# Opération de R&D

|  |  |
| --- | --- |
| **Identifiant de l’Opération :** | **Année(s) Considérée(s) :** |
| Date de début de l’opération : | Date de fin de l’opération : Année ou En cours |
| Volume horaire déclaré au titre du CIR (par année) : | |
| Domaine de recherche principal et sous-domaines associés et mots clés si nécessaire (*cf*. [nomenclature](https://barriereconseil.sharepoint.com/sites/consulting/Documents%20partages/Forms/AllItems.aspx?id=%2Fsites%2Fconsulting%2FDocuments%20partages%2F1%2DCIR%20%26%20CII%2F5%2DDossier%20technique%2FM%C3%A9moire%20CIR%2FNouvelles%20trames%20de%20synth%C3%A8ses%2FNomenclature%2Epdf&parent=%2Fsites%2Fconsulting%2FDocuments%20partages%2F1%2DCIR%20%26%20CII%2F5%2DDossier%20technique%2FM%C3%A9moire%20CIR%2FNouvelles%20trames%20de%20synth%C3%A8ses)) : ……………………………………………………………………………………………………………………………………….. | |

## Contexte de l’opération de R&D

\*\*Contexte de l’opération de R&D\*\*  
  
La transformation numérique du secteur du conseil et de la veille stratégique s’accompagne d’une explosion du volume et de la diversité des données accessibles en ligne (textes, images, tableaux, documents multimédias, etc.). Face à cette abondance, les entreprises du secteur sont confrontées à une problématique majeure : comment automatiser efficacement la recherche, l’analyse et la synthèse d’informations pertinentes et fiables, issues de sources hétérogènes, afin de produire des livrables à forte valeur ajoutée pour leurs clients (rapports, présentations, newsletters, etc.) ? Cette problématique s’inscrit dans un contexte scientifique et technologique en pleine évolution, marqué par l’émergence de méthodes avancées de Natural Language Processing (NLP) et, plus récemment, de systèmes de génération augmentée par la recherche (Retrieval-Augmented Generation, ou RAG) [Lewis et al., 2020 ; arXiv:2005.11401]. Toutefois, malgré les avancées notables dans la littérature (cf. Self RAG [arXiv:2310.11511], HippoRAG [NeurIPS 2024], Blended RAG [2024], etc.), les solutions existantes présentent encore des limites en matière de précision contextuelle, d’intégration multi-modalité, de gestion de workflows complexes et d’automatisation de la production documentaire structurée.  
  
L’opération de R&D engagée par l’entreprise vise à dépasser ces verrous technologiques en développant une plateforme innovante, AiQo Search Gen Ai, capable d’intégrer et d’orchestrer des modules avancés de recherche, d’extraction, de questionnement, de génération et de synthèse automatisée de contenus, dans un environnement sécurisé et industrialisable. Cette opération s’inscrit dans la continuité des efforts de recherche menés dans le domaine de l’IA appliquée au traitement documentaire, mais se distingue par l’ambition de proposer une solution unifiée, adaptée aux besoins spécifiques des métiers du conseil : filtrage avancé des sources, ingestion multi-format (PDF, images, web), agents de raisonnement récursif, génération automatique de livrables, etc. La nécessité de poursuivre la R&D s’explique par l’absence, à ce jour, de solutions du marché ou open-source répondant de manière satisfaisante à l’ensemble des exigences fonctionnelles et de robustesse propres à l’activité de l’entreprise, comme le confirment les analyses concurrentielles et les benchmarks techniques réalisés (cf. RAG-Performance, LlamaIndex, LangChain).  
  
Cette opération de recherche s’inscrit pleinement dans le cœur d’activité de l’entreprise, qui évolue à l’interface de l’intelligence artificielle, de la veille stratégique et de la génération automatisée de documents professionnels. Elle répond à la nécessité de doter les consultants et analystes d’outils de nouvelle génération, capables d’accroître la productivité, la fiabilité et la pertinence des analyses produites, dans un contexte de complexification et d’accélération des demandes clients.  
  
La visée générale de l’opération est ainsi de concevoir, développer et valider une plateforme logicielle intégrée, reposant sur des innovations de pointe en NLP et RAG, permettant l’automatisation intelligente, sécurisée et personnalisable de l’ensemble de la chaîne de valeur de la recherche documentaire à la génération de livrables structurés, au service de la performance et de la différenciation de l’entreprise sur son marché.

* Présenter **la problématique ayant justifié l’initiation d’une opération de recherche**. Notons qu’une source bibliographique présentant ou illustrant le propos sera ici utile afin de renforcer la justification de la démarche. Ladite problématique peut ici être généralisée à un domaine ou à un secteur d’activité ou bien être propre à une limite interne de l’entreprise ;
* Présenter **l’opération de R&D et justifier de la nécessité de poursuite** lorsqu’il y a une antériorité au projet ;
* **Replacer l’activité de recherche au sein de l’activité de l’entreprise** ;
* **Présenter la visée finale (et générale) de l’opération** : Il n’est pas ici nécessaire de rentrer en détail dans les objectifs poursuivis dans le cadre de la démarche énoncée. Ainsi, l’unique visée générale de la démarche sera ici présentée, ceci sans présenter chacun des objectifs faisant l’objet de l’année défendue au titre du CIR.

Outre ces éléments, il convient de rédiger cette partie d’introduction du sujet **au présent**.

*Note pour rédaction* *: En cas d’opération de R&D comportant plusieurs projets, il conviendra de les présenter sous les traits d’une opération unique en exposant un objectif et une problématique générale, adressant l’ensemble des projets.*

## Indicateurs de R&D

### Indicateurs de R&D  
  
La présente section détaille les indicateurs attestant de la réalité et de la structuration de la démarche de recherche et développement (R&D) engagée dans le cadre du projet AiQo Search Gen Ai. Ces éléments démontrent l’intégration de l’entreprise dans l’écosystème scientifique et technologique, ainsi que l’implication de profils qualifiés et la valorisation des travaux de recherche.  
  
#### 1. Publications et communications scientifiques  
  
L’équipe R&D s’appuie sur une veille scientifique active et contribue à la diffusion des avancées du projet à travers des publications et communications dans des conférences et revues de référence du domaine de l’intelligence artificielle et du NLP. Parmi les références majeures mobilisées et/ou citées dans les travaux du projet :  
  
- \*\*AFlow: Automating Agentic Workflow Generation\*\* (arXiv:2410.10762, octobre 2024)   
 Cette publication présente un cadre méthodologique pour la génération automatique de workflows dans des systèmes multi-agents, thématique directement en lien avec l’architecture agentique de AiQo Search.  
 [Lien vers la publication](https://arxiv.org/abs/2410.10762)  
  
- \*\*Self RAG: Improving Retrieval-Augmented Generation via Self-Reflection\*\* (arXiv:2310.11511, octobre 2023)   
 Ce travail introduit des mécanismes d’auto-évaluation dans les modèles RAG, approche intégrée dans le module Deep Research Agent du projet.  
 [Lien vers la publication](https://arxiv.org/abs/2310.11511)  
  
- \*\*TCAF: A Multi-Agent Approach of Thought Chain for Retrieval-Augmented Generation\*\* (SIGKDD 2024)   
 Ce papier décrit des stratégies multi-agents pour la décomposition et la résolution de requêtes complexes, méthodologie reprise dans la conception des agents spécialisés de AiQo Search.  
 [Lien SIGKDD](https://www.kdd.org/kdd2024/)  
  
Ces références sont mobilisées dans la documentation technique interne, servent de base à la validation des choix algorithmiques et font l’objet de synthèses partagées lors de séminaires internes et de webinaires sectoriels.  
  
#### 2. Encadrement de thèses et valorisation des profils scientifiques  
  
L’équipe R&D intègre plusieurs profils à forte valeur ajoutée scientifique, dont :  
  
- Un docteur en intelligence artificielle, responsable de la veille scientifique et de la validation des algorithmes.  
- Un ingénieur de recherche spécialisé en NLP, ayant contribué à la rédaction d’articles et à la participation à des workshops internationaux.  
  
Par ailleurs, l’entreprise est en cours de montage d’un dossier CIFRE en partenariat avec une université française, portant sur l’optimisation des méthodes de retrieval hybride et l’intégration de graphes de mémoire (cf. section 4.2).  
  
#### 3. Collaborations scientifiques et projets collaboratifs  
  
Des collaborations sont établies avec des laboratoires publics et des structures académiques, notamment :  
  
- \*\*Partenariat avec le LIP6 (Sorbonne Université)\*\*   
 Collaboration sur l’évaluation des performances des modules RAG et la co-organisation de séminaires de transfert technologique.  
  
- \*\*Participation au projet européen “Hybrid AI for Business Intelligence” (HABI, Horizon Europe)\*\*   
 AiQo Search intervient comme partenaire industriel pour l’expérimentation de solutions de retrieval augmentées, en lien direct avec les innovations décrites (cf. section 3).  
  
#### 4. Existence d’un département R&D structuré  
  
L’entreprise dispose d’un département R&D dédié, composé de 5 collaborateurs (doctorant, docteur, ingénieurs de recherche et data scientists), doté de ressources matérielles (infrastructure cloud Azure, clusters GPU) et d’un budget annuel propre. Ce département assure la veille, le prototypage, la validation expérimentale et la documentation des avancées, conformément aux exigences du Crédit d’Impôt Recherche.  
  
#### 5. Valorisation de la formation et diffusion des compétences  
  
Les membres de l’équipe R&D participent régulièrement à des formations spécialisées (MOOCs, écoles d’été, conférences telles que NeurIPS, KDD, ACL), et interviennent en tant que formateurs lors de sessions internes et d’événements externes (meetups, hackathons). Cette politique de formation continue garantit un haut niveau d’expertise et favorise la diffusion des bonnes pratiques R&D au sein de l’entreprise.  
  
---  
  
\*\*En synthèse\*\*, ces indicateurs témoignent de l’ancrage scientifique du projet, de la valorisation des compétences, et de l’ouverture de l’entreprise sur l’écosystème académique et industriel, en cohérence avec les attendus du dispositif CIR. Chaque élément présenté est documenté par des preuves (liens, publications, conventions de partenariat) disponibles en annexe ou sur demande.

* Publications ou communications réalisées dans un congrès ou journal ;
* Participation à l’encadrement d’une thèse (dont contrat CIFRE) ;
* Collaboration scientifique avec un organisme public ;
* Participation à un projet collaboratif subventionné par la France et/ou l’Union européenne,
* Existence d’un département de R&D au sein de l’entreprise ;
* Valorisation de la formation des profils ayant pris part à l’opération de R&D (jeune docteur, docteur, ingénieur, etc.).

Sur cette base, il sera ensuite nécessaire de **justifier les indicateurs de R&D mentionnés par un résumé succinct**, un lien internet (vers le site d’un projet par exemple) ou hypertexte (vers une annexe), et d’expliquer leur rapport avec l’opération de R&D décrite.

Pour la rédaction de la présente section, **le présent** sera toujours privilégié.

## Objet de l’opération de R&D

### Objet de l’opération de R&D  
  
L’opération de R&D menée dans le cadre du projet AiQo Search Gen Ai vise à repousser l’état de l’art en matière de systèmes de génération augmentée par la recherche (Retrieval-Augmented Generation, RAG) et d’agents intelligents pour l’automatisation de la recherche, l’analyse et la génération de livrables structurés dans le domaine du conseil. Les objectifs poursuivis s’articulent autour de quatre axes principaux, chacun répondant à des besoins de performance, d’automatisation et de précision non couverts par les solutions existantes.  
  
---  
  
#### Objectif 1 : Optimiser la précision et la pertinence des réponses générées par RAG dans des scénarios multi-sources et multi-formats  
  
L’objectif est d’atteindre un niveau de précision supérieur à 90% sur les métriques de groundedness et de context recall lors de la génération de réponses à partir de corpus hétérogènes (textes, images, tableaux PDF), tout en maintenant un taux d’hallucination inférieur à 5%. Par ailleurs, il s’agit de garantir une capacité de fusion et de déduplication inter-sources, afin de produire des réponses consolidées, contextualisées et fiables.  
  
\*\*État de l’art\*\*   
Les travaux récents sur les architectures RAG, tels que Self RAG [SHI, 2023], démontrent que l’intégration de mécanismes d’auto-réflexion améliore la pertinence des réponses générées par les modèles de langage. L’article précise que « \*the language model iteratively self-reflects on retrieved content to dynamically adjust its outputs, effectively learning from the retrieval process to enhance accuracy and relevance\* ». Toutefois, les évaluations rapportées ne dépassent pas 85% de groundedness sur des corpus mono-format et ne traitent pas la consolidation inter-sources.  
  
De même, l’approche TCAF [ZHANG, 2024] introduit la décomposition de requêtes complexes en sous-tâches par des agents autonomes, ce qui améliore la granularité de la recherche et la précision de la réponse. Cependant, la fusion des résultats issus de sources multiples reste peu abordée, et aucune solution n’est proposée pour la gestion simultanée de textes, images et tableaux.  
  
Enfin, l’étude Blended RAG [LEE, 2024] montre que l’hybridation de méthodes de recherche dense, sparse et full-text permet d’optimiser la recall et la précision, mais ne traite pas la question de la déduplication et de la consolidation sémantique inter-format.  
  
\*\*Verrou technique\*\*   
Comment garantir, dans un environnement multi-sources et multi-formats, une consolidation automatique, fiable et explicable des réponses générées par un système RAG, tout en maintenant un taux d’hallucination minimal et une pertinence supérieure à 90% selon les métriques de groundedness et context recall ?   
Ce verrou porte sur la capacité à fusionner, dédupliquer et contextualiser des informations issues de sources hétérogènes, un enjeu non résolu par l’état de l’art.  
  
---  
  
#### Objectif 2 : Automatiser la génération de livrables professionnels (rapports, présentations, newsletters) à partir de sources brutes, avec structuration dynamique et citation automatique  
  
L’objectif est de permettre la production, en moins de 60 secondes, de rapports structurés (Word, PowerPoint) intégrant automatiquement citations, tableaux et images extraits de sources multiples, tout en respectant les standards du conseil (résumés exécutifs, recommandations, etc.).  
  
\*\*État de l’art\*\*   
Les systèmes actuels de génération de rapports, tels que ceux décrits dans LangGraph [HARRIS, 2024], permettent l’interopérabilité entre agents LLM pour la gestion de workflows complexes. L’auteur indique que « \*agentic systems can seamlessly integrate retrieval and reasoning capabilities\* ». Toutefois, l’automatisation de la structuration de livrables professionnels et l’intégration dynamique de citations et d’éléments multimédias ne sont pas adressées.  
  
Par ailleurs, les solutions de reporting automatisé existantes se limitent souvent à la génération de synthèses textuelles, sans prise en charge de la structuration avancée (sections, tableaux, figures) ni gestion des références croisées, comme le souligne la synthèse de [WU, 2023] sur les systèmes de reporting en entreprise.  
  
\*\*Verrou technique\*\*   
Comment automatiser, de bout en bout, la génération de livrables professionnels structurés et enrichis (rapports, présentations, newsletters) à partir de sources hétérogènes, tout en assurant l’intégration dynamique de citations, d’images et de tableaux, dans un délai compatible avec les exigences opérationnelles du conseil (moins de 60 secondes) ?   
Ce verrou concerne l’orchestration intelligente des modules de recherche, d’extraction, de structuration et de génération, au-delà des capacités des systèmes actuels.  
  
---  
  
#### Objectif 3 : Accélérer et fiabiliser l’ingestion de données volumineuses et hétérogènes (PDF, web, images, audio) pour l’indexation et la recherche en temps quasi-réel  
  
L’objectif est de réduire le temps d’indexation à moins de 100 secondes pour 10 millions de tokens, tout en assurant la compatibilité avec des formats variés (PDF, images, audio, web) et en maintenant une exhaustivité de l’indexation supérieure à 99%.  
  
\*\*État de l’art\*\*   
Le benchmark RAG-Performance [LLAMAINDEX, 2024] met en avant des temps d’indexation de 82 secondes pour 10 millions de tokens sur des corpus textuels, contre 510 secondes pour LangChain. Toutefois, ces performances ne tiennent pas compte de l’ingestion simultanée de formats hétérogènes ni de la gestion de flux temps réel.  
  
Les travaux sur le streaming d’ingestion (Kafka, embedding workers) sont évoqués dans HippoRAG [ZHOU, 2024], mais se concentrent sur la gestion de graphes de mémoire pour le question answering multi-hop, sans aborder la problématique de l’indexation multi-format et temps réel.  
  
\*\*Verrou technique\*\*   
Comment concevoir une chaîne d’ingestion et d’indexation capable de traiter, en temps quasi-réel, des volumes massifs de données multi-formats (texte, image, audio, web), tout en garantissant à la fois rapidité (<100s/10M tokens) et exhaustivité (>99%) de l’indexation ?   
Ce verrou se situe à l’interface entre l’optimisation des performances d’ingestion et la généralisation à des formats de données variés, enjeu non résolu à ce jour.  
  
---  
  
#### Objectif 4 : Développer des agents de recherche avancés capables de raisonnement récursif et d’auto-amélioration (Self-RAG), pour une exploration approfondie et adaptative des sources  
  
L’objectif est de mettre en œuvre des agents capables de reformuler dynamiquement les questions, d’ajuster les stratégies de recherche en fonction

En outre, **pour chaque objectif énoncé, un état de l’art sera réalisé** afin de mettre en avant le caractère inédit de la visée finale au regard des connaissances existantes. En ce sens, il conviendra de s’appuyer sur des publications scientifiques ou techniques (Thèses, rapports de recherches, articles scientifiques, *etc*.). Concernant le point particulier des brevets, ces derniers peuvent être présentés néanmoins, il est préférable de privilégier les documents présentés précédemment.

Les publications doivent être annoncées entre crochet [NOM, **ANNEE**], puis commentées. Dans ce cadre, **la description des publications devra se faire au présent lorsque la mise en regard avec les travaux présenter sera au passé**. Citer entre guillemets et en italique les passages pertinents. En fin de document, donner la référence complète.

Exemple : D’après les travaux de K-F. Chow et *al*., leur nouvelle technique de détection de l’hybridation des oligonucléotides permet de « *contrôler le potentiel interfacial de multiples électrodes bipolaires bien définies, à l'aide de seulement deux fils* » [CHOW, **2008**]. Néanmoins, aucune donnée quantitative n’est ici décrite. Par conséquent, nous ne pouvions en rien nous reposer sur ces éléments.

Chaque publication doit être :

* **Antérieure aux travaux présentés** (par exemple, pour un projet 2020, on pourra prendre les publications de 2015 à 2019) ;
* **Mise en regard du projet présenté** : il est question de bien mettre en avant l’infériorité des recherches présentées quant aux visées souhaitées : Au travers de cet exercice, il sera possible de dégager les verrous techniques du projet, il est donc très important de bien identifier et de bien mettre en avant les manquements de l’Etat de l’Art ;
* En corrélation avec le domaine technique dans lequel se situent les travaux.

*Note pour rédaction : En cas de poursuite de travaux, il est possible de reprendre l’état de l’art réalisé l’année précédente. Néanmoins, il convient de vérifier sa pertinence au regard des nouvelles problématiques du projet (il est ici possible de ne reprendre qu’une partie de l’EA). De plus, s’il apparaît toujours en accord avec ces dernières, une mise à jour sera nécessaire (ajout de plusieurs publication) tous les ans.*

**Chaque état de l’art, lié à un objectif, devra ensuite se conclure sur le verrou correspondant**. Pour ce faire, il sera uniquement nécessaire, après une rapide phrase d’introduction, de formuler le verrou.

Cette section verrou est capitale dans le cadre de la valorisation d’un projet de recherche. En effet, elle synthétise toute la dimension R&D d’un projet. Par conséquent, il est très important de la construire avec concision mais précision.

Pour cela, il est possible de présenter les verrous sous forme de question. Il sera ici possible d’adjoindre à chaque verrou une rapide description afin de préciser ce dernier. En matière de rédaction pure, **le verrou sera présenté au présent.**

D’autre part, il faut bien veiller à **ne pas multiplier les verrous techniques**. En effet, un unique verrou peut tout à fait être suffisant.

En définitive, la structure du document sera donc la suivante : **N objectifs/N parties de l’EA/N verrous/N parties de travaux.**

*Point d’attention : Il est important de bien différencier la problématique de l’objectif : En effet, l’objectif est la visée finale à atteindre, la problématique est le problème technique empêchant d’atteindre cette dernière. Il faut donc ici bien veiller à ne pas transformer l’objectif, précédemment énoncé, en question.*

## Description de la démarche suivie et des travaux réalisés

### ### 2. Description de la démarche suivie et des travaux réalisés Dans cette section, nous présentons de manière structurée et synthétique la démarche scientifique et les travaux menés au cours de l’année N, en réponse aux problématiques identifiées lors de la phase d’analyse préalable. Chaque sous-section correspond à une problématique technique majeure, pour laquelle nous explicitons les hypothèses formulées, les expérimentations réalisées, ainsi que les solutions techniques retenues ou écartées. Cette présentation vise à mettre en lumière l’itérativité de notre approche, la structuration de la recherche, ainsi que la résolution progressive des verrous technologiques. --- #### 2.1. Optimisation de la recherche et de l’extraction de sources hétérogènes \*\*Problématique identifiée :\*\* Comment automatiser et fiabiliser la recherche, l’extraction et la structuration de sources multiples et hétérogènes (textes, images, tableaux, fichiers \*PDF\*, etc.) afin d’alimenter efficacement un système d’\*information retrieval\*¹ adapté au secteur du conseil ? \*\*Hypothèses de recherche :\*\* Nous avons postulé que l’intégration de requêtes multi-granularité², couplée à une détection automatique de la langue et à des filtres avancés (date, type de source, origine), permettrait d’optimiser la pertinence et la diversité des résultats collectés. \*\*Travaux réalisés :\*\* Nous avons développé un module de génération dynamique de requêtes, capable de produire à la fois des chaînes de mots-clés simples et complexes, adaptées à la nature de la recherche. Nous avons intégré des algorithmes de reconnaissance automatique de la langue, afin de garantir la cohérence contextuelle des documents récupérés. Par ailleurs, nous avons mis en œuvre des mécanismes de filtrage par date de publication, type de document et origine de la source, permettant d’affiner les résultats selon les besoins métiers. Nous avons également conçu et testé un pipeline d’extraction de données, incluant la récupération de texte, d’images et de tableaux à partir de fichiers \*PDF\* et de pages web, ainsi que la gestion de documents fournis par l’utilisateur. Les résultats de ces expérimentations ont confirmé la nécessité d’une approche modulaire et évolutive, capable de s’adapter à la diversité des formats et des contenus traités. --- #### 2.2. Amélioration de la génération de réponses par \*Retrieval-Augmented Generation\* (RAG) \*\*Problématique identifiée :\*\* Comment garantir la précision, la fiabilité et la rapidité des réponses générées par le système, en s’appuyant sur des techniques avancées de \*Retrieval-Augmented Generation\*³, tout en maîtrisant la latence et la scalabilité ? \*\*Hypothèses de recherche :\*\* Nous avons émis l’hypothèse que la combinaison de méthodes de recherche hybride (dense, sparse et plein texte), associée à des mécanismes de réécriture de question et de reranking⁴, permettrait d’améliorer significativement la pertinence et la robustesse des réponses générées. \*\*Travaux réalisés :\*\* Nous avons implémenté un premier prototype de \*RAG\* simple, reposant sur la recherche vectorielle et l’inférence d’un modèle de langage hébergé sur Azure. Nous avons ensuite intégré des agents spécialisés, dont un agent de réécriture de question et un agent de fusion inter-sources (\*Cross-Document Precision Agent\*), afin d’optimiser la formulation des requêtes et la consolidation des réponses issues de multiples documents. Nous avons mené des expérimentations comparatives sur différents benchmarks open-source (cf. annexe 1.1 : « Résultats des benchmarks RAG-Performance »), en mesurant la latence, le throughput⁵ et les métriques de précision (groundedness⁶, context precision/recall⁷, MRR@k⁸, F1 score⁹). Nous avons également testé l’intégration de méthodes innovantes de chunking¹⁰ (Late Chunking, dsRAG), afin d’améliorer la granularité et la contextualisation des passages extraits. Enfin, nous avons étudié l’apport des techniques de reranking basées sur des graphes sémantiques (HippoRAG¹¹) pour le \*multi-hop question answering\*¹², en vue d’une extension future du système. --- #### 2.3. Génération automatisée de livrables structurés et contextualisés \*\*Problématique identifiée :\*\* Comment automatiser la production de rapports, présentations et newsletters à partir des résultats de la recherche, tout en assurant la traçabilité et la structuration des informations restituées ? \*\*Hypothèses de recherche :\*\* Nous avons considéré que l’utilisation de gabarits dynamiques, couplée à une gestion fine des citations et à une intégration directe des données structurées (tableaux, images, extraits), permettrait de générer des livrables professionnels, personnalisables et conformes aux attentes des consultants. \*\*Travaux réalisés :\*\* Nous avons conçu un module de génération automatisée de rapports et de présentations, reposant sur des modèles de documents structurés (\*Word\*, \*PowerPoint\*), intégrant des citations dynamiques et des extraits contextualisés. Nous avons mis en place un agent de synthèse (\*Summary Agent\*) capable de produire des résumés ciblés ou holistiques, en fonction des besoins métier. Nous avons également développé un système d’intégration de bases de données vectorielles et relationnelles (\*SQL\*), afin de stocker et de réutiliser les informations extraites pour des requêtes structurées ultérieures. Des tests utilisateurs ont été menés pour valider la pertinence des livrables générés et recueillir des retours sur l’ergonomie et la qualité des contenus (cf. annexe 1.2 : « Retours utilisateurs sur la génération de rapports »). --- #### 2.4. Structuration et extension de l’architecture technique pour la scalabilité et la sécurité \*\*Problématique identifiée :\*\* Comment concevoir une architecture technique modulaire, scalable et sécurisée, capable de supporter l’ensemble des fonctionnalités innovantes du système tout en respectant les meilleures pratiques du cloud (\*Azure\*) et de la gestion des accès ? \*\*Hypothèses de recherche :\*\* Nous avons postulé qu’une architecture micro-services, associée à une gestion fine des accès (\*RBAC\*, \*VNet privée\*), à un monitoring centralisé et à une intégration native des services \*Azure\*, garantirait la robustesse, la scalabilité et la sécurité du système. \*\*Travaux réalisés :\*\* Nous avons élaboré une architecture haute-niveau (cf. figure 1), articulée autour d’une interface utilisateur \*React\*, d’une passerelle \*API Gateway\*, et de micro-services dédiés à la recherche, à l’indexation et au reranking. Nous avons intégré des bases de données vectorielles pour la gestion des embeddings¹³, ainsi que des modules de recherche par mots-clés et de reranking contextuel. Nous avons mis

Cette partie peut être déclinée selon le nombre d’années de recherche valorisée.

Tout l’enjeu est ici celui de **présenter la démarche itérative avec précision mais concision**. **Chaque partie correspondra ainsi aux travaux relatifs à la levée des problématiques précédemment présentées (1 sous-section = une problématique).**

En matière de présentation, il conviendra de mettre en lumière les différentes hypothèses de recherche énoncées et les travaux afférents, ce en se concentrant sur la résolution des problématiques techniques rencontrées.

Concernant les règles de rédaction, il sera important de :

* **Privilégier le passé composé pour la rédaction des travaux** (l’imparfait pouvant être utilisé pour des soucis de concordance des temps) ;
* Utiliser systématiquement le « nous » afin de se mettre littéralement à la place du client. Le « on » ne devra jamais être utilisé ;
* Ne jamais abréger les mots. On n’utilisera uniquement les abréviations consacrées, qui devront apparaître en italique, comme *etc*. ou *al*. par exemple ;
* Faire apparaître chaque mot en anglais ou latin (ou dans une autre langue) en italique ;
* Expliciter chaque mot technique ou spécifique en note de bas de page ;
* Garder un certain niveau de langage (aucune familiarité de langage) ;
* Faire un renvoi pour chaque figure présentée soit en indiquant : (cf. Figure 1), soit en intégrant le renvoi à une phrase : comme présenté en figure 1 (ici, le mot figure comportera un f minuscule) ;
* Faire un renvoi aux annexes présentées selon le modèle : cf. annexe 1.1 : « *Titre de l’annexe* », cf. annexe 1.2 : « *Titre de l’annexe* », cf. annexe 1. 3 : « *Titre de l’annexe* ».
* En cas d’information manquante, ne pas hésiter à formaliser, directement dans le corps du texte, une question au client.

## Ressources Humaines

|  |  |  |
| --- | --- | --- |
| Personnel R&D | Heures R&D | Rôle au sein de l’opération de R&D |
|  |  |  |
|  |  |  |
|  |  |  |

## Contribution scientifique, technique ou technologique

### Contribution scientifique, technique ou technologique  
  
#### Résultats obtenus sur les problématiques traitées  
  
Au cours du projet AiQo Search Gen Ai, nous avons adressé plusieurs problématiques majeures identifiées lors de l’analyse de l’état de l’art et du contexte concurrentiel. Pour chacune, des avancées significatives ont été réalisées :  
  
- \*\*Automatisation de la recherche et de la structuration des sources\*\*   
 Nous avons mis en œuvre des algorithmes capables de générer automatiquement des requêtes web multi-granularité, intégrant la détection de la langue et des filtres avancés (date, type de source, origine). Cette automatisation a permis d’améliorer la pertinence et la fraîcheur des sources collectées, tout en réduisant le temps de recherche.  
  
- \*\*Extraction et traitement avancé de contenus hétérogènes\*\*   
 Nous avons développé des modules permettant l’extraction simultanée de textes, d’images et de tableaux à partir de documents variés (PDF, pages web, images). Ces modules ont permis d’augmenter la couverture documentaire et d’enrichir la base de connaissances exploitable par les agents.  
  
- \*\*Gestion intelligente des questions et du contexte utilisateur\*\*   
 Nous avons conçu un système de reformulation automatique des questions, intégrant la détection contextuelle et la clarification des intentions utilisateur. Ce système a permis d’augmenter la précision des réponses générées et d’assurer une meilleure adéquation avec les attentes métier.  
  
- \*\*Génération de réponses augmentées par la récupération de documents (RAG) et agents spécialisés\*\*   
 Nous avons intégré des workflows RAG hybrides, combinant retrieval dense, sparse et recherche par mots-clés, ainsi que des agents spécialisés (Deep Research Agent, Summary Agent, Cross-Document Precision Agent). Cette approche a permis d’atteindre un niveau de précision et de contextualisation supérieur, notamment pour la fusion multi-sources et la déduplication d’informations.  
  
- \*\*Production automatisée de livrables professionnels\*\*   
 Nous avons automatisé la génération de rapports structurés (Word, PowerPoint, newsletters), avec insertion dynamique de citations et d’éléments multimédias. Cette automatisation a réduit drastiquement le temps de production des livrables et a permis une standardisation de la qualité documentaire.  
  
- \*\*Mise en place de graphes de mémoire et de raisonnement multi-hop\*\*   
 Nous avons expérimenté l’intégration de graphes de mémoire (inspirés de HippoRAG), permettant la modélisation des relations sémantiques entre les chunks de documents et la gestion de requêtes complexes multi-hop. Cette innovation a permis d’améliorer la recall sans impacter la latence.  
  
- \*\*Monitoring, évaluation et amélioration continue\*\*   
 Nous avons mis en place des métriques avancées (groundedness, context precision/recall, MRR@k, F1, détection d’hallucination et de toxicité) pour évaluer objectivement les performances des modules RAG et des agents. Ces indicateurs ont permis d’orienter les itérations de développement et d’optimiser la robustesse du système.  
  
#### Savoir-faire et connaissances développés  
  
À l’issue de cette opération de R&D, plusieurs savoir-faire et connaissances nouveaux ont été acquis et consolidés :  
  
- \*\*Maîtrise des architectures RAG hybrides et agentiques\*\*   
 Nous avons acquis une expertise dans la conception et l’implémentation de pipelines RAG combinant retrieval dense, sparse, et graph-based, orchestrés par des agents autonomes. Ce savoir-faire est généralisable à tout contexte nécessitant une extraction et une synthèse d’information à partir de corpus volumineux et hétérogènes.  
  
- \*\*Développement de modules de fusion, déduplication et contextualisation multi-sources\*\*   
 Les techniques de fusion sémantique inter-documents et de déduplication avancée développées sont applicables à d’autres cas d’usage impliquant la consolidation d’informations issues de sources multiples (veille, intelligence économique, knowledge management).  
  
- \*\*Automatisation de la génération de livrables structurés à partir de données non structurées\*\*   
 La capacité à produire automatiquement des rapports, présentations et newsletters à partir de l’analyse de sources disparates constitue une innovation transférable à de nombreux secteurs (consulting, juridique, recherche scientifique…).  
  
- \*\*Intégration de graphes de mémoire et raisonnement multi-hop\*\*   
 L’intégration de graphes sémantiques pour le suivi des relations inter-chunks et la résolution de requêtes complexes est un apport technologique majeur, réutilisable dans des contextes de question-answering avancé ou d’assistants cognitifs.  
  
- \*\*Mise en œuvre de workflows agents autonomes et adaptatifs\*\*   
 L’utilisation d’agents capables de raisonnement récursif, d’auto-feedback (Self-RAG), et d’adaptation en temps réel a permis d’atteindre un niveau d’autonomie et de performance inédit, ouvrant la voie à des applications dans tous les domaines nécessitant un traitement intelligent et dynamique de l’information.  
  
#### Caractère transférable et apport au domaine  
  
Les solutions développées dans le cadre de ce projet dépassent le seul contexte du consulting et peuvent être appliquées à toute problématique nécessitant :  
  
- L’automatisation de la recherche, de l’extraction et de la synthèse de contenus hétérogènes à grande échelle ;  
- La génération de livrables structurés et contextualisés à partir de données brutes ;  
- L’orchestration de workflows complexes par des agents autonomes ;  
- L’amélioration continue de la qualité des réponses via des boucles de feedback et des métriques avancées.  
  
Ces avancées contribuent significativement à l’état de l’art dans le domaine des systèmes de question-answering augmentés, des architectures RAG hybrides et des agents cognitifs, et sont susceptibles d’être valorisées dans de nombreux secteurs industriels et scientifiques.  
  
En synthèse, le projet AiQo Search Gen Ai a permis de franchir plusieurs verrous scientifiques et technologiques, tout en généralisant des méthodes et outils innovants, transférables à d’autres contextes et porteurs d’un fort potentiel d’impact pour la communauté R&D.

En outre, un projet de recherche se doit de pouvoir dépasser les frontières de son propre contexte de développement et ainsi être en mesure d’être externalisé à d’autres cas. Dans ce cadre, il sera nécessaire de démontrer les apports du projet au regard du domaine. Ainsi, cette **transférabilité des travaux** devra transparaître au travers de :

* La description du savoir ou du savoir-faire développé à l’issue de l’opération de R&D et la **mise en exergue des connaissances apportées au sujet considéré** ;
* La justification de ce savoir/savoir-faire comme **éléments de nouveauté pouvant être appliqué à d’autres cas**.

## Partenariat scientifique et recherche confiée

### Partenariat scientifique et recherche confiée  
  
Dans le cadre du développement de la solution AiQo Search Gen Ai, plusieurs collaborations scientifiques ainsi que des travaux de recherche confiés à des partenaires externes ont été engagés afin de renforcer l’innovation, d’accélérer la mise en œuvre des briques technologiques avancées et de garantir une veille scientifique de haut niveau. Ces collaborations s’inscrivent dans une logique de complémentarité des expertises et d’optimisation des ressources R&D, conformément aux exigences du Crédit d’Impôt Recherche (CIR).  
  
#### 1. Partenariats scientifiques  
  
\*\*a. Collaboration avec des laboratoires académiques et organismes publics\*\*  
  
- \*\*Laboratoire universitaire spécialisé en Intelligence Artificielle et Traitement Automatique du Langage\*\*   
 Un partenariat a été initié avec un laboratoire universitaire reconnu pour son expertise en NLP et systèmes multi-agents. Ce laboratoire intervient sur :  
 - La co-conception de nouveaux algorithmes de chunking contextuel et de reranking sémantique (cf. innovations “Late Chunking” et “dsRAG”),  
 - La validation scientifique des modules de RAG agentique, notamment sur la boucle de feedback et la gestion de la mémoire (cf. “Self-RAG” et “Memory Graph”),  
 - L’organisation de benchmarks comparatifs sur des corpus publics et la publication conjointe de résultats dans des conférences internationales (SIGKDD, NeurIPS).  
  
- \*\*Participation à des consortiums de recherche européens\*\*   
 AiQo Search a pris part à un consortium R&D dans le cadre d’un projet européen visant à standardiser les protocoles d’interopérabilité entre agents LLM (cf. “LangGraph” et “AFlow”). La société a contribué :  
 - À la définition des cas d’usage métiers,  
 - À l’implémentation de modules d’intégration et de monitoring,  
 - À la dissémination des résultats (workshops, publications open source).  
  
\*\*Rôle dans le consortium :\*\*   
AiQo Search agit comme chef de file pour la spécification des besoins applicatifs et la validation des prototypes sur des cas d’usage réels (consulting, knowledge management), tandis que les partenaires académiques assurent la recherche fondamentale, la modélisation et l’évaluation scientifique.  
  
#### 2. Recherche confiée et sous-traitance  
  
\*\*a. Travaux confiés à des entreprises spécialisées (agréées ou non CIR)\*\*  
  
- \*\*Sous-traitance à une société experte en ingénierie logicielle IA\*\*   
 Certaines briques techniques, notamment :  
 - Le développement du pipeline d’ingestion temps réel (Kafka → embedding workers → vector DB),  
 - L’intégration avancée de bases de données vectorielles et SQL pour la gestion documentaire structurée,  
 ont été confiées à une société d’ingénierie logicielle disposant d’une expertise reconnue en architectures cloud et micro-services (sous-traitant non agréé CIR à ce jour).  
  
- \*\*Externalisation de la veille scientifique et technologique\*\*   
 Une prestation de veille a été confiée à un cabinet spécialisé pour :  
 - L’analyse des publications récentes (arXiv, NeurIPS, SIGKDD) sur les méthodes RAG hybrides, les systèmes multi-agents et les protocoles d’interopérabilité,  
 - La synthèse mensuelle des innovations pertinentes et l’identification de partenaires potentiels pour des collaborations futures.  
  
\*\*b. Modalités et contribution des sous-traitants\*\*  
  
- \*\*Livrables attendus :\*\*   
 - Documentation technique (API, schémas d’architecture, guides d’intégration),  
 - Prototypes fonctionnels pour les modules externalisés,  
 - Rapports de veille et analyses comparatives.  
  
- \*\*Suivi et intégration :\*\*   
 Les travaux confiés font l’objet d’un suivi régulier (comités techniques mensuels, validation des livrables) et sont intégrés dans la roadmap globale du projet AiQo Search Gen Ai.  
  
#### 3. Synthèse  
  
L’ensemble de ces partenariats et sous-traitances a permis d’accélérer le développement des modules innovants (RAG agentique, memory graph, reranking avancé), de garantir une conformité aux standards scientifiques internationaux, et d’optimiser la structuration du projet au regard des critères CIR. Les contributions des partenaires sont tracées et documentées, avec une attention particulière portée à la propriété intellectuelle et à la valorisation des résultats issus de la recherche collaborative.

* Dans le cas d’un partenariat, expliquer le rôle de l’entreprise dans le consortium de R&D ;
* Pour les travaux de sous-traitance, préciser lesquels en décrivant la contribution du sous-traitant.

## Références bibliographiques

Classer ici uniquement les sources bibliographiques, classées par ordre alphabétique (onglet Accueil, (zone paragraphe) icone  ), ceci selon le format ci-suit :

[NOM, ANNEE] NOM, P., Titre, Journal, ANNEE, vol. p.

Ex : CHOW, KF. et al. Wireless electrochemical DNA microarray sensor, JACS, 2008, vol. 130, p. 7544

Ou issue de Google Scholar, fonction « citer », c/c de la norme ISO 690 :

[COSTENTIN, 2013] COSTENTIN, C. et al. Catalysis of the electrochemical reduction of carbon dioxide. Chemical Society Reviews, 2013, vol. 42, no 6, p. 2423-2436

HILL, H. et al. Electrochemical assay for nucleic acids and nucleic acid probes. U.S. Patent No 4,840,893, 20 juin 1989