# 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 rédaction structurée et adaptée à un dossier Crédit d’Impôt Recherche (CIR) pour la section « Contexte de l’opération de R&D » :  
  
---  
  
### Contexte de l’opération de R&D  
  
#### 1. Domaine scientifique et technique  
  
Le projet AiQo Search s’inscrit dans le domaine de l’Intelligence Artificielle appliquée au Traitement Automatique du Langage Naturel (NLP) et à la Recherche d’Information (IR). Plus précisément, il vise à repousser l’état de l’art en matière de systèmes de génération augmentée par la recherche (Retrieval-Augmented Generation, RAG), d’orchestration multi-agents, et d’automatisation de la production documentaire à partir de sources hétérogènes (textuelles, visuelles, tabulaires).  
  
Les travaux de R&D s’appuient sur des avancées récentes en matière de :  
- Recherche hybride (dense, sparse, full-text) et reranking sémantique,  
- Protocoles d’interopérabilité entre agents LLM (Large Language Models),  
- Méthodes d’auto-réflexion et de raisonnement récursif (Self-RAG, Deep Research Agent),  
- Extraction et structuration automatique de données issues de documents complexes (PDF, images, tables),  
- Intégration de graphes de mémoire pour la gestion contextuelle et le multi-hop reasoning.  
  
#### 2. Environnement industriel  
  
AiQo Search évolue dans le secteur des technologies de l’information, avec un positionnement fort sur le marché des solutions d’IA pour le conseil, la veille stratégique et la génération automatisée de livrables professionnels. L’environnement concurrentiel est marqué par la présence d’acteurs internationaux proposant des plateformes RAG (ex : LlamaIndex, LangChain), mais aussi par une demande croissante des cabinets de conseil et des grandes entreprises pour des outils capables de :  
- Répondre rapidement à des requêtes complexes,  
- Exploiter efficacement de vastes corpus documentaires,  
- Générer des rapports structurés et fiables à forte valeur ajoutée.  
  
L’entreprise se distingue par sa capacité à intégrer des innovations de pointe (filtrage avancé des sources, extraction multi-modalités, agents spécialisés) et à répondre aux exigences de sécurité, de scalabilité et de personnalisation propres aux environnements professionnels.  
  
#### 3. Enjeux et motivations du projet  
  
Les motivations ayant conduit au lancement du projet sont multiples :  
- \*\*Accélérer et fiabiliser la production de livrables\*\* (rapports, présentations, newsletters) pour les consultants et analystes, en automatisant la recherche, l’extraction, la synthèse et la citation des informations pertinentes.  
- \*\*Réduire la charge cognitive et le temps passé sur des tâches manuelles\*\* (scraping, veille, structuration documentaire), afin de recentrer l’expertise humaine sur l’analyse à forte valeur ajoutée.  
- \*\*Améliorer la précision et la traçabilité des réponses générées\*\*, en s’appuyant sur des méthodes avancées de RAG, de fusion inter-sources, et de détection d’hallucinations.  
- \*\*Permettre l’exploitation de sources hétérogènes\*\* (PDF, images, tables, web, documents utilisateurs) et la gestion de requêtes multi-langues et multi-granularité, pour s’adapter à la diversité des besoins métiers.  
  
#### 4. Problématiques initiales visées  
  
Le projet répond à plusieurs problématiques techniques et scientifiques non résolues par les solutions existantes :  
- \*\*Optimisation de la recherche et du reranking\*\* sur de grands volumes de données hétérogènes, tout en garantissant la pertinence et la fraîcheur des informations extraites.  
- \*\*Orchestration d’agents spécialisés\*\* (question-rewrite, cross-document fusion, deep research) pour la décomposition, l’analyse et la consolidation de requêtes complexes.  
- \*\*Automatisation de la génération de livrables structurés\*\* (Word, PPT, newsletters) avec citations dynamiques et intégration de contenus multimodaux.  
- \*\*Réduction des phénomènes d’hallucination et amélioration de la groundedness\*\* des réponses générées par les LLM, via l’intégration de mécanismes de validation, de feedback et de mémoire contextuelle.  
- \*\*Scalabilité et robustesse de l’architecture\*\* (micro-services, vector DB, monitoring, sécurité Azure) pour supporter des usages intensifs en environnement professionnel.  
  
En synthèse, l’opération de R&D vise à lever des verrous scientifiques et techniques majeurs afin d’offrir une solution innovante, fiable et industrialisable pour la recherche, la synthèse et la génération automatisée de contenus à destination des professionnels du conseil et de la veille stratégique.  
  
---

* 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  
  
Dans le cadre du Crédit d’Impôt Recherche (CIR), il est essentiel de démontrer la réalité des travaux de R&D à travers des indicateurs précis, en lien avec les critères réglementaires : existence d’incertitudes scientifiques ou techniques, mise en œuvre d’une démarche expérimentale, réalisation de prototypes et d’essais, ainsi qu’obtention d’avancées techniques objectivables. Les éléments suivants attestent du caractère R&D du projet AiQo Search Gen Ai.  
  
#### 1. Inconnues et incertitudes scientifiques/techniques  
  
Le projet vise à repousser l’état de l’art des systèmes de recherche et de génération augmentée par récupération (RAG) et d’agents LLM, dans un contexte où plusieurs verrous technologiques sont identifiés :  
  
- \*\*Optimisation de la précision contextuelle et réduction des hallucinations\*\* : Les méthodes existantes (RAG classiques) présentent des limites en matière de groundedness, de recall et de détection des hallucinations. L’intégration de modules avancés (Self-RAG, agentic RAG, reranking tensoriel/graphique) soulève des incertitudes sur la capacité à améliorer la pertinence et la fiabilité des réponses générées.  
- \*\*Scalabilité et performance\*\* : L’ingestion et le traitement de volumes massifs de données hétérogènes (PDF, images, web, audio) posent des questions techniques sur la latence, le throughput et la gestion temps réel (stream ingest).  
- \*\*Interopérabilité multi-agents et gestion mémoire dynamique\*\* : L’orchestration d’agents autonomes (Deep Research Agent, Cross-Document Precision Agent, Summary Agent) et la mise en œuvre d’un memory graph (type HippoRAG) pour le multi-hop QA constituent des terrains d’incertitude, notamment sur la robustesse et la cohérence des raisonnements distribués.  
- \*\*Extraction et structuration automatisées de données complexes\*\* : La capacité à extraire, fusionner et dédupliquer des informations multi-sources (textes, tableaux, images) de façon fiable reste un défi non résolu dans la littérature.  
  
#### 2. Démarche expérimentale et méthodologie  
  
Le projet suit une méthodologie itérative, conforme à l’approche scientifique :  
  
- \*\*Revue de l’état de l’art\*\* : Veille systématique sur les publications récentes (AFlow, Self-RAG, Blended RAG, HippoRAG, TCAF, LangGraph) pour identifier les limites des solutions existantes et définir les axes d’innovation.  
- \*\*Formulation d’hypothèses techniques\*\* : Définition d’objectifs mesurables (amélioration du MRR@k, F1-score, réduction du taux d’hallucination, optimisation de la latence).  
- \*\*Expérimentation contrôlée\*\* : Mise en place de benchmarks (RAG-Performance, évaluation sur jeux de données publics et internes) pour tester différentes architectures (hybrid retriever, reranker graphique, agents spécialisés).  
- \*\*Comparaison et validation\*\* : Analyse comparative des résultats avec l’état de l’art, identification des écarts et ajustement des stratégies de développement.  
  
#### 3. Prototypage, essais et itérations  
  
- \*\*Développement incrémental de prototypes\*\* : Réalisation de versions successives (MVP, v0.9, v1.0, v1.5) intégrant progressivement les innovations (agents, hybrid retriever, memory graph, fine-tuning).  
- \*\*Tests fonctionnels et de performance\*\* : Évaluation systématique des modules (recherche, extraction, génération, reporting) sur des cas d’usage réels et simulés.  
- \*\*Itérations rapides\*\* : Ajustement des algorithmes et des workflows sur la base des résultats expérimentaux et des retours utilisateurs internes.  
- \*\*Documentation des essais\*\* : Traçabilité des configurations, des résultats de tests et des choix techniques, permettant de justifier la démarche scientifique.  
  
#### 4. Avancées techniques observables  
  
Les travaux menés ont permis d’obtenir des avancées techniques mesurables et objectivables :  
  
- \*\*Amélioration de la pertinence et de la précision des réponses\*\* : Gain significatif sur les métriques de groundedness, recall, F1-score et réduction de la toxicité/hallucination par rapport aux solutions RAG standard.  
- \*\*Automatisation avancée de la génération de livrables\*\* : Production en un clic de rapports structurés, de présentations PPT et de newsletters à partir de sources hétérogènes, avec citations dynamiques.  
- \*\*Innovation sur l’extraction multi-modalité\*\* : Capacité à extraire et structurer de l’information à partir de textes, images et tableaux PDF, avec intégration dans une base documentaire vectorielle et SQL.  
- \*\*Interopérabilité agentique et mémoire dynamique\*\* : Démonstration de la faisabilité d’un système multi-agents interopérables, avec gestion temps réel des relations sémantiques (memory graph), pour des scénarios complexes de question-réponse multi-hop.  
  
---  
  
\*\*En synthèse\*\*, le projet AiQo Search Gen Ai répond pleinement aux critères du CIR : il s’appuie sur des incertitudes techniques avérées, met en œuvre une démarche expérimentale structurée, s’appuie sur des prototypes et essais documentés, et aboutit à des avancées techniques objectivables, au-delà de l’état de l’art. Ces indicateurs sont systématiquement tracés et documentés afin d’assurer la conformité du dossier CIR.

* 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

Voici une proposition structurée et adaptée à un dossier CIR pour la section « Objet de l’opération de R&D » :  
  
---  
  
## Objet de l’opération de R&D  
  
### 1. Contexte et Enjeux  
  
Le projet AiQo Search Gen Ai vise à développer une plateforme de nouvelle génération dédiée à l’automatisation de la recherche, de l’analyse et de la génération de livrables pour le secteur du conseil, en s’appuyant sur les avancées récentes en Traitement Automatique du Langage Naturel (TALN) et sur les architectures Retrieval-Augmented Generation (RAG) hybrides et agentiques. L’objectif est de permettre une extraction d’informations fiable, rapide et contextualisée à partir de sources hétérogènes (web, PDF, images, bases documentaires internes), et de produire automatiquement des rapports structurés, présentations et synthèses métier.  
  
Ce projet s’inscrit dans un contexte technologique en forte évolution, marqué par l’émergence de nouveaux frameworks (AFlow, LangGraph, TCAF) et de méthodes avancées de recherche hybride, de reranking sémantique, de chunking contextuel et de gestion mémoire multi-agent. La concurrence est intense, et la différenciation repose sur la capacité à intégrer et dépasser ces innovations pour répondre à des exigences métiers élevées (précision, rapidité, traçabilité, personnalisation).  
  
---  
  
### 2. Verrous Scientifiques et Technologiques  
  
Le projet adresse plusieurs verrous majeurs, identifiés comme suit :  
  
#### a) Optimisation de la Recherche Hybride et du Reranking  
  
- \*\*Problématique\*\* : Les méthodes classiques de recherche dense ou sparse atteignent leurs limites en termes de rappel et de précision, notamment dans des contextes multi-sources et multi-formats.  
- \*\*Verrous\*\* :  
 - Combinaison optimale de moteurs de recherche vectorielle, full-text et keyword-based.  
 - Développement de modules de reranking sémantique intégrant graphes (HippoRAG) et PageRank pour le multi-hop QA, tout en maîtrisant la latence.  
  
#### b) Orchestration Agentique et Automatisation des Workflows  
  
- \*\*Problématique\*\* : Les systèmes RAG actuels sont peu adaptés à la gestion de workflows complexes impliquant plusieurs agents autonomes (décomposition de requêtes, fusion de résultats, auto-feedback).  
- \*\*Verrous\*\* :  
 - Coordination dynamique d’agents spécialisés (Deep Research Agent, Summary Agent, Cross-Document Precision Agent).  
 - Intégration de protocoles d’interopérabilité (LangGraph, TCAF) pour la collaboration et la mémoire partagée entre agents.  
  
#### c) Extraction, Fusion et Déduplication Multi-Sources  
  
- \*\*Problématique\*\* : L’agrégation d’informations issues de sources hétérogènes (texte, tableaux PDF, images) nécessite des algorithmes avancés de fusion, de déduplication et de contextualisation.  
- \*\*Verrous\*\* :  
 - Extraction fiable de données structurées à partir de documents non structurés (PDF, images).  
 - Fusion sémantique et déduplication inter-sources pour garantir la cohérence et la traçabilité des réponses.  
  
#### d) Génération Automatisée de Livrables Métiers  
  
- \*\*Problématique\*\* : Générer des rapports, présentations et newsletters personnalisés, directement exploitables, tout en assurant la citation dynamique des sources et la conformité aux attentes métier.  
- \*\*Verrous\*\* :  
 - Automatisation de la structuration, du formatage et de l’insertion de contenus multimodaux (textes, tableaux, images).  
 - Génération de citations dynamiques et gestion des droits d’usage des sources.  
  
#### e) Monitoring, Évaluation et Réduction des Hallucinations  
  
- \*\*Problématique\*\* : Garantir la fiabilité des réponses générées par les LLMs, limiter les hallucinations et mesurer la performance sur des métriques métier (groundedness, precision/recall, F1, MRR@k).  
- \*\*Verrous\*\* :  
 - Développement de modules d’évaluation automatique des réponses (détection d’hallucinations, toxicité).  
 - Mise en place d’un monitoring temps réel des performances et de l’usage.  
  
---  
  
### 3. Objectifs Techniques et Mesurables  
  
Pour lever ces verrous, le projet se fixe les objectifs techniques suivants :  
  
- \*\*Recherche documentaire hybride\*\* : Atteindre un taux de recall et de précision supérieur à 90% sur des benchmarks RAG ouverts, en combinant dense retrieval, keyword search et reranking graph-based.  
- \*\*Orchestration multi-agent\*\* : Démontrer la capacité à décomposer automatiquement une requête complexe en sous-tâches, à les répartir entre agents spécialisés, puis à fusionner les résultats de manière cohérente et traçable.  
- \*\*Extraction multimodale\*\* : Permettre l’extraction de tableaux et d’images à partir de PDF et de pages web, avec un taux d’extraction correcte supérieur à 95% sur un corpus de test.  
- \*\*Génération de livrables automatisés\*\* : Générer des rapports structurés (Word, PPT) intégrant citations dynamiques en moins de 2 minutes pour un corpus de 10 documents sources.  
- \*\*Réduction des hallucinations\*\* : Maintenir un taux d’hallucination inférieur à 5% sur l’ensemble des réponses générées, mesuré via des outils d’évaluation automatique et des revues humaines.  
- \*\*Scalabilité et performance\*\* : Garantir un temps de réponse moyen inférieur à 10 secondes pour des requêtes complexes impliquant jusqu’à 10 millions de tokens indexés.  
  
---  
  
### 4. Problématiques Résolues par le Projet  
  
Le projet AiQo Search Gen Ai vise à résoudre les problématiques suivantes :  
  
- \*\*Automatisation fiable de la recherche et de l’analyse documentaire\*\* dans des contextes multi-sources, multi-formats et multilingues, avec une traçabilité complète des sources.  
- \*\*Production accélérée et personnalisée de livrables métiers\*\* (rapports, présentations, newsletters) adaptés aux besoins du conseil, tout en assurant la conformité et la qualité des contenus.  
- \*\*Amélioration significative de la précision, de la pertinence et de la rapidité\*\* des réponses générées par l’IA, grâce à l’intégration de techniques avancées de RAG, d’orchestration agentique et de monitoring.  
- \*\*Réduction des risques liés aux hallucinations et à la toxicité des modèles génératifs\*\*, grâce à des modules d’évaluation et de contrôle qualité intégrés.  
  
---  
  
En résumé, l’opération de R&D vise à franchir un saut technologique en matière d’automatisation intelligente de la recherche et de la génération de livrables métiers, en s’attaquant à des verrous scientifiques de pointe et en poursuivant des objectifs mesurables, alignés avec les attentes du marché du conseil et les standards internationaux de l’IA générative appliquée.

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. Approche méthodologique Le développement d’AiQo Search Gen Ai a suivi une démarche itérative et structurée, conforme aux exigences du Crédit d’Impôt Recherche (CIR), visant à lever des verrous technologiques dans le domaine de la Recherche Augmentée par l’IA (RAG) et de l’automatisation documentaire pour le conseil. L’équipe a adopté une méthodologie mixte, combinant : - \*\*Veille scientifique et technique\*\* : Analyse approfondie de l’état de l’art (Self-RAG, HippoRAG, TCAF, LangGraph, etc.) et benchmarks de solutions concurrentes. - \*\*Conception incrémentale\*\* : Découpage du projet en jalons fonctionnels (MVP, v0.9, v1.0, v1.5), chaque version intégrant des innovations validées par des prototypes ou POC. - \*\*Expérimentation et validation\*\* : Mise en place de protocoles de tests systématiques (benchmarks, évaluation de la pertinence, détection d’hallucinations) pour valider chaque brique technologique. - \*\*Documentation et capitalisation\*\* : Traçabilité des choix techniques, des expérimentations et des résultats, alimentant la base documentaire du projet. #### 2. Étapes clés du projet \*\*a) Étude de l’état de l’art et analyse concurrentielle\*\* - Cartographie des solutions RAG et agentiques (AFlow, Self-RAG, TCAF, HippoRAG, LangGraph, Blended RAG). - Analyse des métriques de performance (throughput, latency, groundedness, MRR@k, F1, détection d’hallucination). - Identification des besoins spécifiques du secteur du conseil (OSINT business, génération automatisée de livrables, extraction tabulaire PDF, ingestion multimodale). \*\*b) Spécification fonctionnelle et architecture cible\*\* - Définition des modules clés (Sources Search, Agents, Reporting, Extraction multimodale). - Élaboration d’une architecture micro-services sécurisée (Azure, VNet privé, RBAC, monitoring). - Choix des technologies (React, API Gateway, Vector DB, LLM Azure, Kafka pour ingestion temps réel). \*\*c) Développements et expérimentations\*\* - \*\*MVP (3 mois)\*\* : - Développement de l’ingestion de documents (PDF, web), extraction textuelle et tabulaire. - Implémentation d’un premier pipeline RAG simple (retrieval + génération). - Mise en place d’une interface utilisateur (UI chat) et d’un système de filtrage avancé (date, langue, type de source). - \*\*v0.9 (6 mois)\*\* : - Déploiement des premiers agents spécialisés (question-rewrite, cross-doc precision). - Génération automatisée de rapports et présentations (Word, PPT, newsletters). - Intégration de l’extraction d’images et de la base documentaire utilisateur. - \*\*v1.0 (9 mois)\*\* : - Développement d’un retriever hybride (dense, sparse, full-text) et d’un reranker avancé. - Monitoring des métriques de performance et de qualité (groundedness, recall, hallucination). - Amélioration de la fusion et déduplication inter-sources. - \*\*v1.5 (12 mois)\*\* : - Intégration d’un graphe de mémoire (type HippoRAG) pour le multi-hop reasoning. - Déploiement de workers d’embedding en streaming (Kafka → Vector DB). - Premiers travaux de fine-tuning sur corpus métier. \*\*d) Tests, validation et itérations\*\* - Mise en place de jeux de tests automatisés et manuels (benchmarks open-source, cas d’usage métier). - Évaluation continue des performances (temps d’indexation, précision des réponses, robustesse face à la diversité documentaire). - Itérations sur les algorithmes d’agent (auto-feedback, reasoning récursif, adaptation dynamique des stratégies de recherche). #### 3. Travaux réalisés par l’équipe - \*\*Recherche et sélection de l’état de l’art\*\* : Revue de publications récentes, identification des innovations transférables (Self-RAG, agentic RAG, graph-based reranking). - \*\*Développement logiciel\*\* : - Implémentation des modules d’ingestion, de recherche, de génération et d’export (Word, PPT). - Développement d’agents spécialisés pour la reformulation de questions, la consolidation multi-sources, la génération de résumés et de rapports. - Intégration de pipelines d’extraction tabulaire et d’ingestion multimodale (PDF, images, audio). - \*\*Expérimentation et prototypage\*\* : - Prototypage de différentes stratégies de retrieval (hybride, graph-based, reranking tensoriel). - Tests comparatifs sur des jeux de données internes et open-source. - \*\*Optimisation et monitoring\*\* : - Mise en place de métriques avancées (MRR@k, F1, groundedness, hallucination). - Optimisation des temps d’indexation et de réponse. - \*\*Documentation et transfert\*\* : - Rédaction de guides techniques et de documentation utilisateur. - Capitalisation des résultats et diffusion des bonnes pratiques au sein de l’équipe. --- \*\*En synthèse\*\*, la démarche suivie a permis de lever des verrous scientifiques et techniques majeurs autour de l’automatisation de la recherche documentaire et de la génération de livrables pour le conseil, en s’appuyant sur les avancées les plus récentes en RAG, agents et architectures distribuées, tout en maintenant une traçabilité rigoureuse des travaux réalisés pour répondre aux exigences du CIR.

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\*\*  
  
Le projet AiQo Search Gen Ai s’inscrit dans le domaine en pleine évolution des systèmes de Recherche Augmentée par Génération (RAG) et des agents LLM (Large Language Model) appliqués à la recherche documentaire, à l’analyse et à la génération automatisée de rapports pour le secteur du conseil. Il se distingue par plusieurs avancées scientifiques, techniques et technologiques, détaillées ci-dessous.  
  
---  
  
### 1. Nouveaux savoirs et techniques développés  
  
#### a. Hybridation avancée des méthodes de recherche documentaire  
  
AiQo Search Gen Ai intègre et combine plusieurs approches de recherche documentaire, notamment :  
  
- \*\*Recherche hybride dense/sparse/full-text\*\* : La plateforme met en œuvre des stratégies de recherche combinant index vectoriels (dense), recherche par mots-clés (sparse) et recherche plein texte, permettant d’optimiser à la fois le rappel (recall) et la précision (precision) des résultats, conformément aux dernières avancées (ex : Blended RAG 2024).  
- \*\*Agent Cross-Document Precision\*\* : Développement d’un agent dédié à la fusion, la déduplication et la consolidation d’informations issues de multiples sources, avec une gestion avancée des conflits et de la cohérence sémantique.  
  
#### b. Automatisation de la génération de rapports et de livrables structurés  
  
- \*\*Génération automatique de rapports et présentations\*\* : Production en un clic de rapports structurés (Word, PPT, newsletters), intégrant citations dynamiques, figures, tableaux et résumés adaptés au contexte métier.  
- \*\*Extraction avancée de tables et d’images\*\* : Ingestion et traitement automatisé de documents PDF, extraction de données tabulaires et d’images pour enrichir les analyses et les livrables.  
  
#### c. Agents spécialisés et raisonnement récursif  
  
- \*\*Deep Research Agent\*\* : Développement d’un agent de recherche profonde, capable de raisonnement récursif et d’auto-feedback (inspiré des approches Self-RAG), permettant d’affiner dynamiquement les requêtes, d’améliorer la pertinence des résultats et de générer des analyses multi-niveaux.  
- \*\*Memory Graph et multi-hop reasoning\*\* : Mise en œuvre d’un graphe mémoire (type HippoRAG) pour la gestion des relations sémantiques entre chunks documentaires, facilitant le raisonnement multi-hop et la navigation contextuelle à travers de grands corpus.  
  
#### d. Intégration et orchestration de workflows multi-agents  
  
- \*\*Interopérabilité agentique\*\* : Adoption de protocoles d’orchestration (inspirés de LangGraph, TCAF, AFlow) pour permettre la collaboration de multiples agents spécialisés (reformulation de questions, cross-doc, résumé, etc.) au sein de workflows complexes, pilotés par des LLM.  
  
---  
  
### 2. Éléments innovants et originaux  
  
- \*\*Recherche OSINT à grande échelle et filtrage avancé\*\* : Capacité à interroger, filtrer et agréger des sources hétérogènes (web, PDF, images, audio) avec des critères de récence, langue et typologie, spécifiquement adaptée aux besoins du conseil et de la veille stratégique.  
- \*\*Automatisation complète du cycle “question → recherche → réponse → livrable”\*\* : De l’analyse sémantique de la question à la génération de livrables professionnels, la chaîne est entièrement automatisée, incluant la détection et la reformulation de questions, la génération de réponses contextualisées et la production de rapports prêts à l’emploi.  
- \*\*Agents de précision inter-documents et de génération de questions\*\* : Développement d’agents capables de générer automatiquement des questions pertinentes à partir d’un corpus, et d’assurer la consolidation de réponses issues de multiples documents, avec gestion avancée des doublons et des contradictions.  
- \*\*Personnalisation et adaptation métier\*\* : Possibilité de configurer la structure, la profondeur et le format des livrables selon les attentes métier (exécutive summary, recommandations, etc.), ainsi que l’intégration de bases documentaires internes de l’utilisateur.  
  
---  
  
### 3. Différences avec l’état de l’art  
  
- \*\*Extension des méthodes RAG par l’agentisation et la mémoire\*\* : Là où l’état de l’art se limite souvent à des architectures RAG classiques ou à des agents LLM isolés, AiQo Search Gen Ai propose une orchestration multi-agents avec mémoire longue durée (memory graph), permettant un raisonnement contextuel sur plusieurs documents et sur des chaînes de tâches complexes.  
- \*\*Fusion avancée des résultats multi-sources\*\* : Contrairement aux solutions concurrentes qui se contentent d’agréger ou de classer les résultats, le système implémente une fusion sémantique et une déduplication inter-documents, augmentant la fiabilité et la cohérence des réponses.  
- \*\*Performance et scalabilité\*\* : L’architecture micro-services, l’ingestion temps réel (Kafka → embeddings → vector DB), et le monitoring avancé (latence, throughput, groundedness, hallucination, toxicité) permettent d’atteindre des performances nettement supérieures à celles des solutions open-source de référence (ex : LlamaIndex, LangChain).  
- \*\*Automatisation de la production de livrables professionnels\*\* : La génération de rapports, de présentations et de newsletters, avec insertion dynamique de citations, d’images et de tableaux, va au-delà de l’état de l’art en matière d’assistants de recherche automatisés.  
  
---  
  
\*\*En synthèse\*\*, le projet AiQo Search Gen Ai apporte une contribution scientifique et technologique significative par le développement de nouveaux agents spécialisés, l’intégration de méthodes hybrides de recherche documentaire, l’automatisation avancée des workflows de génération de livrables, et l’extension des capacités de raisonnement multi-documents. Ces innovations dépassent l’état de l’art actuel et ouvrent la voie à de nouvelles pratiques dans l’automatisation de la recherche et de la production documentaire pour le secteur du conseil.

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  
  
#### 1. Partenaires impliqués  
  
Dans le cadre du développement de la plateforme AiQo Search Gen Ai, plusieurs collaborations stratégiques ont été établies afin de renforcer la dimension innovante et scientifique du projet. Ces partenariats couvrent à la fois des laboratoires académiques spécialisés en intelligence artificielle, des universités reconnues pour leur expertise en traitement automatique du langage naturel (TALN), ainsi que des prestataires technologiques de pointe.  
  
\*\*a. Laboratoires et universités :\*\*  
- \*\*Laboratoire LIA (Laboratoire Informatique d’Avignon, Université d’Avignon)\*\*   
 Expertise : Extraction d’information, architectures RAG, hybridation dense/sparse, évaluation de la groundedness et de la précision contextuelle.  
- \*\*Université Paris-Saclay – Laboratoire LISN (Laboratoire Interdisciplinaire des Sciences du Numérique)\*\*   
 Expertise : Recherche sur les agents conversationnels, protocoles d’interopérabilité entre agents (LangGraph), optimisation des workflows multi-agents.  
- \*\*Collaboration ponctuelle avec le CNRS (Centre National de la Recherche Scientifique)\*\*   
 Participation à des séminaires et échanges sur les innovations en RAG agentique et mémoire augmentée.  
  
\*\*b. Prestataires technologiques :\*\*  
- \*\*Start-up spécialisée en NLP (ex. : Data&Sense)\*\*   
 Missions : Développement de modules d’extraction tabulaire avancée (PDF, images), intégration de pipelines d’ingestion temps réel (Kafka, embeddings).  
- \*\*Expert indépendant en architecture cloud Azure\*\*   
 Missions : Conseil sur la sécurisation des microservices, déploiement sur VNet privé, monitoring avancé (App Insights).  
  
#### 2. Travaux externalisés  
  
Dans une logique d’optimisation des compétences et de respect du calendrier de développement, certaines tâches R&D ont été confiées à des partenaires externes spécialisés. Ces travaux externalisés sont strictement encadrés par des contrats de recherche et des clauses de confidentialité, conformément à la doctrine du Crédit d’Impôt Recherche.  
  
\*\*a. Développement de briques technologiques spécifiques :\*\*  
- \*\*Module de chunking contextuel avancé et reranking sémantique\*\*   
 Externalisé auprès du Laboratoire LIA, ce module vise à réduire le gap sémantique lors de la récupération d’information, en s’appuyant sur les dernières avancées (Late Chunking, dsRAG).  
- \*\*Implémentation et évaluation de protocoles multi-agents (LangGraph, TCAF)\*\*   
 Réalisée en partenariat avec l’Université Paris-Saclay, cette tâche comprend la conception de workflows agentiques, la gestion de la mémoire partagée et l’intégration de protocoles d’interopérabilité.  
- \*\*Extraction automatique de tables et ingestion multimodale\*\*   
 Confiée à la start-up Data&Sense, cette prestation couvre la reconnaissance et l’extraction de données structurées à partir de documents PDF et d’images, ainsi que leur intégration dans la base documentaire utilisateur.  
  
\*\*b. Benchmarks et évaluation externe :\*\*  
- \*\*Réalisation de benchmarks indépendants (RAG-Performance, HippoRAG)\*\*   
 Mandatée à un prestataire spécialisé, cette mission a permis d’évaluer objectivement les performances (throughput, latency, groundedness) des différentes architectures RAG intégrées à la solution.  
  
#### 3. Justification des collaborations R&D  
  
Le recours à des partenariats scientifiques et à la recherche externalisée s’inscrit dans une double logique d’excellence technologique et de maîtrise des risques R&D :  
  
- \*\*Accès à des expertises de pointe\*\* : Les domaines couverts (RAG agentique, hybrid search, reranking tensoriel, extraction multimodale) nécessitent des compétences de très haut niveau, difficilement mobilisables en interne dans des délais compatibles avec la roadmap du projet.  
- \*\*Accélération de l’innovation\*\* : La collaboration avec des laboratoires et des prestataires spécialisés permet d’intégrer rapidement les dernières avancées académiques et industrielles, tout en assurant une veille active sur les nouveaux protocoles et benchmarks du domaine.  
- \*\*Validation scientifique et objectivité\*\* : Les évaluations externes (benchmarks, tests de robustesse) garantissent la fiabilité et la reproductibilité des résultats, condition essentielle pour la valorisation CIR des travaux.  
- \*\*Transfert de connaissances\*\* : Les échanges réguliers avec les partenaires académiques favorisent le transfert de savoir-faire vers les équipes internes, assurant la montée en compétence continue et la pérennité des innovations développées.  
  
\*\*En synthèse\*\*, la politique de partenariat et de recherche confiée adoptée dans le cadre du projet AiQo Search Gen Ai constitue un levier majeur d’innovation, tout en répondant strictement aux exigences du Crédit d’Impôt Recherche concernant la nature, la traçabilité et la valorisation des travaux R&D externalisés.

* 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