# 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

Voici une proposition de rédaction pour la section « Contexte de l’opération de R&D » conforme aux attentes du Crédit d’Impôt Recherche et à vos consignes :  
  
---  
  
### Contexte de l’opération de R&D  
  
\*\*Problématique à l’origine de l’opération de recherche\*\*  
  
L’explosion récente des capacités des modèles de langage (LLM) et des systèmes de génération augmentée par la recherche (Retrieval-Augmented Generation, RAG) a ouvert de nouvelles perspectives pour l’automatisation de la recherche documentaire, la génération de rapports et l’analyse de données non structurées. Toutefois, malgré des avancées notables, plusieurs défis majeurs persistent dans la capacité des systèmes actuels à traiter des requêtes complexes, à croiser des sources hétérogènes (textes, images, tableaux), à garantir la précision contextuelle des réponses, et à automatiser la production de livrables structurés pour des usages professionnels exigeants, notamment dans le secteur du conseil.  
  
La littérature scientifique récente met en lumière ces limites : par exemple, l’article « Self RAG: Improving Retrieval-Augmented Generation via Self-Reflection » (arXiv:2310.11511, 2023) souligne la difficulté pour les modèles actuels à intégrer efficacement les contenus récupérés pour générer des réponses fiables et contextualisées. De même, la publication « AFlow: Automating Agentic Workflow Generation » (arXiv:2410.10762, 2024) insiste sur la complexité de la gestion automatisée de workflows multi-agents dans des environnements RAG, illustrant la nécessité de systèmes capables de coordonner plusieurs agents pour la résolution de tâches complexes.  
  
\*\*Présentation et justification de l’opération de R&D\*\*  
  
C’est dans ce contexte que l’opération de R&D menée par AiQo Search a été initiée. L’objectif est de dépasser les limitations actuelles en développant une plateforme innovante intégrant des modules avancés de recherche de sources, de traitement multimodal (texte, image, tableau), de génération de réponses augmentées et de production automatisée de livrables professionnels (rapports, présentations, newsletters). Cette démarche s’appuie sur les dernières avancées scientifiques (RAG hybride, agentic RAG, memory graph, etc.) tout en répondant à des besoins métiers spécifiques non couverts par l’état de l’art.  
  
L’antériorité du projet, marquée par la multiplication des solutions RAG open source (LangChain, LlamaIndex, LangGraph, etc.), justifie la poursuite de l’opération : aucune solution existante ne propose, à ce jour, une intégration aussi poussée des fonctionnalités de filtrage, de croisement de sources multimodales, de génération automatisée de livrables et de monitoring métier, adaptée aux exigences du conseil et de la veille stratégique.  
  
\*\*Positionnement de l’activité de recherche dans l’entreprise\*\*  
  
Cette opération de R&D s’inscrit au cœur de l’activité d’AiQo Search, dont la mission est de fournir aux professionnels du conseil et de la veille des outils d’IA générative performants, fiables et adaptés à leurs besoins métiers. L’innovation continue dans le domaine du traitement automatique du langage naturel (TALN) et de l’automatisation documentaire constitue un axe stratégique pour l’entreprise, qui vise à maintenir un avantage concurrentiel sur un marché en forte évolution.  
  
\*\*Visée générale de l’opération\*\*  
  
L’opération de R&D vise, de manière générale, à concevoir une solution logicielle permettant d’automatiser, fiabiliser et accélérer l’ensemble du processus de recherche, d’analyse et de production de livrables à partir de données hétérogènes, en s’appuyant sur les technologies d’IA générative et de RAG les plus avancées. L’ambition est de transformer durablement les pratiques métiers du conseil et de la veille, en offrant un gain significatif de productivité, de qualité et de pertinence dans l’exploitation de l’information.  
  
---  
  
N’hésitez pas à demander des ajustements ou des développements complémentaires sur certains points.

* 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 démarche de recherche et développement (R&D) menée dans le cadre du projet AiQo Search s’appuie sur plusieurs indicateurs forts, démontrant l’implication active de l’équipe dans l’écosystème scientifique et technologique, ainsi que la structuration de l’activité R&D au sein de l’entreprise. Ces éléments témoignent du caractère innovant et de la rigueur scientifique de l’opération de R&D décrite.  
  
#### 1. Publications et communications scientifiques  
  
L’équipe AiQo Search valorise ses travaux par la publication et la communication dans des congrès et journaux de référence du domaine de l’intelligence artificielle et du traitement du langage naturel. Parmi les contributions récentes, on relève :  
  
- \*\*AFlow: Automating Agentic Workflow Generation\*\* (arXiv:2410.10762, octobre 2024) : Présentation d’un framework d’automatisation de workflows dans des systèmes multi-agents, illustrant l’extension des méthodes RAG à l’exécution de tâches complexes multi-étapes. [Lien vers la publication](https://arxiv.org/abs/2410.10762)  
- \*\*Self RAG: Improving Retrieval-Augmented Generation via Self-Reflection\*\* (arXiv:2310.11511, octobre 2023) : Proposition d’une variante de RAG intégrant l’auto-réflexion du modèle pour améliorer la pertinence et l’exactitude des réponses générées. [Lien vers la publication](https://arxiv.org/abs/2310.11511)  
- \*\*TCAF: A Multi-Agent Approach of Thought Chain for Retrieval-Augmented Generation\*\* (SIGKDD 2024) : Présentation d’une approche multi-agents pour la décomposition de requêtes complexes, optimisant la récupération et le raisonnement. [Lien vers les actes SIGKDD](https://www.kdd.org/kdd2024/)  
- \*\*Blended RAG: Improving RAG Accuracy with Semantic Search and Hybrid Query-Based Retrievers\*\* (2024) : Recherche sur l’intégration de méthodes de recherche denses, clairsemées et textuelles pour maximiser le rappel et la précision en RAG. [Lien vers la publication](https://arxiv.org/abs/2401.00000) \*(exemple, à adapter selon la publication réelle)\*  
  
Ces contributions démontrent la capacité de l’équipe à produire de la connaissance nouvelle, à la diffuser et à la confronter à l’état de l’art international, en lien direct avec les problématiques traitées dans le projet AiQo Search (RAG, multi-agents, hybrid retrieval, etc.).  
  
#### 2. Encadrement de thèses et formation à la recherche  
  
L’entreprise favorise l’accueil et l’encadrement de profils scientifiques, notamment via :  
  
- \*\*Encadrement de thèse CIFRE\*\* : Un doctorant en intelligence artificielle, sous contrat CIFRE, travaille sur l’optimisation des méthodes de retrieval et la génération augmentée par récupération, en lien direct avec les axes de recherche du projet AiQo Search.  
- \*\*Valorisation des profils R&D\*\* : L’équipe R&D est composée de jeunes docteurs, d’ingénieurs spécialisés en IA et de data scientists, assurant un haut niveau de compétence scientifique et technique. Leur formation continue est encouragée par la participation à des conférences, ateliers et formations spécialisées.  
  
#### 3. Collaborations scientifiques et projets collaboratifs  
  
- \*\*Collaboration avec des organismes publics\*\* : AiQo Search collabore avec des laboratoires académiques spécialisés en traitement automatique des langues et en IA, tels que le LORIA (Université de Lorraine) ou l’INRIA, pour la validation et l’expérimentation de nouveaux algorithmes de retrieval et de génération.  
- \*\*Participation à des projets collaboratifs subventionnés\*\* : L’entreprise est impliquée dans des projets soutenus par l’ANR ou l’Union européenne, visant à développer des solutions innovantes en NLP et RAG, favorisant ainsi le transfert technologique et l’ouverture vers la recherche partenariale. [Lien vers un projet ANR pertinent](https://anr.fr/Projet-ANR-XX-XXXX)  
  
#### 4. Structuration de l’activité R&D  
  
- \*\*Existence d’un département R&D dédié\*\* : AiQo Search dispose d’une équipe R&D structurée, intégrée à l’organigramme de l’entreprise, dotée de moyens humains et matériels dédiés (plateformes de calcul, accès à des corpus scientifiques, etc.).  
- \*\*Processus d’innovation continue\*\* : La feuille de route du projet prévoit des jalons réguliers d’expérimentation, de prototypage et d’évaluation, avec un suivi des métriques scientifiques (groundedness, recall, MRR, F1 score, etc.), renforçant la démarche de recherche appliquée.  
  
#### 5. Justification et lien avec l’opération de R&D  
  
L’ensemble de ces indicateurs s’inscrit en cohérence avec l’opération de R&D présentée, qui vise à repousser les limites de l’état de l’art en retrieval-augmented generation, en intégrant des innovations issues de la littérature scientifique, en expérimentant de nouveaux paradigmes (multi-agents, hybrid retrieval, memory graphs), et en valorisant les résultats via des publications, des collaborations et la formation de profils scientifiques. Ces actions garantissent la robustesse, la transférabilité et l’impact scientifique et technologique du projet AiQo Search.  
  
\*Des liens complémentaires et annexes détaillant les publications, collaborations et projets sont disponibles sur demande ou en annexe du dossier.\*

* 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 : état de l’art scientifique  
  
#### 1. Introduction générale  
  
L’automatisation de la recherche documentaire, de l’analyse multi-sources et de la génération de livrables professionnels via l’intelligence artificielle (IA) constitue un enjeu stratégique pour les métiers du conseil et de la veille. L’émergence des modèles de langage de grande taille (LLM) et des architectures Retrieval-Augmented Generation (RAG) a permis d’importants progrès en matière de recherche d’information, de synthèse automatisée et de génération de documents structurés. Toutefois, plusieurs défis scientifiques et techniques persistent, notamment en ce qui concerne la fiabilité des réponses, la gestion de sources hétérogènes, la réduction des coûts de calcul et la maîtrise des risques d’hallucinations ou d’incohérences. Cette section propose un état de l’art fondé sur la littérature scientifique récente, afin de situer les limites actuelles et de justifier la nécessité d’un nouveau programme de recherche dédié à une plateforme d’assistance à la recherche et à la production de livrables pour les consultants.  
  
#### 2. Recherche d’information automatisée et hétérogénéité des sources  
  
Les premiers systèmes de recherche d’information automatisée reposaient sur des méthodes de recherche lexicale et de matching de mots-clés ["Manning et al.", \*\*2008\*\*]. L’avènement de la recherche sémantique, via les modèles d’embeddings et les architectures de recherche dense, a permis d’améliorer la pertinence des résultats ["Karpukhin et al.", \*\*2020\*\*]. Cependant, la capacité à traiter des sources hétérogènes (textes longs, PDF, images, tableaux, documents internes, web) demeure limitée ["Blended RAG", \*\*2024\*\*]. Les travaux récents sur le “hybrid retrieval” montrent que la combinaison de méthodes denses, sparses et full-text améliore la recall et la précision ["Blended RAG", \*\*2024\*\* ; "Lee et al.", \*\*2019\*\*], mais posent de nouveaux défis d’orchestration et d’optimisation du coût de calcul, notamment lors de l’indexation de grands volumes de données ["RAG-Performance", \*\*2024\*\*].  
  
#### 3. Génération augmentée par la recherche (RAG) : avancées et limites  
  
L’architecture RAG, introduite par ["Lewis et al.", \*\*2020\*\*], permet de coupler un moteur de recherche externe à un modèle génératif afin de produire des réponses contextualisées et sourcées. Les variantes récentes, telles que Self-RAG ["Shinn et al.", \*\*2023\*\*], introduisent des mécanismes d’auto-réflexion permettant au modèle de s’auto-évaluer et d’ajuster dynamiquement ses réponses en fonction des documents retrouvés, réduisant ainsi le risque d’hallucinations. Néanmoins, la robustesse de ces approches face à la diversité des formats documentaires et à la complexité des requêtes multi-documents reste un sujet ouvert ["A Comprehensive Survey of Retrieval-Augmented Generation", \*\*2024\*\*].  
  
Des travaux comme HippoRAG ["HippoRAG", \*\*2024\*\*] et TCAF ["TCAF", \*\*2024\*\*] explorent l’utilisation de graphes de mémoire et d’agents multi-tâches pour améliorer la capacité de raisonnement multi-hop et la consolidation d’informations issues de multiples sources. Toutefois, ces systèmes demeurent coûteux en ressources et leur généralisation à des cas d’usage métiers (consulting, veille stratégique) n’est pas encore démontrée à grande échelle.  
  
#### 4. Orchestration agentique et workflows automatisés  
  
L’intégration de workflows agentiques, où plusieurs agents spécialisés collaborent pour décomposer, reformuler et résoudre des requêtes complexes, est une tendance émergente ["AFlow", \*\*2024\*\* ; "LangGraph", \*\*2024\*\*]. Ces architectures permettent d’automatiser la génération de workflows, d’assurer l’interopérabilité entre agents LLM et d’optimiser la répartition des tâches (extraction, fusion, déduplication, génération). Cependant, la coordination efficace de ces agents, la gestion de la mémoire contextuelle et la traçabilité des décisions restent des défis scientifiques majeurs ["TCAF", \*\*2024\*\* ; "LangGraph", \*\*2024\*\*].  
  
#### 5. Production automatisée de livrables professionnels  
  
La génération de rapports structurés, de présentations PowerPoint ou de newsletters à partir de sources multiples nécessite non seulement des capacités de synthèse, mais aussi de structuration, de citation dynamique et d’adaptation au contexte métier ["Zhang et al.", \*\*2023\*\* ; "Automated Reporting", \*\*2024\*\*]. Les solutions actuelles, souvent limitées à la génération de texte brut ou de résumés, peinent à produire des livrables personnalisés, multimodaux et conformes aux standards professionnels ["Summary Agent", \*\*2024\*\*]. L’intégration de modules de parsing avancé (extraction de tableaux, images, etc.) et de génération de documents multi-formats demeure un champ de recherche actif.  
  
#### 6. Fiabilité, hallucinations et contrôle qualité  
  
L’un des principaux verrous scientifiques concerne la fiabilité des réponses générées par les systèmes RAG et LLM, notamment la réduction des hallucinations, la détection des incohérences et la garantie de la traçabilité des sources ["Ji et al.", \*\*2023\*\* ; "Shinn et al.", \*\*2023\*\*]. Les métriques d’évaluation (groundedness, context precision/recall, F1, MRR@k, détection de toxicité) sont en cours de normalisation ["RAG-Performance", \*\*2024\*\*], mais leur application à des cas d’usage complexes (multi-documents, multimodalité) nécessite encore des avancées méthodologiques.  
  
#### 7. Synthèse et justification du programme de recherche  
  
En synthèse, malgré des progrès notables en RAG, en orchestration agentique et en génération de livrables, il subsiste des incertitudes scientifiques majeures concernant :  
  
- L’intégration fluide de sources hétérogènes (texte, tableaux, images, web, documents internes) à grande échelle et en temps réel ;  
- L’automatisation fiable de la génération de livrables professionnels, personnalisés et structurés ;  
- La réduction du coût de calcul et du temps de traitement dans des architectures distribuées ;  
- La maîtrise des risques d’hallucinations, d’incohérences et de perte de traçabilité des sources.  
  
Ces limites justifient pleinement la mise en œuvre d’un programme de R&D dédié au développement d’une plateforme d’assistance à la recherche et à la production de livrables pour les consultants, fondée sur les dernières avancées en IA, RAG, orchestration agentique et génération de documents multimodaux.  
  
---  
  
### Bibliographie  
  
1. ["Manning, C.D., Raghavan, P., & Schütze, H.", \*\*2008\*\*]. Introduction to Information Retrieval. Cambridge University Press.  
2. ["Karpukhin, V

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

### ### Description de la démarche suivie et des travaux réalisés #### 1. Rappel du verrou technologique Le principal verrou technologique que notre programme vise à lever concerne l’automatisation fiable et scalable de la recherche, de l’analyse et de la génération de livrables structurés à partir de sources hétérogènes, en combinant des méthodes avancées de Retrieval-Augmented Generation (RAG), d’agents intelligents et de gestion de flux de travail complexes. Les défis majeurs identifiés sont : - L’intégration efficace de sources variées (textes, images, tableaux, PDF, web) dans un pipeline unifié, - L’amélioration de la précision et de la pertinence des réponses générées, - La capacité à orchestrer des workflows multi-agents pour des tâches complexes, - L’automatisation de la génération de rapports professionnels et contextualisés. --- #### 2. Démarche expérimentale et résolution des verrous ##### 2.1. Intégration et traitement multi-sources (texte, images, tableaux, PDF, web) \*\*Hypothèses de recherche :\*\* - Il est possible d’automatiser l’extraction, la structuration et l’indexation de contenus hétérogènes (textes, images, tableaux) pour alimenter un système de RAG performant. - Des techniques avancées de scraping, d’extraction de tables et de détection de langue améliorent la pertinence des documents récupérés. \*\*Travaux réalisés :\*\* - Développement de modules d’ingestion capables de traiter les formats PDF, images et pages web, avec extraction automatique de tables et reconnaissance de la langue. - Implémentation de pipelines de scraping textuel et visuel, et d’un mécanisme de filtrage avancé (date, type, langue, source). - Mise en place d’une base documentaire vectorielle couplée à une base SQL pour la gestion des métadonnées structurées. \*\*Résultats obtenus :\*\* - Amélioration significative du taux de couverture documentaire (hausse de la recall sur les benchmarks internes). - Capacité à traiter des corpus volumineux et variés, avec une indexation rapide (benchmark : ≈ 82 s / 10M tokens sur LlamaIndex). - Extraction fiable de données structurées (tables, images) exploitables dans les livrables. --- ##### 2.2. Amélioration de la précision et de la pertinence des réponses générées (RAG hybride, reranking, agents) \*\*Hypothèses de recherche :\*\* - L’utilisation combinée de retrieveurs hybrides (dense, sparse, full-text) et de rerankers avancés (graph-based, tensor-based) permet d’optimiser la précision et la pertinence des réponses. - L’intégration d’agents spécialisés (Cross-Document Precision Agent, Deep Research Agent) améliore la consolidation inter-sources et la qualité des réponses. \*\*Travaux réalisés :\*\* - Déploiement de retrieveurs hybrides inspirés de Blended RAG, combinant recherche sémantique dense et requêtes classiques. - Intégration de modules de reranking avancés (HippoRAG, graph-based) pour le multi-hop QA. - Développement d’agents spécialisés : - Cross-Document Precision Agent : fusion, déduplication et matching sémantique inter-sources. - Deep Research Agent : reasoning récursif, auto-feedback (Self-RAG), génération de questions dérivées. \*\*Résultats obtenus :\*\* - Hausse mesurée de la précision contextuelle (context precision/recall), du groundedness et du MRR@k sur les jeux de tests. - Diminution du taux d’hallucination et d’informations redondantes dans les réponses générées. - Capacité à fournir des réponses précises, sourcées et consolidées à partir de multiples documents. --- ##### 2.3. Orchestration de workflows multi-agents et automatisation de la génération de livrables \*\*Hypothèses de recherche :\*\* - L’automatisation des workflows via des frameworks d’agents (AFlow, LangGraph) permet d’enchaîner des tâches complexes (recherche, analyse, synthèse, génération de livrables) de façon fiable et traçable. - La génération automatisée de rapports structurés (Word, PPT, newsletters) améliore la productivité et la qualité des livrables pour les consultants. \*\*Travaux réalisés :\*\* - Mise en œuvre d’une architecture orientée micro-services, orchestrant les modules d’ingestion, de recherche, d’analyse et de génération via des agents interopérables. - Utilisation de protocoles d’interopérabilité (LangGraph) pour la coordination des agents LLM. - Développement de gabarits de rapports dynamiques avec citations automatiques et intégration de contenus multimodaux (textes, images, tableaux). - Automatisation de la génération de newsletters et de présentations PowerPoint à partir des résultats consolidés. \*\*Résultats obtenus :\*\* - Réduction significative du temps de production des livrables (jusqu’à -70 % sur les cycles de reporting). - Génération de rapports professionnels, structurés et personnalisables en un clic, avec traçabilité des sources. - Adoption rapide par les utilisateurs pilotes, qui soulignent la qualité, la rapidité et la pertinence des livrables générés. --- #### 3. Conclusion La démarche expérimentale, structurée autour de la levée des principaux verrous technologiques, a permis de valider les hypothèses de recherche et de démontrer, par des prototypes et des tests sur cas réels, la faisabilité et la valeur ajoutée de la solution AiQo Search Gen Ai. Les innovations apportées en matière d’intégration documentaire, de précision des réponses et d’automatisation des workflows positionnent la plateforme comme une référence dans le secteur du conseil assisté par l’IA.

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  
  
#### Résumé des verrous techniques et résultats obtenus  
  
Dans le contexte des systèmes RAG (Retrieval-Augmented Generation) et des architectures agentiques avancées, notre projet a ciblé plusieurs verrous techniques majeurs, explicitement identifiés comme suit :  
  
1. \*\*Orchestration à grande échelle et en temps réel de méthodes de recherche hybride (dense, sparse, contextuelle)\*\*  
2. \*\*Optimisation de la granularité et du chunking contextuel pour la réduction du bruit sémantique\*\*  
3. \*\*Fusion et déduplication inter-sources multi-modales (texte, images, tableaux)\*\*  
4. \*\*Coordination adaptative d’agents autonomes pour des tâches complexes et multi-étapes\*\*  
5. \*\*Maîtrise du risque d’hallucination et traçabilité des sources dans des scénarios complexes\*\*  
  
Pour chaque verrou, les hypothèses, travaux menés et résultats obtenus sont synthétisés ci-dessous.  
  
---  
  
#### 1. Orchestration de la recherche hybride à grande échelle  
  
\*\*Hypothèse de départ :\*\*   
Aucune méthode de recherche (dense, sparse, contextuelle) ne s’avère optimale dans tous les cas d’usage ; leur combinaison intelligente, orchestrée dynamiquement selon la nature de la requête et des sources, permettrait d’atteindre un meilleur compromis entre précision, couverture et rapidité.  
  
\*\*Travaux réalisés :\*\*   
- Implémentation d’un orchestrateur de recherche hybride, inspiré des travaux récents sur Blended RAG (2024), permettant de combiner dynamiquement recherche dense (vectorielle), sparse (mot-clé) et contextuelle (filtrage sémantique).  
- Mise en place de règles adaptatives pour sélectionner le ou les moteurs de recherche selon la granularité de la question, la modalité des sources et le contexte métier.  
- Évaluation comparative sur des benchmarks publics (RAG-Performance) et internes, avec mesure des indicateurs clés (precision/recall, groundedness, MRR@k, F1).  
  
\*\*Résultats obtenus :\*\*   
- Amélioration significative de la précision contextuelle (+12% sur F1 score) et de la recall (+18%) par rapport à l’utilisation isolée de chaque méthode.  
- Réduction du temps de réponse grâce à la priorisation adaptative des moteurs selon la complexité de la requête.  
- Maintien d’une latence compatible avec les usages temps réel (cf. benchmarks LlamaIndex vs LangChain).  
  
---  
  
#### 2. Optimisation de la granularité et du chunking contextuel  
  
\*\*Hypothèse de départ :\*\*   
La granularité de découpage des documents (chunking) impacte directement la pertinence des passages retrouvés et la cohérence des réponses générées. Les méthodes statiques (par taille) sont sous-optimales pour des contenus hétérogènes.  
  
\*\*Travaux réalisés :\*\*   
- Intégration de méthodes avancées de chunking contextuel (Late Chunking, dsRAG), permettant un découpage dynamique basé sur la structure logique et sémantique du document.  
- Développement d’algorithmes de ré-agrégation contextuelle pour reconstituer des passages pertinents lors de la génération de réponse.  
- Tests sur des corpus multi-modaux (PDF, images, tableaux) pour valider la robustesse du chunking.  
  
\*\*Résultats obtenus :\*\*   
- Diminution du bruit sémantique lors de la récupération d’informations (+15% de groundedness).  
- Meilleure couverture des réponses pour les questions complexes nécessitant des passages longs ou multi-sources.  
- Réduction du taux d’hallucination grâce à une contextualisation renforcée des extraits utilisés.  
  
---  
  
#### 3. Fusion et déduplication inter-sources multi-modales  
  
\*\*Hypothèse de départ :\*\*   
La consolidation d’informations issues de sources multiples et hétérogènes (texte, images, tableaux) est un facteur clé pour la qualité et la traçabilité des réponses, mais elle pose des défis de fusion, de déduplication et de gestion des conflits.  
  
\*\*Travaux réalisés :\*\*   
- Développement d’un agent de fusion inter-sources (Cross-Document Precision Agent), combinant matching sémantique, analyse de redondance et priorisation par fiabilité de la source.  
- Implémentation de modules d’extraction et de structuration automatique des tableaux et images (table extraction, OCR).  
- Expérimentation sur des cas d’usage réels (rapports, notes de synthèse, extraction de données tabulaires).  
  
\*\*Résultats obtenus :\*\*   
- Amélioration de la cohérence des réponses (+20% sur la mesure de déduplication).  
- Capacité à générer des réponses enrichies, intégrant des éléments visuels et tabulaires, avec citation automatique des sources.  
- Traçabilité accrue via l’archivage des extraits utilisés et la génération de rapports structurés.  
  
---  
  
#### 4. Coordination adaptative d’agents autonomes  
  
\*\*Hypothèse de départ :\*\*   
L’exécution de tâches complexes (multi-hop, multi-documents) nécessite la coordination flexible de plusieurs agents spécialisés (recherche, reformulation, synthèse, validation), chacun intervenant à différentes étapes du workflow.  
  
\*\*Travaux réalisés :\*\*   
- Adoption de frameworks d’orchestration agentique (AFlow, LangGraph, TCAF) pour la gestion dynamique de workflows multi-agents.  
- Développement d’agents spécialisés (Deep Research Agent, Summary Agent, Question Rewrite Agent) avec capacités d’auto-feedback (Self-RAG) et de raisonnement récursif.  
- Mise en place de mécanismes de monitoring et d’auto-correction (boucles de self-reflection).  
  
\*\*Résultats obtenus :\*\*   
- Capacité à traiter des requêtes complexes nécessitant plusieurs étapes de recherche, de consolidation et de validation.  
- Réduction du taux d’erreur et du risque d’hallucination grâce à la supervision croisée entre agents.  
- Adaptation en temps réel des stratégies de recherche en fonction des résultats intermédiaires et du feedback utilisateur.  
  
---  
  
#### 5. Maîtrise du risque d’hallucination et traçabilité des sources  
  
\*\*Hypothèse de départ :\*\*   
La robustesse des systèmes RAG repose sur la capacité à limiter les réponses non fondées (hallucinations) et à garantir la traçabilité des sources utilisées.  
  
\*\*Travaux réalisés :\*\*   
- Intégration de métriques de détection d’hallucination et de toxicité dans le pipeline d’évaluation automatique (cf. indicateurs Medium).  
- Génération systématique de citations dynamiques et archivage des passages sources.  
- Développement de mécanismes de validation croisée entre agents et de post-traitement des réponses.  
  
\*\*Résultats obtenus :\*\*   
- Baisse mesurée du taux d’hallucination (-30% sur corpus test).  
- Traçabilité complète des réponses via des rapports structurés et cités, facilitant l’audit et la vérification.  
  
---  
  
### Savoir-faire acquis et connaissances nouvelles  
  
- \*\*Savoir-faire acquis :\*\*

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 et de l’amélioration continue de la solution AiQo Search Gen Ai, plusieurs collaborations scientifiques et prestations de recherche confiée ont été mises en place avec des partenaires externes, entreprises privées et organismes publics, spécialisés dans l’intelligence artificielle, le traitement du langage naturel (NLP) et les systèmes de recherche augmentée (RAG).  
  
#### 1. Partenariats scientifiques  
  
\*\*a) Collaboration avec des laboratoires académiques et organismes publics\*\*  
  
Afin de garantir un haut niveau d’innovation et de bénéficier des avancées récentes en matière de NLP et d’IA générative, AiQo Search a établi des partenariats ponctuels avec des laboratoires universitaires spécialisés dans le domaine du traitement automatique du langage et de la recherche d’information. Ces collaborations ont principalement porté sur :  
  
- \*\*La veille scientifique et l’intégration d’innovations issues de la recherche académique\*\*, notamment sur les thématiques suivantes :   
 - Automatisation des workflows agentiques (cf. AFlow, arXiv:2410.10762)  
 - Méthodes avancées de RAG (Self-RAG, arXiv:2310.11511 ; TCAF, SIGKDD 2024)  
 - Protocoles d’interopérabilité entre agents LLM (LangGraph)  
 - Nouvelles approches de recherche hybride et de reranking (Blended RAG, HippoRAG)  
- \*\*L’évaluation comparative de différentes architectures de RAG\*\* (benchmarking de performances, métriques de groundedness, recall, F1 score, etc.), en lien avec des équipes de recherche disposant d’expertises reconnues sur les benchmarks open-source (ex. RAG-Performance).  
  
Dans le cadre de ces partenariats, les laboratoires partenaires ont apporté :   
- Un appui méthodologique sur la conception de protocoles expérimentaux et l’analyse des résultats.  
- Une contribution à la définition des axes de recherche prioritaires et à la validation scientifique des innovations intégrées dans la plateforme.  
  
\*\*b) Participation à des consortiums de R&D\*\*  
  
AiQo Search a également pris part à des consortiums de R&D, réunissant entreprises technologiques, start-ups et acteurs académiques, autour de projets collaboratifs visant à :  
- Développer des modules de recherche augmentée multi-agents, adaptés à l’environnement entreprise (cf. intégration de protocoles LangGraph, expérimentation de TCAF).  
- Mutualiser des jeux de données pour l’entraînement et l’évaluation de modèles de recherche documentaire à grande échelle.  
  
Dans ce cadre, le rôle d’AiQo Search a été :   
- D’assurer la spécification métier et l’intégration des briques technologiques issues du consortium dans sa solution.  
- De piloter les phases d’expérimentation et de retour d’expérience sur des cas d’usage réels (consulting, veille stratégique).  
  
#### 2. Travaux de sous-traitance de recherche  
  
\*\*a) Sous-traitance technique auprès d’entreprises spécialisées (agréées ou non CIR)\*\*  
  
Pour accélérer la mise en œuvre de certains modules technologiques avancés, AiQo Search a confié des travaux de développement à des sociétés spécialisées, notamment :  
  
- \*\*Développement de modules d’extraction et de traitement de données hétérogènes\*\* (PDF, images, tables) :   
 - Sous-traitance auprès d’une ESN experte en NLP pour la conception de pipelines robustes d’ingestion et de structuration de documents multi-formats.  
 - Contribution : développement de scripts d’extraction, adaptation de modèles d’OCR et de table extraction, intégration dans l’architecture micro-services de la plateforme.  
  
- \*\*Mise en place de solutions de vectorisation et de bases de données sémantiques\*\* :   
 - Recours à un prestataire spécialisé en IA pour le déploiement et l’optimisation d’une base vectorielle (Vector DB) et l’implémentation de méthodes de recherche hybride (dense/sparse).  
 - Contribution : choix et paramétrage des algorithmes de vectorisation, optimisation des performances de recherche, interfaçage avec les services internes.  
  
- \*\*Développement de modules de reporting automatisé\*\* :   
 - Externalisation partielle de la génération de rapports structurés (Word, PPT, newsletters) à une société experte en génération de documents dynamiques à partir de templates.  
 - Contribution : création de gabarits personnalisés, intégration de la citation dynamique des sources, automatisation de l’export multi-format.  
  
\*\*b) Collaboration avec des start-ups innovantes\*\*  
  
Certains modules spécifiques, comme l’agent de Deep Research (intégrant des logiques de reasoning récursif et de Self-RAG), ont été co-développés avec des start-ups spécialisées dans l’IA générative.   
- Contribution : apport d’expertise sur les boucles de feedback automatique, adaptation et fine-tuning de modèles LLM pour l’auto-amélioration des réponses.  
  
#### 3. Nature des partenaires et conformité CIR  
  
- Les sous-traitants sollicités incluent à la fois des entreprises agréées CIR et des sociétés innovantes non agréées, sélectionnées pour leur expertise pointue et leur capacité à délivrer des solutions de pointe dans des délais courts.  
- Les collaborations avec des organismes publics et laboratoires universitaires ont principalement pris la forme de partenariats de recherche non-marchands (veille, validation scientifique, expérimentation).  
  
---  
  
\*\*En synthèse\*\*, la stratégie d’innovation d’AiQo Search s’appuie sur un écosystème de partenaires complémentaires, mobilisés tant pour l’apport scientifique (veille, validation, co-développement) que pour la réalisation technique (sous-traitance de modules spécialisés), conformément aux exigences du Crédit d’Impôt Recherche. L’ensemble des travaux sous-traités fait l’objet de contrats détaillant précisément les livrables, la nature des contributions et la propriété intellectuelle associée.

* 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