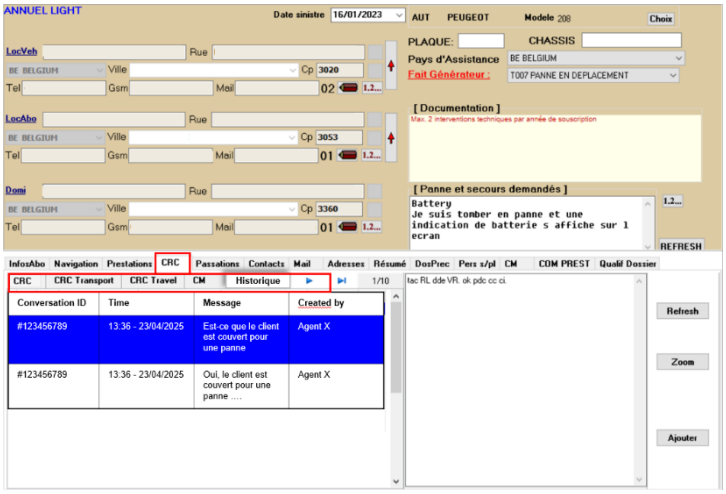


Figure 41: Emplacement de stockage dans STAR

Le seul impact pour l'agent sera dès lors la possibilité de choisir la méthode qu'il souhaite pour vérifier les informations (à la fois pour les couvertures et pour les procédures).

4.10.1.2 Contexte client

Pour travailler efficacement, le chatbot a besoin de connaître le contexte (dossier client).

Les données principales nécessaires sont :

- les références du produit (CG) : ces références permettent de restreindre les recherches aux seuls documents pertinents (sources) ;
- les garanties du contrat ;
- les dates de couverture ;

- le pays d'assistance ;
- la date du sinistre.

A noter que c'est la mise à disposition de ce contexte qui a conduit à l'obligation de rédiger un DPIA. On peut cependant affirmer que les contraintes de finalité et de proportionnalité sont respectées, puisque ces données ne permettent pas en tant que telles d'identifier le client et qu'elles sont nécessaires pour que le chatbot puisse répondre de manière pertinente. La conversation sera stockée dans le dossier client, mais ne contiendra jamais plus de données personnelles que les données déjà présentes dans ce dossier. Le traitement lié au chatbot n'augmente pas le risque au niveau de la protection des données personnelles.

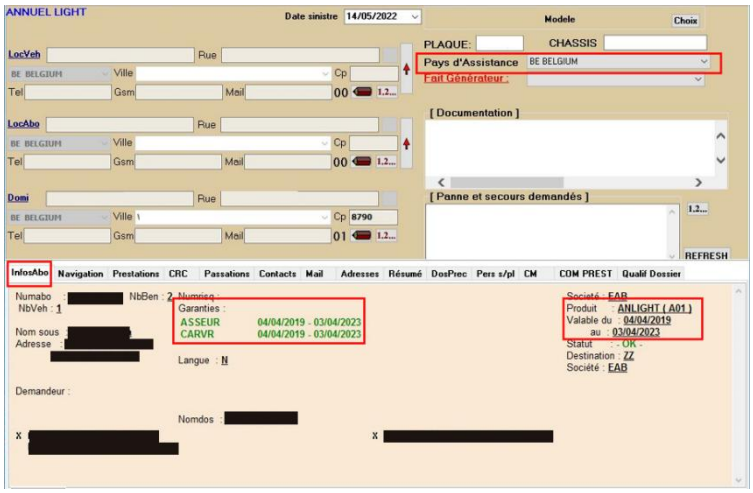


Figure 42: Informations récupérées dans STAR

Parmi les données du contexte, le nom du produit sera utilisé pour extraire uniquement les documents associés à ce dernier, ce qui permet de réduire considérablement le nombre de chunks à analyser.

4.10.1.3 Interactions du chatbot avec les systèmes de gestion

Le schéma ci-dessous a pour objectif d'expliquer l'ensemble des interactions entre le chatbot et les autres systèmes de gestion de manière à le personnaliser au contexte du dossier, à évaluer les réponses et à historiser les résultats.

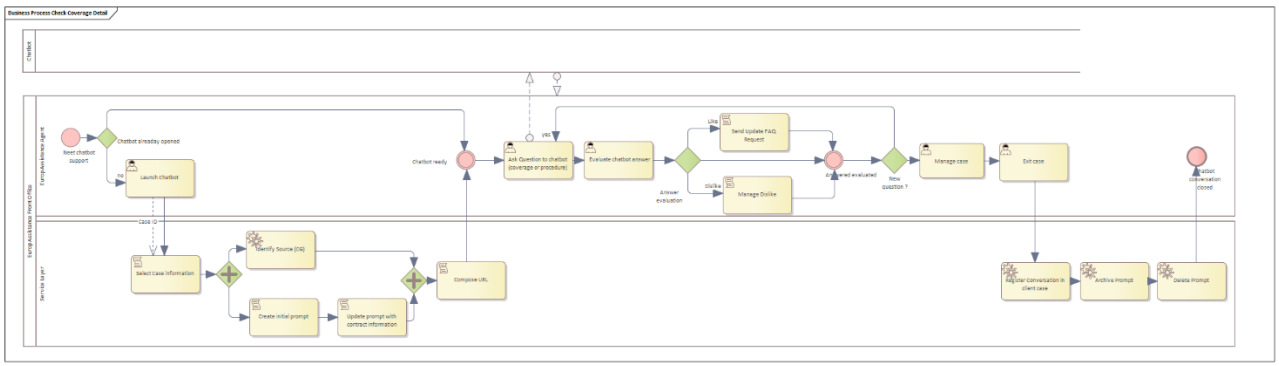


Figure 43: Processus de la solution avec les informations dans le prompt

Sur base des analyses détaillées que j'ai réalisées, une série de scripts ont été développés pour permettre de fournir au chatbot le contexte du dossier client (voir annexe 7 : Query Client's information).

Lorsque l'agent d'assistance clique sur le bouton du chatbot à partir de son système de gestion STAR, cela déclenche les étapes suivantes (voir Annexe 8 : Scripts du chatbot)

- la récupération des données du dossier client via une requête SQL sur base de l'ID du case (dossier) ;
- la sélection via un script des documents (sources) sur base du nom du produit (CG) ;
- en parallèle de la préparation des sources, la création du prompt (étape technique obligatoire) puis son update avec d'une part les instructions générales et d'autre part les données spécifiques au dossier ;
- l'ensemble des données ainsi récupérées est ajouté à l'URL de manière à ce que la conversation se fasse dans un contexte personnalisé.

Le chatbot est prêt et l'agent d'assistance peut poser des questions soit sur les couvertures, soit sur les procédures à suivre.

Chaque réponse reçue peut être évaluée par l'agent. Si l'agent considère que la réponse a un intérêt général, il peut la liker. Ces réponses seront dès lors reprises dans une liste de FAQ potentielles (autre processus). Si la réponse est erronée, il peut la disliker. Cette réponse sera analysée par les gestionnaires du chatbot pour améliorer celui-ci (amélioration continue) via un processus spécifique.

Lorsque l'agent a terminé son travail et qu'il quitte le dossier, l'ensemble de la conversation est stocké dans le dossier du client, le prompt lui-même est archivé (changement de statut qui le rend non visible), puis supprimé (suppression physique).

4.11 Formation

Même si le système a été conçu pour être le plus intuitif possible et qu'une enquête a montré qu'une majorité d'agents étaient positifs à l'introduction d'un chatbot, il est essentiel de les former, car cela impacte les procédures de travail.

On remarque dans le tableau ci-dessous que les agents plus expérimentés portent moins d'intérêt pour le chatbot que les autres. Cela est en partie expliqué par le fait qu'ils se sont habitués avec le temps à utiliser Athena malgré sa faible ergonomie.

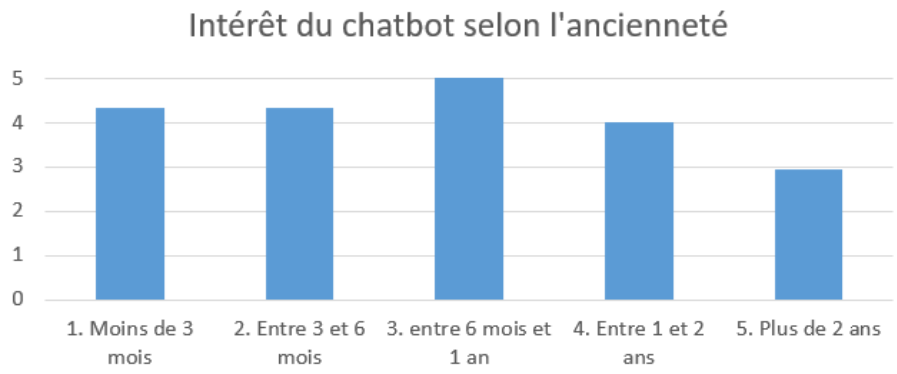


Figure 44: Intérêt du chatbot selon l'ancienneté (Europ Assistance 6, 2025)

La formation pour les agents est estimée à 2 heures et devra permettre de :

- leur présenter l'interface ;
- leur expliquer le fonctionnement du chatbot et la manière dont il est personnalisé ;
- leur expliquer comment formuler les questions pour obtenir des réponses pertinentes ;
- leur montrer pourquoi et comment évaluer les réponses ;
- leur montrer comment retrouver l'historique des conversations.

Etant donné le rôle de support première ligne des « learning specialists », une formation plus avancée devra être prévue, en particulier concernant la formulation des questions et l'évaluation des réponses.

4.12 MAINTENANCE, SUIVI, AMÉLIORATION CONTINUE

Comme pour toute application EA, un problème technique devra être signalé via un ticket aux équipes de support. Aujourd'hui, les niveaux de services associés au chatbot ne sont pas encore définis.

Au niveau fonctionnel, un principe d'amélioration continue est prévu via l'analyse des « dislike » et via un formulaire qui sera mis à disposition des agents pour collecter leurs problèmes et leurs suggestions d'améliorations.

Un suivi des KPI (cf infra section 4.6) devra permettre de vérifier que les hypothèses du ROI sont correctes et de prendre des mesures correctives lorsque nécessaire.

Les technologies autour de l'IA évoluent très rapidement. Un éveil technologique par rapport à ces évolutions (nouveaux modèles d'extraction de texte, de chunking, d'embedding, ou de modèles LLM) devrait être mis en place pour permettre une optimisation du chatbot (amélioration de la qualité et réduction des coûts).

5 BILAN ET PERSPECTIVE DU PROJET

5.1 ANALYSE CRITIQUE ET MISE EN PERSPECTIVE

5.1.1 Evaluation des résultats obtenus et degré de réalisation des objectifs

5.1.1.1 Statut du projet

Dans son état actuel, la solution chatbot intégrée dans STAR tourne sur l'environnement de test. Les tests utilisateurs n'ont pas encore démarré en raison d'un problème de clé au niveau des appels API.

Au niveau de la mise à disposition des données sources, seule une partie des données a été traitée. On doit dès lors considérer la solution comme un POC et non comme une solution prête pour la mise en production.

Pour pouvoir démarrer réellement en production, les étapes suivantes doivent encore être réalisées :

- préparation des sources de données et upload ;
    - les outils permettant de préparer les sources pour le contrôle de couverture sont disponibles mais les résultats devront être vérifiés manuellement (estimation de 5 mandays pour les CG), puis chargés dans le système (automatique),
- réception d'une API-key de la Holding pour permettre l'utilisation des API ;
- formation des utilisateurs ;
- mise en place du support.

5.1.1.2 Difficultés rencontrées et limitations

Mon ambition était de déployer la solution en production avant la fin de mon stage. Néanmoins, plusieurs éléments ont engendré du retard.

- Niveau de priorité du projet
    - La mise en œuvre du chatbot n'est pas considérée par EAB comme un projet prioritaire. Les ressources sont dès lors assignées selon le principe du best effort. Les développeurs étant souvent indisponibles, le planning souhaité pour l'intégration du chatbot, n'a pas pu être respecté.
    - L'intégration étant sur le chemin critique du projet, ce retard a également eu un impact sur l'organisation des formations et du support.
- Manque de réactivité et/ou de ressources de la Holding
    - La solution se base sur une interface et des API fournies par la Holding. Si cette mise à disposition a permis à EAB de lancer son initiative de chatbot, la Holding n'a pas apporté le support nécessaire pour la mise en œuvre effective : délais trop importants pour la mise à disposition des outils de tests à grande échelle (benchmark), manque de réactivité pour corriger les bugs, délais trop importants pour fournir les clés d'autorisation, refus d'envisager la mise à disposition d'autres types de RAG ou la modification de certaines API (ajout du conversation id dans l'URL), ...

En ce qui concerne la qualité des résultats du chatbot, j'ai souligné dans la section 4.4 l'importance de la qualité des sources. L'effort résiduel pour traiter les conditions générales propres à EAB est limité (voir les 5 mandays cités ci-dessus). Par contre, la préparation des procédures stockées dans Athena sous forme de page HTML nécessiterait un effort important. Aucun budget n'est actuellement disponible pour réaliser cette activité. Bien que le coût d'investissement serait moindre, le gain de temps estimé est plus faible car le chatbot ne pourrait pas traiter les procédures. Ce nouveau ROI serait de 63% sur la même période et atteindrait le break-even point fin février. Cependant en termes de profit, on passerait de 61000 euros à seulement 25000 euros soit près de 36000 euros en moins. (Voir Annexe 4.2 : ROI corrigé).

Le temps de réponse moyen actuel pour une question adressée au chatbot est de 5 secondes. Cette limitation est liée au modèle utilisé. Une optimisation ne pourrait être réalisée qu'en utilisant des modèles plus performants, mais également plus coûteux. Si EAB envisage une telle piste, un nouvel ROI devra être calculé.

La solution mise en œuvre se base sur l'ouverture d'un chatbot à partir de l'application STAR. Une intégration plus forte dans STAR pourrait améliorer encore l'ergonomie. Cette solution n'a pas été retenue d'une part par manque de budget et d'autre part en raison de l'architecture vieillissante de l'application STAR.

5.1.1.3 Conclusion

Différentes pistes d'amélioration de la solution mise en œuvre existent, mais nécessiteraient un investissement plus conséquent. Néanmoins, il s'agit d'une initiative intéressante qui permettra à EAB de faire un premier pas dans la découverte des nouvelles technologies de l'IA.

5.1.2 Recensement des difficultés rencontrées sur les plans :

5.1.2.1 Méthodologique

Historiquement, EAB travaille avec une méthode de gestion de projet Waterfall. Cependant, pour l'optimisation du chatbot, mon manager a souhaité appliquer une méthode de type Agile sans pour autant que tous les rôles soient remplis ni que toutes les best practices soient réellement suivies. En pratique, l'équipe agile n'était composée que d'un autre stagiaire et de moi-même.

- Pas de « product owner » désigné par EA, ce qui va à l'encontre de l'objectif de la méthode Agile, à savoir une étroite collaboration entre le business et les personnes en charge de l'analyse et/ou du développement.
- Pas de respect de la comitologie agile.
    - Daily scrum remplacé par des échanges ad-hoc,
    - Sprint review toutes les 3 semaines entre les 2 stagiaires,
    - Sprint backlog sans la présence du product owner ni de la holding,
    - Participation à des réunions weekly rassemblant les représentants de tous les projets de l'équipe CCA durant lesquelle un statut est donné aux managers et les difficultés ou demandes de ressources peuvent être remontées.

5.1.2.2 Humain

Un des défis du projet est l'adoption de la nouvelle technologie par les équipes. A ce sujet, une enquête montre que les points de vue sont très différents. Certaines personnes sont totalement réticentes et craignent que cette technologie remplace à terme les agents ou rende leur travail inintéressant. A l'opposé, d'autres agents sont enthousiastes et voient cette solution comme un moyen de soulager leur charge de travail, surtout durant les périodes de pic.

Quoi qu'il en soit, il est important que des formations soient organisées afin de leur expliquer comment fonctionne l'outil et la technologie sous-jacente, et surtout comment l'utiliser de manière optimale.

L'objectif du management est certes de limiter les coûts, mais plus dans l'optique d'être capable de faire face à l'augmentation des appels à effectifs équivalents que dans l'optique d'une diminution du personnel. Etant donné le métier d'Europ Assistance, le contact avec le client reste un gage de qualité.

5.1.2.3 Technique

Un seul modèle RAG a été mis à disposition par la Holding, ce qui a limité les possibilités d'optimisation. L'utilisation de modèles hybrides aurait en effet permis d'augmenter la qualité des résultats.

De même, seule la recherche par similarité de type cosinus (recherche sémantique) a été mise à disposition pour les vérifications de couverture et de procédure, alors qu'une recherche hybride (combinaison d'une recherche lexicale avec une recherche sémantique) aurait permis d'optimiser les résultats. Quant à la recherche dans les FAQ, seule la recherche lexicale a été mise à disposition alors que la recherche hybride aurait également amélioré les résultats.

Pour optimiser le chatbot, une demande a été adressée à la holding afin de mettre à disposition un script pour traiter les requêtes en batch et accélérer ainsi les tests. Un script a effectivement été mis à disposition, mais tardivement et durant une période trop limitée pour que ce script apporte réellement une plus-value.

Enfin, les API proposées par la Holding ne permettent pas de passer en paramètre l'identifiant de la conversation, ce qui nous a obligés à trouver une alternative, à savoir générer un nouveau prompt pour chaque conversation, à le mettre à jour avec le contexte du dossier puis à archiver et à supprimer le prompt (4 appels d'API au lieu d'un seul). Ce point est détaillé dans la section 4.10.

Comme déjà indiqué préalablement, l'autre enjeu a été le nettoyage des sources. Idéalement, des outils de gestion plus évolués auraient dû être disponibles pour que l'on dispose immédiatement de ces sources dans un format exploitable et avec une qualité suffisante. Ceci est particulièrement le cas pour les procédures stockées dans Athena sous format HTML.

5.1.3 Proposition de pistes d'amélioration

5.1.3.1 Structure et format des CG

Aujourd'hui, les CG sont rédigées selon des critères marketing, pour répondre aux besoins des clients, et leur structure est très diverse en fonction des partenaires.

Pour être efficace, le chatbot a besoin que ses documents sources, dont le CG, aient une mise en page standardisée :

- une structure simple et organisée ;
- un contenu sans références ;
- un format PDF.

Idéalement, il faudrait disposer d'un référentiel unique capable de générer des formats différents selon la finalité : un format pour le client et un format pour le chatbot.

La mise en place d'une telle solution peut cependant s'avérer très onéreuse, et Europ Assistance n'a pas dégagé aujourd'hui les budgets nécessaires à une telle transformation.

Il serait néanmoins intéressant d'intégrer cette réflexion lors de l'élaboration de nouveaux produits.

5.1.3.2 Format des procédures

De la même manière, les procédures devraient être disponibles à la fois sous forme de page HTML Athena et sous forme de PDF pour le chatbot.

5.1.3.3 Modèles de chunking

Les modèles de chunking utilisés dans la solution sont suffisants pour autant que les données sources soient correctement nettoyées, en particulier au niveau de la suppression des référencements. Etant donné la qualité effective des sources, les modèles agentiques seraient plus efficaces. Le modèle agentique mis à disposition par la holding n'est cependant pas de qualité suffisante et devrait être amélioré.

5.1.3.4 Retrieval

La méthode de retrieval actuelle par similarité, mise en place actuellement, présente des limites lorsque les chunks sont trop grands (plus le chunk est grand, plus on perd en précision).

Pour résoudre ce problème, une approche serait de :

- générer deux vecteurs par chunk ;
    - Un vecteur basé sur ADA2 (modèle sémantique),
    - Un vecteur basé sur TF-IDF (modèle lexical),
- effectuer deux recherches par similarités en générant également 2 vecteurs de recherche ;
- combiner les résultats (union, intersection, … : à affiner selon les résultats souhaités).

Cette même approche serait intéressante à la fois pour les contrôles de couvertures, pour les procédures et pour les FAQ.

5.1.3.5 RAG

Aujourd'hui, seul le RAG simple avec mémoire est utilisé. Un RAG est modulaire par essence et le RAG simple actuel pourrait être combiné avec d'autres modules comme un « CRAG » ou un « Self RAG ».

Le self RAG permettrait de répondre en particulier au problème de référencement en permettant de lancer une nouvelle récupération (les textes référencés dans notre cas) lorsqu'il lui manque des informations pour générer sa réponse.

Un modèle CRAG permettrait d'améliorer le traitement des chunks en les découpant en segments et en se focalisant uniquement sur les segments pertinents.

5.1.3.6 Intégration

Le mécanisme actuel permettant de personnaliser le système nécessite 4 appels d'API : 2 appels pour créer et mettre à jour le prompt et, en fin de conversation, 2 appels pour archiver et supprimer le prompt.

Une solution alternative permettrait de remplacer toutes ces étapes par l'appel d'une seule API. Un conversation ID serait passé dans le « request body » de l'API, puis ajouté à l'URL permettant d'envoyer un message dans une conversation spécifique, puis d'ouvrir directement cette conversation via l'URL personnalisée. De plus, en cas de rafraichissement de la page, cela n'ouvrirait pas de nouvelle conversation.

Le schéma ci-dessous reprend l'ensemble des interactions entre le chatbot et les autres systèmes de gestion de manière à le personnaliser au contexte du dossier, à évaluer les réponses et à historiser les résultats, mais avec la méthode alternative.

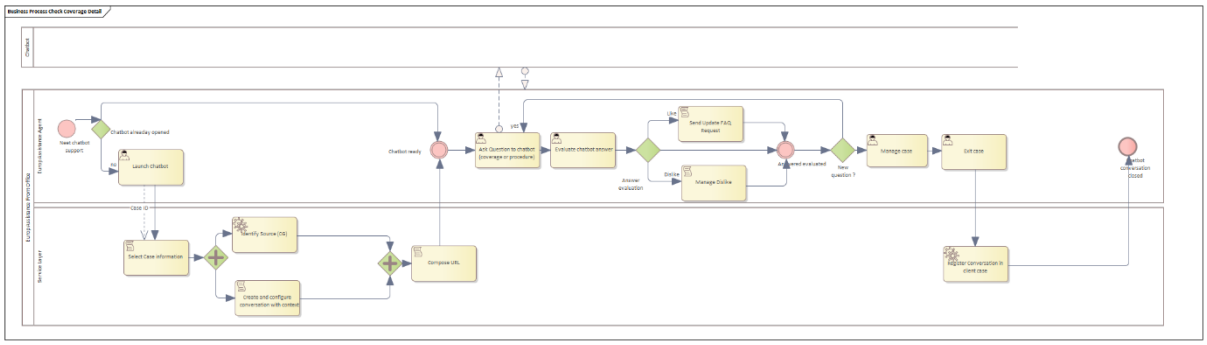


Figure 45 : Processus de la solution avec les informations envoyées par message

5.2 PERSPECTIVES DU PROJET

5.2.1 Inscription du projet au niveau stratégique

La Holding investit dans l'intelligence artificielle et met différents outils à disposition de ses filiales. Europ Assistance Belgique profite de cette opportunité pour renforcer ses objectifs stratégiques, à savoir automatiser au maximum ses processus tout en garantissant le meilleur service pour ses clients. Dans le cadre du projet, il s'agit du processus de recherche d'information par les agents d'assistance.

5.2.2 Etendre le scope du projet

5.2.2.1 Chatbot accessible au client

La mise à disposition du chatbot pour le client final permettrait à Europ Assistance de diminuer le nombre d'appels non lié à des sinistres. Le client trouverait plus facilement les réponses à ses questions qu'en consultant une FAQ.

L'approche serait de limiter les documents sources aux seules CG et aux FAQ. Seul un faible effort serait nécessaire pour configurer ce chatbot client.

Une diminution du nombre d'appels permettrait de diminuer les temps d'attente, d'améliorer les KPI associés, d'assurer une meilleure satisfaction à la fois des partenaires et des clients finaux. Cette amélioration de la satisfaction pourrait se refléter dans les scores NPS et attirer de nouveaux clients.

Ce même chatbot pourrait être intégré dans des phases ultérieures aux sites des partenaires d'Europ Assistance (pour autant que les CG de ceux-ci soient chargées dans le système).

5.2.2.2 Chatbot mis à disposition du service claims

Le contrôle des couvertures est également essentiel dans le travail du service claims. La mise à disposition d'un chatbot capable d'adresser des questions sur base des CG pourrait également être bénéfique.

Les gestionnaires claims utilisent par contre une autre application que les chargés d'assistance. La partie intégration doit dès lors être refaite.

5.2.2.3 Chatbot pour la formation

Une autre utilisation potentielle du chatbot serait la mise à disposition des supports de formation. Cela permettrait aux nouveaux employés de poser facilement leurs questions même en l'absence de formateur.

5.2.3 Développement futur de nouvelles fonctionnalités

L'intégration du chatbot dans les applications de gestion pourrait être encore améliorée. Le système pourrait adresser automatiquement les questions au chatbot sur base du type de sinistre et intégrer le résultat dans le dossier. Si le client est couvert, une deuxième question pourrait collecter la procédure d'application.

Une nouvelle fonctionnalité du chatbot, ou plutôt du LLM, pourrait permettre de résumer les CRC (comptes rendus des actions précédemment prises dans un dossier). Ceci serait particulièrement intéressant pour les dossiers complexes qui contiennent un grand nombre de CRC comme les dossiers médicaux.

De plus, il serait envisageable d'aller encore plus loin en intégrant des fonctionnalités « speech to text » qui permettraient de capter les échanges entre l'agent et le client, en complétant directement les informations nécessaires dans le dossier pour fournir ensuite une réponse détaillée en fonction de la demande.

CONCLUSION

Europ Assistance, comme de nombreuses entreprises, a souhaité mettre en œuvre une solution basée sur l'intelligence artificielle pour améliorer le service offert à ses clients et augmenter sa productivité. La Belgique a profité d'une interface développée par la Holding pour lancer un projet de chatbot à destination des chargés d'assistance.

Les technologies basées sur l'intelligence artificielle font rêver, mais ne constituent pas une solution miracle. Il est important de rester réaliste lorsque l'on calcule le ROI. Il ne faut ni sous-estimer le temps nécessaire pour configurer de tels systèmes, ni surestimer les gains réels. Dans le cadre du projet, une des grandes difficultés rencontrées a été la qualité des documents utilisés par le chatbot pour répondre aux requêtes des utilisateurs. Une mauvaise qualité ou une structure trop complexe des documents freine la qualité des résultats. C'est pour cette raison qu'en cours de projet, le scope a été réduit en excluant les conditions générales des partenaires jugées trop complexes et que le traitement des procédures Athena a été postposé.

Quant aux gains, ils vont dépendre de la performance du système et de l'acceptation de la solution par les utilisateurs.

Dans ce contexte, la décision de suivre les activités de paramétrisation du chatbot en mode agile a été judicieuse, puisqu'elle a permis d'ajuster le scope au fur et à mesure.

D'un point de vue projet, la Holding doit être considérée comme le fournisseur de la solution. Contrairement à ce qui aurait été fait avec un fournisseur externe, aucun contrat n'a été signé et dès lors aucun SLA fixé pour le traitement des tickets introduits, qu'ils concernent des corrections ou des demandes d'évolutions. Malgré les « escalations » introduites via le sponsor, ce manque de réaction rapide a eu des impacts négatifs sur le planning.

On peut également noter que le projet n'a pas été géré comme un projet prioritaire. Marketing n'a pas obtenu de budget pour améliorer la qualité des documents sources et l'intégration du chatbot dans l'application STAR a été planifiée en « best effort », ce qui a également eu des conséquences sur le planning.

Le planning initial, prévoyant une mise en production avant la fin du stage, n'a pas pu être respecté. Même si les tests d'intégration ont été concluants, étant donné que les UAT (user acceptance test) n'ont pas encore débuté, il n'a pas été possible, dans le cadre de ce travail, de vérifier si les bénéfices escomptés ont pu totalement ou partiellement être atteints.

Un dernier point d'attention est le change management. Malgré une enquête montrant qu'une majorité d'agents d'assistance voit l'introduction du chatbot de

manière positive, les agents les plus expérimentés pourraient considérer cette technologie comme une menace pour leur travail. Le change management requiert une attention constante non seulement pendant le projet, mais également après la mise en production.

D’un point de vue personnel, ce stage m’a permis de mettre en pratique les connaissances acquises durant mes études et m’a apporté une belle expérience, à la fois humaine, méthodologique et technologique.

BIBLIOGRAPHIE

- ASQ. (s.d.). Problem Solving. https://asq.org/quality-resources/problem-solving?srsltid=AfmBOoog6nK8qlHutvAZyQuRzwZmCuphMNF3gsuAp0qmqMU09-LXuqXs
- Assuralia 1. (2024, 2 septembre). *Top 15 assistance avec part de marché*. https://files.assuralia.be/stats/FR/04_composition-du-marche/04_05_top15-assistance.htm
- Assuralia 2. (2024, 22 octobre). *Chiffres clés pour la branche assistance*. https://files.assuralia.be/stats/FR/02_chiffres-cles-par-branche/02_05_chiffres-cles-assistance.htm
- Barnard, J. (2023, 22 décembre). What is embedding? IBM. https://www.ibm.com/think/topics/embedding
- Belcic, I. (2024, 21 octobre). *What is RAG (retrieval augmented generation)?* . IBM. https://www.ibm.com/think/topics/retrieval-augmented-generation
- Bengesi, S., El-Sayed, H., Houkpati, Y., Irungu, J., Oladunni, T. & Sarker, MD K. (2024, 6 mai). Advancements in Generative AI: A comprehensive review of GANs, GPT, autoencoders, diffusion model, and transformers. IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/10521640
- Cambridge Dictionary 1. *(2025). artificial intelligence*. https://dictionary.cambridge.org/dictionary/english/artificial-intelligence
- Cambridge Dictionary 2. (2025). Chunking. https://dictionary.cambridge.org/fr/dictionnaire/anglais/chunking
- Cambridge Dictionary 3. (2025). *KPI*. https://dictionary.cambridge.org/dictionary/english/kpi
- ChatGPT info. (2025, 1 février). *Quelle est la taille du modèle GPT-4 ?* ChatGPT Info. https://chatgpt-info.fr/taille-modele-gpt4/
- Chen, S., Chen, T., Ma, K., Wang, H., Yu, D, Yu, W., Zhang, H. & Zhao, X. (2024, 4 octobre). Dense X retrieval: What retrieval granularity should we use?. In EMNLP 2024 Main Conference [Conference-proceeding]. https://arxiv.org/pdf/2312.06648
- Data Camp. (s.d.). *Cosinus Similarité*. https://assets.datacamp.com/production/repositories/4966/datasets/ec0fa4795484baf3a19c3f345755a85457a2aae4/cosine.png
- Davies, A. (2025, April 11). What is Agile Methodology? - DevTeam.Space. DevTeam.Space. https://www.devteam.space/blog/what-is-an-agile-methodology/
- eKomi. (2025, 16 février). *Europ Assistance Belgium*. Consulté 18 avril 2025. Avis clients Europ Assistance Belgium – avis francophones - Notation : 4.5 sur la base de 3527 avis clients et expériences pour europ-assistance.be/fr
- Esmaielbeiki, R. (2023, 28 novembre). *BERT, GPT and BART: a short comparison*. Medium. https://medium.com/@reyhaneh.esmailbeigi/bert-gpt-and-bart-a-short-comparison-5d6a57175fca
- Europ Assistance 1. (2024). Assistance *Leakage* [Mail]. Europ Assistance
- Europ Assistance 2. (2024). *Présentation de l'entreprise* [PowerPoint]. Europ Assistance
- Europ Assistance 3. (2024, 10 décembre). [Facture]. Europ Assistance
- Europ Assistance 1. (2025). *Organigramme* [Page HTML]. Europ Assistance

- Europ Assistance 2. (2025). Modèles d'extraction et chunking. [Pdf]
- Europ Assistance 3. (2025). *Qui sommes-nous ?*. https://www.europ-assistance.fr/fr/qui-sommes-nous
- Europ Assistance 4. (2025, 31 janvier). *About us | Europ Assistance*. https://www.europ-assistance.com/about-us/
- Europ Assistance 5. (2025, 31 février). [Power BI]. Europ Assistance
- Europ Assistance 6. (2025, 31 mars). [Enquête]. Europ Assistance
- European Parliament, (2025, 19 février).EU *AI Act: first regulation on artificial intelligence | Topics*. (2023, August 6). Topics | European Parliament. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence
- Farley, P. (2024, 17 octobre). OCR - Optical Character Recognition - Azure AI services. Microsoft Learn. https://learn.microsoft.com/en-us/azure/ai-services/computer-vision/overview-ocr#ocr-engine
- Farnschläder, T. (2025, 27 janvier). AI Hallucination : Un guide avec des exemples. https://www.ibm.com/think/topics/retrieval-augmented-generation
- Ferrer, J. (2024. 9 janvier). How Transformers Work: A Detailed Exploration of Transformer Architecture. https://www.datacamp.com/tutorial/how-transformers-work
- Gite, S., Rawat, U., Kumar, S., Saini, B., Bhatt, A., Kotecha, K., & Naik, N. (2024, 19 août). *Unfolding Conversational Artificial Intelligence: A Systematic review of datasets, techniques and challenges in developments*. https://www.espublisher.com/journals/articlehtml/engineered-science/10.30919-es1210
- Grigore. (2025, 31 mars). *What is a Good Net Promoter* Score? (2025 NPS Benchmark). Retently CX. https://www.retently.com/blog/good-net-promoter-score/
- Gutowska, A. (2025, 20 janvier). Implement RAG chunking strategies with LangChain and watsonx.ai. IBM. https://www.ibm.com/think/tutorials/chunking-strategies-for-rag-with-langchain-watsonx-ai?mhsrc=ibmsearch_a&mhq=chunking%20strategies%20for%20rag%20tutorial
- IBM 1. (2024, 18 avril). Qu'est-ce que la reconnaissance optique de caractères (OCR) ?. https://www.ibm.com/fr-fr/think/topics/optical-character-recognition
- IBM 2. (2025). *Qu'est-ce qu'un chatbot ?*. https://www.ibm.com/fr-fr/think/topics/chatbots
- Ionos, (2023, 23 octobre). *Le Prompt Engineering : explication*. https://www.ionos.fr/digitalguide/sites-internet/creation-de-sites-internet/prompt-engineering/
- Jac. (2025, 22 février). Qu'est-ce qu'une API ? Définition, Types et fonctionnement. *Explorez le web, IA, Technologies, Outils innovants, Derniers produits High-tech*. https://critiqueplus.com/technologie/quest-ce-quune-api-definition-types-et-fonctionnement/
- Jumelle, M. (2024, 2 février). Fine-tuning de LLM : tout savoir. *Formation Tech et Data en ligne | Blent.ai*. https://blent.ai/blog/a/fine-tuning-llm
- Karpathy, A. (2023, 17 janvier). *Let's build GPT: from scratch, in code, spelled out*. [Video]. YouTube. https://www.youtube.com/watch?v=kCc8FmEb1nY

- Kavlakoglu, E., Stryker, C. (2024, 9 août). *Qu'est-ce que l'intelligence artificielle (IA) ?*. IBM. https://www.ibm.com/fr-fr/think/topics/artificial-intelligence
- Kavlakoglu, E., Vaish, R. (2020). *NLP vs. NLU vs. NLG: the differences between three natural language processing concepts*. IBM. https://www.ibm.com/think/topics/nlp-vs-nlu-vs-nlg
- Kelly, C. (2025, 1 février). *8 Retrieval Augmented Generation (RAG) Architectures you should know in 2025*. Humanloop. https://humanloop.com/blog/rag-architectures
- Larousse, É. (2025). *Définitions : chatbot - Dictionnaire de français Larousse*. https://www.larousse.fr/dictionnaires/francais/chatbot/188506
- Mbiya, J.C. (s.d.). *API REST : HTTP Status Codes et leurs significations*. Letecode. https://www.letecode.com/api-rest-http-status-codes-et-leurs-significations
- Métais, J.-F. (2025). Prix et performance Marketing [Syllabus imprimé]. ICHEC
- Microsoft Azure (s.d.). *Azure OpenAL Service – Tarification*. Consulté 15 mars 2025 https://azure.microsoft.com/fr-fr/pricing/details/cognitive-services/openai-service/
- Microsoft 1. (2025). Copilot [Grand Modèle linguistique]. https://m365.cloud.microsoft/chat/?auth=2
- Microsoft 2. (2025). *What are large language models (LLMs)?*. https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-are-large-language-models-llms?msockid=13c166e8b4f968701c127200b5b169bf
- Mrbullwinkle. (2025, 26 mars). Azure OpenAI Service embeddings tutorial - Azure OpenAI. Microsoft Learn. https://learn.microsoft.com/en-us/azure/ai-services/openai/tutorials/embeddings?tabs=python-new%2Ccommand-line&pivots=programming-language-powershell
- Murel, J., Noble, J. (2024, 16 décembre). What is LLM Temperature ?.IBM. https://www.ibm.com/think/topics/llm-temperature
- Negre, E. (2013, 17 octobre). *Comparaison de textes: quelques approches*... In LAMSADE [Report]. https://hal.science/hal-00874280/document
- Nollevaux, G. (2024). *Gestion de projet informatique Gestion de projets digitaux* [Syllabus imprimé]. ICHEC
- NVIDIA (s.d.). What is a Vector Database?. https://www.nvidia.com/en-eu/glossary/vector-database/
- OpenAI Help Center. (s.d.). *Best practices for prompt engineering with the OpenAI API*. https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-the-openai-api
- Oracle 1. (2025). *Explore Chunking Techniques and Examples*. Oracle. https://docs.oracle.com/en/database/oracle/oracle-database/23/vecse/explore-chunking-techniques-and-examples.html
- Oracle 2. (2025, 23 avril). *Generate Vector Embeddings*. https://docs.oracle.com/en/database/oracle/oracle-database/23/vecse/generate-vector-embeddings-node.html
- Oracle 3. (2025, 23 avril). *Perform Exact Similarity Search*. https://docs.oracle.com/en/database/oracle/oracle-database/23/vecse/perform-exact-similarity-search.html

- Oracle 4. (2025, 23 avril). *Perform Multi-Vector Similarity Search*. https://docs.oracle.com/en/database/oracle/oracle-database/23/vecse/perform-multi-vector-similarity-search.html
- Oracle 5. (2025, 23 avril). *Understand Approximate Similarity Search Using Vector Indexes*. https://docs.oracle.com/en/database/oracle/oracle-database/23/vecse/understand-approximate-similarity-search-using-vector-indexes.html
- Oracle 6. (2025, 23 avril). *Understand Hybrid search*. https://docs.oracle.com/en/database/oracle/oracle-database/23/vecse/understand-hybrid-search.html
- Plutora. (2019, 18 septembre) *Water-Scrum-Fall is a real prevalent phenomen*. https://www.plutora.com/blog/water-scrum-fall
- Sharma, H. (2024, 19 juin). *Softmax Temperature*. Medium. https://medium.com/@harshit158/softmax-temperature-5492e4007f71
- Staff 1, C. (2025, 18 mars). What is TF-IDF? Coursera. https://www.coursera.org/articles/what-is-tfidf
- Staff 2, C. (2025, 31 mars). *Service-Level Agreement (SLA): why it's important and how to write one*. Coursera. https://www.coursera.org/articles/sla?msockid=13c166e8b4f968701c127200b5b169bf
- Tankoua Yojuen, W. I., (2025, 29 avril). Architecture Chatbot [Interview]. Europ Assistance
- Teaganne, F. (2023, 5 septembre). *5 types de chatbots et comment choisir le bon pour votre entreprise*. IBM. https://www.ibm.com/fr-fr/think/topics/chatbot-types
- UCLouvain, (2024, 13 mars). *Applications Didactiques de Physique Les vecteurs*. https://sites.uclouvain.be/didac-physique/didacphys/rappels/math/vecteurs.html
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Google Brain, Google Research, Gomez, A. N., University of Toronto, Kaiser, Ł., & Polosukhin, I. (2023). *Attention is all you need*. 31st Conference on Neural Information Processing Systems (NIPS 2017). https://arxiv.org/pdf/1706.03762.pdf
- Wolford, B. (2024, 29 août). *What is GDPR, the EU's new data protection law?* GDPR.eu. https://gdpr.eu/what-is-gdpr/

GLOSSAIRE

API	Application Programming Interface
Athena	Bibliothèque des connaissances interne de EAB
CES	Customer Effort Score
DPIA	Data Protection Impact Assessment
EA	Europ Assistance (Group)
EAB	Europ Assistance Belgique
GPT	Generative Pre-Trained Transformer
IA	Intelligence artificielle
KPI	Key Performance Indicator
LLM	Large language Model
LS	Learning Specialist
NLG	Nalutal Language Generative
NLP	Natural Language Processing
NLU	Natural Language Understanding
NPS	Net Promoter Score
PSS	Problem Solving Solution
RAG	Retrieval Augmented generation
RGPD / GDPR	Règlement Général sur la Protection des Données
SLA	Service Level Agreement
STAR	Outil de travail interne pour l'assistance de EAB
TDD	Test Driven Development
TF-IDF	Term frequency – Inverse document frequency
UAT	User Acceptance Test
UE	Union Européenne