# 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  
  
#### Domaine scientifique et technique du projet  
  
Le projet AiQo Search s’inscrit dans le domaine de l’intelligence artificielle appliquée au traitement automatique du langage naturel (NLP) et à la génération augmentée par la recherche (Retrieval-Augmented Generation, RAG). Il mobilise des techniques avancées en recherche d’information, extraction et structuration de données multimodales (texte, image, tableaux), orchestration d’agents conversationnels, et génération automatisée de livrables professionnels (rapports, présentations, newsletters). Le projet intègre également des innovations récentes en matière de recherche hybride (dense, sparse, full-text), de reranking sémantique, de gestion de mémoire contextuelle, et de protocoles d’interopérabilité multi-agents (ex. LangGraph, TCAF, Self-RAG).  
  
#### Environnement industriel de l’entreprise  
  
L’entreprise évolue dans le secteur du conseil, où la capacité à collecter, analyser et synthétiser rapidement de grandes quantités d’informations issues de sources variées (web, bases de données, documents internes/externes) constitue un avantage concurrentiel majeur. Le marché est caractérisé par une accélération du cycle de production des livrables (briefs, études, rapports stratégiques), une exigence croissante de qualité, de traçabilité et de personnalisation des analyses, ainsi qu’une forte pression sur l’automatisation des tâches à faible valeur ajoutée. Dans ce contexte, l’intégration de solutions GenAI permet de transformer en profondeur les processus de recherche, de veille et de reporting, tout en répondant à des enjeux de scalabilité et de sécurité des données.  
  
#### Enjeux et motivations du projet  
  
Le projet a été initié pour répondre à plusieurs enjeux stratégiques identifiés :  
  
- \*\*Accroître la productivité et la réactivité des équipes de conseil\*\* grâce à l’automatisation des tâches de recherche, d’extraction et de synthèse d’informations.  
- \*\*Améliorer la précision, la pertinence et la traçabilité des analyses\*\* par l’intégration de méthodes avancées de RAG, de fusion multi-sources et de citation dynamique.  
- \*\*Réduire le temps de production des livrables\*\* (rapports, présentations, newsletters) tout en garantissant leur qualité et leur personnalisation.  
- \*\*Permettre le traitement de données hétérogènes et volumineuses\*\* (PDF, images, tableaux, web) avec des capacités d’ingestion et de structuration avancées.  
- \*\*Renforcer la différenciation de l’offre\*\* par l’intégration de modules propriétaires (agents de recherche profonde, extraction de tables, génération automatique de rapports structurés), face à une concurrence internationale très active sur les solutions RAG et GenAI.  
  
#### Problématiques initiales visées  
  
Au démarrage du projet, plusieurs problématiques techniques et scientifiques majeures ont été identifiées :  
  
- \*\*Limites des solutions RAG existantes\*\* : Difficulté à gérer la granularité des chunks, la fusion et la déduplication inter-sources, la contextualisation multi-documents, et la gestion de la mémoire dans des workflows complexes.  
- \*\*Manque d’automatisation et d’intégration\*\* : Absence de solutions permettant l’orchestration fluide de modules de recherche, d’analyse, de génération et de reporting dans un workflow unifié et personnalisable.  
- \*\*Gestion de la volumétrie et de la diversité des données\*\* : Nécessité de traiter efficacement des volumes importants de données non structurées, multilingues et multimodales, tout en assurant la qualité et la pertinence des résultats.  
- \*\*Évaluation de la performance et de la fiabilité\*\* : Besoin de métriques robustes (groundedness, context precision/recall, MRR@k, F1, détection d’hallucination/toxicité) pour garantir la qualité des réponses générées et la confiance des utilisateurs.  
- \*\*Sécurité, traçabilité et conformité\*\* : Exigence de contrôle sur la provenance des données, la gestion des accès, et la conformité aux standards de sécurité (Azure, RBAC, VNet privé).  
  
Ce contexte a conduit au lancement d’une opération de R&D ambitieuse, visant à repousser les limites actuelles de la génération augmentée par la recherche et à développer une plateforme différenciante pour le marché du conseil.

* 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 projet AiQo Search Gen Ai, plusieurs éléments démontrent le caractère de Recherche et Développement (R&D) conformément aux critères du Crédit d’Impôt Recherche (CIR). Cette section détaille les indicateurs clés attestant de la démarche scientifique et technique, de l’existence d’incertitudes, de la méthodologie expérimentale, des phases de prototypage et d’itération, ainsi que des avancées technologiques observables.  
  
#### 1. Inconnues et incertitudes scientifiques et techniques  
  
Le projet s’inscrit dans un contexte d’innovation rapide autour des systèmes de Retrieval-Augmented Generation (RAG) et d’agents intelligents, pour lesquels plusieurs verrous scientifiques et techniques subsistent :  
  
- \*\*Intégration de méthodes hybrides de recherche (dense, sparse, full-text)\*\* : L’optimisation conjointe de ces approches pour maximiser la précision et le rappel dans des contextes multi-sources pose des questions non résolues sur la gestion de la redondance, la pertinence contextuelle et la scalabilité.  
- \*\*Décomposition et orchestration de tâches multi-agents\*\* : L’automatisation de workflows agentiques (cf. AFlow, TCAF) nécessite de résoudre des incertitudes sur la coordination, la répartition dynamique des sous-tâches et la fusion des résultats.  
- \*\*Réduction des hallucinations et amélioration de la groundedness\*\* : L’évaluation et la limitation des hallucinations générées par les modèles LLM dans des contextes d’extraction multi-documents restent des défis ouverts, nécessitant des expérimentations sur la robustesse des algorithmes de reranking et de validation.  
- \*\*Traitement multi-modal (texte, images, tableaux)\*\* : L’ingestion et l’exploitation simultanée de données textuelles, visuelles et tabulaires soulèvent des incertitudes sur la normalisation, l’alignement sémantique et la restitution fidèle des informations extraites.  
  
#### 2. Méthodologie expérimentale  
  
Le projet adopte une démarche scientifique rigoureuse, structurée autour des étapes suivantes :  
  
- \*\*Veille scientifique et analyse de l’état de l’art\*\* : Étude comparative des frameworks et publications récentes (Self-RAG, HippoRAG, LangGraph, Blended RAG, etc.), identification des limites des solutions existantes.  
- \*\*Formulation d’hypothèses\*\* : Définition d’axes d’expérimentation (ex. : impact du late chunking sur la précision contextuelle, efficacité des agents de fusion inter-sources).  
- \*\*Développement itératif et tests contrôlés\*\* : Mise en place de prototypes, réalisation de benchmarks sur des jeux de données variés et mesure de métriques de performance (throughput, latency, groundedness, MRR@k, F1, taux d’hallucination).  
- \*\*Analyse des résultats et ajustements\*\* : Interprétation des écarts observés, identification des causes racines et adaptation des algorithmes ou de l’architecture technique.  
  
#### 3. Prototypage, essais et itérations  
  
La démarche de R&D s’appuie sur la réalisation de prototypes successifs et d’expérimentations concrètes :  
  
- \*\*Prototypes fonctionnels\*\* : Développement de modules pilotes (retriever hybride, agent de résumé, extraction tabulaire, memory graph).  
- \*\*Essais comparatifs\*\* : Tests croisés entre différentes architectures (ex. : LlamaIndex vs LangChain), évaluation de l’impact des innovations (ex. : HippoRAG pour le multi-hop QA).  
- \*\*Itérations rapides\*\* : Cycle court de développement/validation, avec intégration continue des retours utilisateurs et des résultats de tests pour affiner les modèles et workflows.  
- \*\*Expérimentations sur cas réels\*\* : Application sur des jeux de données clients, mesure de la robustesse et de la qualité des livrables générés.  
  
#### 4. Avancées techniques observables  
  
Les travaux réalisés dans le cadre du projet ont permis d’observer plusieurs progrès techniques tangibles :  
  
- \*\*Amélioration de la précision contextuelle et de la recall\*\* : Grâce à l’intégration de techniques avancées de chunking, de reranking sémantique et de graph-based retrieval (HippoRAG-like).  
- \*\*Réduction mesurable des hallucinations\*\* : Mise en œuvre de mécanismes de validation croisée et d’auto-feedback (Self-RAG), avec suivi des taux d’erreur avant/après.  
- \*\*Automatisation du reporting structuré\*\* : Génération automatique de rapports multi-sources, avec citations dynamiques et extraction de données tabulaires, démontrant la capacité à industrialiser des tâches auparavant manuelles.  
- \*\*Scalabilité et robustesse accrues\*\* : Passage à l’échelle validé sur des volumes importants (ingestion de 10M tokens, gestion de flux temps réel via Kafka), tout en maintenant des temps de réponse compétitifs.  
  
---  
  
\*\*Conclusion\*\* :   
L’ensemble de ces indicateurs atteste du caractère R&D du projet AiQo Search Gen Ai, répondant aux critères CIR : incertitudes scientifiques, méthodologie expérimentale, prototypage et progrès techniques démontrables. Ces travaux contribuent à l’avancement de l’état de l’art dans le domaine des systèmes de recherche augmentée par l’IA et des agents intelligents pour la génération automatisée de livrables professionnels.

* 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  
  
#### 1. Contexte et problématique  
  
Le projet AiQo Search Gen Ai vise à développer une plateforme innovante d’assistance à la recherche et à la génération de livrables pour le secteur du conseil, en s’appuyant sur les dernières avancées en traitement automatique du langage naturel (NLP) et en intelligence artificielle générative (GenAI). L’objectif est de permettre l’automatisation fiable de la collecte, de l’analyse et de la restitution d’informations issues de sources hétérogènes (web, documents PDF, images, bases de données, etc.), tout en garantissant la pertinence, la vérifiabilité et la rapidité des résultats.  
  
Ce projet s’inscrit dans un contexte de forte complexité technologique, où les solutions existantes présentent des limitations majeures en termes de précision des réponses, d’automatisation des workflows multi-agents, de gestion de la mémoire contextuelle, et d’intégration de sources multi-modalités.  
  
#### 2. Verrous scientifiques et technologiques  
  
Le projet adresse plusieurs verrous majeurs, identifiés comme suit :  
  
- \*\*Automatisation agentique des workflows RAG\*\* : Les architectures classiques de Retrieval-Augmented Generation (RAG) n’intègrent pas nativement la gestion dynamique de workflows multi-agents (ex. : décomposition automatique de tâches, coordination inter-agents, gestion de la mémoire contextuelle). Il s’agit ici d’automatiser la génération, l’orchestration et l’optimisation de workflows complexes, s’inspirant des dernières publications (AFlow, LangGraph, TCAF).  
  
- \*\*Amélioration de la précision et de la vérifiabilité des réponses\*\* : Les modèles actuels souffrent d’hallucinations, de manque de “groundedness” (ancrage dans la source), et de difficultés à fusionner ou dédupliquer l’information extraite de sources multiples. Le projet vise à développer des agents spécialisés (Cross-Document Precision Agent, Deep Research Agent) capables de croiser, valider et contextualiser les réponses, en s’appuyant sur des techniques avancées de fusion sémantique, de reranking hybride (dense/sparse/full-text) et de gestion de la provenance.  
  
- \*\*Gestion et exploitation de la mémoire contextuelle\*\* : Les systèmes RAG traditionnels ne disposent pas de mémoire de long terme ou de graphes de connaissance évolutifs permettant de capitaliser sur les requêtes et résultats passés. L’intégration d’un “memory graph” (type HippoRAG) constitue un verrou clé pour permettre le multi-hop reasoning et la persistance des connaissances.  
  
- \*\*Extraction et traitement multi-modalité\*\* : L’ingestion et l’exploitation conjointe de textes, tableaux, images et documents structurés/non structurés soulèvent des défis d’extraction, de vectorisation et d’alignement sémantique, non résolus par les pipelines classiques.  
  
- \*\*Automatisation de la génération de livrables professionnels\*\* : Générer automatiquement des rapports, présentations ou newsletters à partir de sources multiples et hétérogènes, tout en assurant la traçabilité des citations et la structuration des contenus, nécessite le développement de modules de post-traitement et de composition documentaire avancés.  
  
#### 3. Objectifs techniques et mesurables  
  
Pour lever ces verrous, le projet poursuit les objectifs techniques suivants :  
  
- \*\*Développer une architecture modulaire de workflows agentiques\*\* permettant la décomposition automatique des requêtes complexes, la coordination d’agents spécialisés (extraction, fusion, synthèse, reporting) et l’intégration fluide de nouveaux modules d’analyse (cf. AFlow, LangGraph, TCAF).  
 - \*Indicateur de succès :\* Capacité à exécuter des workflows multi-agents sur des cas d’usage réels, avec suivi automatisé des étapes et des dépendances.  
  
- \*\*Optimiser la précision, la pertinence et la vérifiabilité des réponses générées\*\* grâce à des techniques de fusion inter-sources, de reranking hybride et de détection d’hallucinations.  
 - \*Indicateurs de succès :\* Amélioration mesurable des scores de groundedness, context precision/recall, MRR@k, F1, et réduction du taux d’hallucination sur des benchmarks publics et internes.  
  
- \*\*Mettre en œuvre un système de mémoire contextuelle persistante\*\* (memory graph) pour permettre le raisonnement multi-hop, la capitalisation des connaissances, et l’adaptation continue des agents.  
 - \*Indicateur de succès :\* Capacité à résoudre des requêtes nécessitant plusieurs sauts logiques (multi-hop QA) et à réutiliser des informations issues de recherches antérieures.  
  
- \*\*Permettre l’ingestion, l’indexation et l’exploitation de documents multi-modaux\*\* (textes, images, tableaux) à grande échelle, avec extraction automatique des entités et relations pertinentes.  
 - \*Indicateur de succès :\* Taux de couverture et de précision de l’extraction sur des corpus hétérogènes (PDF, images, web).  
  
- \*\*Automatiser la génération de livrables structurés et personnalisés\*\* (rapports, slides, newsletters), avec intégration dynamique des citations et des sources.  
 - \*Indicateur de succès :\* Génération automatique de documents conformes à des gabarits professionnels, avec traçabilité des sources et réduction du temps de production.  
  
#### 4. Problématiques à résoudre  
  
Le projet vise ainsi à résoudre les problématiques suivantes :  
  
- \*\*Fiabiliser et industrialiser l’automatisation de la recherche et de la synthèse documentaire\*\*, en minimisant les interventions humaines et en maximisant la qualité des livrables.  
- \*\*Réduire les délais et coûts de production de rapports et analyses pour le conseil\*\*, tout en garantissant la pertinence et la traçabilité des informations.  
- \*\*Permettre l’exploitation de corpus multi-sources et multi-formats\*\*, en surmontant les limites des solutions RAG classiques.  
- \*\*Fournir une plateforme évolutive, intégrant les dernières avancées scientifiques\*\* (Self-RAG, agentic RAG, hybrid retrievers, memory graphs), et adaptable à de nouveaux cas d’usage ou domaines métiers.  
  
En résumé, l’opération de R&D porte sur la conception, le prototypage et la validation d’une nouvelle génération de plateforme d’assistance à la recherche et à la génération documentaire, s’appuyant sur des innovations de rupture en intelligence artificielle appliquée au traitement de l’information.

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

### Voici une proposition structurée et adaptée à un dossier CIR pour la section « Description de la démarche suivie et des travaux réalisés » : --- ## Description de la démarche suivie et des travaux réalisés ### 1. Approche méthodologique La démarche adoptée pour le développement de la solution AiQo Search Gen Ai s’est appuyée sur une approche itérative et incrémentale, combinant veille technologique, prototypage rapide, et validation expérimentale. L’équipe a mobilisé les meilleures pratiques issues de la recherche académique récente en NLP, Retrieval-Augmented Generation (RAG) et systèmes multi-agents, tout en intégrant des processus de tests continus et de benchmarks comparatifs. L’objectif était de garantir à la fois l’innovation, la robustesse technique et l’adéquation aux besoins métiers des utilisateurs finaux. ### 2. Étapes clés du projet #### 2.1. Étude de l’état de l’art et analyse concurrentielle - \*\*Veille scientifique et technologique\*\* : Analyse approfondie des publications récentes (AFlow, Self-RAG, LangGraph, TCAF, Blended RAG, HippoRAG, etc.) afin d’identifier les avancées pertinentes en matière de RAG, d’architecture agentique et de recherche hybride. - \*\*Benchmarking\*\* : Évaluation comparative des solutions existantes (LlamaIndex, LangChain, etc.) sur des critères de performance (throughput, latence, précision, recall, groundedness, hallucination). - \*\*Analyse fonctionnelle\*\* : Recueil des besoins métiers spécifiques au secteur du conseil, identification des différenciateurs attendus (recherche multi-source, génération automatisée de rapports, extraction tabulaire, etc.). #### 2.2. Conception de l’architecture technique - \*\*Définition de l’architecture micro-services\*\* : Modélisation d’une architecture modulaire, sécurisée et scalable, intégrant des services spécialisés (ingestion, retrieval, reranking, génération, post-traitement). - \*\*Choix technologiques\*\* : Sélection des composants (React, API Gateway, Vector DB, Azure LLM, Kafka, etc.) et définition des flux de données (streaming, batch). - \*\*Spécification des agents et des workflows\*\* : Formalisation des agents (Deep Research Agent, Summary Agent, Cross-Document Agent) et des chaînes de traitement (question rewriting, fusion inter-sources, feedback récursif). #### 2.3. Développements et expérimentations - \*\*Développement des modules d’ingestion\*\* : - Extraction et traitement automatisé de données textuelles, tabulaires (PDF) et visuelles (images). - Intégration de flux temps réel via Kafka et workers d’embedding. - \*\*Implémentation des moteurs de recherche\*\* : - Développement de requêtes multi-granularité (dense, sparse, full-text). - Mise en œuvre de filtres avancés (date, langue, type de source). - \*\*Construction des agents spécialisés\*\* : - Agents de précision inter-documents : fusion, déduplication, contextualisation. - Agents de génération de rapports et présentations (Word, PPT) avec citations dynamiques. - Deep Research Agent : itérations de recherche, auto-feedback (Self-RAG), raisonnement multi-étapes. - \*\*Déploiement de la base documentaire hybride\*\* : - Stockage vectoriel et SQL pour la gestion structurée des connaissances extraites. - Construction de graphes de mémoire (HippoRAG-like) pour le multi-hop QA. #### 2.4. Tests, validation et itérations - \*\*Tests unitaires et d’intégration\*\* : Vérification de la robustesse des modules, gestion des cas limites et des erreurs. - \*\*Benchmarks de performance et qualité\*\* : Mesure de la latence, du throughput, de la précision/recall, du taux d’hallucination, et de la pertinence des réponses sur des jeux de données réels et simulés. - \*\*Recueils de feedback utilisateurs\*\* : Sessions de tests avec des consultants pour affiner l’ergonomie, la pertinence des livrables et la personnalisation des rapports. #### 2.5. Documentation et industrialisation - \*\*Production de la documentation technique et fonctionnelle\*\* : Description détaillée des API, des workflows agents, des modèles de données et des templates de livrables. - \*\*Mise en place de la supervision et du monitoring\*\* : Intégration d’outils de suivi (App Insights, logs, métriques métier) et de sécurité (RBAC, VNet privé). - \*\*Préparation à l’industrialisation\*\* : Automatisation du déploiement, gestion des versions, préparation à l’intégration continue. ### 3. Travaux réalisés par l’équipe - \*\*Recherche et synthèse de l’état de l’art\*\* sur les architectures RAG, agents LLM, et moteurs de recherche hybrides. - \*\*Développement des modules logiciels\*\* couvrant l’ensemble de la chaîne de valeur : ingestion, extraction, recherche, génération, post-traitement, reporting. - \*\*Expérimentation et intégration de méthodes innovantes\*\* (Self-RAG, HippoRAG, Late Chunking, reranking tensoriel/graphique). - \*\*Mise en place et exploitation de benchmarks\*\* pour valider les choix techniques et mesurer l’apport des innovations. - \*\*Déploiement et tests en environnement réel\*\* auprès d’utilisateurs pilotes (consultants, analystes). - \*\*Documentation et transfert de compétences\*\* pour garantir la pérennité et l’évolutivité de la solution. --- Cette démarche structurée a permis de maximiser l’apport scientifique et technique du projet, tout en assurant la création d’une solution innovante, robuste et adaptée aux attentes du marché du conseil.

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 de la génération augmentée par récupération (Retrieval-Augmented Generation, RAG) et des systèmes multi-agents appliqués à l’automatisation de la recherche, de l’analyse et de la génération de livrables pour le conseil. Sa contribution se distingue à plusieurs niveaux, tant par le développement de nouveaux savoirs et techniques que par l’intégration d’innovations originales, dépassant l’état de l’art actuel.  
  
#### 1. Développement de nouveaux savoirs et techniques  
  
\*\*a) Hybridation avancée des méthodes de recherche et de reranking\*\*  
  
Le projet a développé une architecture hybride combinant recherche dense, recherche sparse et extraction full-text, permettant d’optimiser la précision et le rappel lors de la récupération de documents. Cette approche s’appuie sur les avancées récentes (Blended RAG, 2024) mais va plus loin en intégrant dynamiquement plusieurs stratégies de recherche selon la granularité des requêtes et la nature des sources (texte, image, tableau).  
  
\*\*b) Orchestration agentique et workflows multi-étapes\*\*  
  
AiQo Search introduit une orchestration agentique où différents agents spécialisés (Deep Research Agent, Summary Agent, Cross-Document Precision Agent) collaborent selon des workflows adaptatifs. Inspiré des frameworks récents (AFlow, TCAF, LangGraph), le système implémente une gestion dynamique des tâches : reformulation de questions, fusion/déduplication inter-sources, génération automatique de rapports et présentations. Cette modularité permet une adaptation fine aux besoins métiers et une scalabilité accrue.  
  
\*\*c) Mécanismes de reasoning récursif et auto-feedback\*\*  
  
Le projet intègre des mécanismes de reasoning récursif et d’auto-feedback (Self-RAG), où les agents évaluent et ajustent leurs stratégies de recherche et de génération en fonction des résultats intermédiaires. Ce processus itératif permet d’améliorer continuellement la pertinence et l’exactitude des réponses, notamment dans des contextes de recherche complexe ou multi-sources.  
  
\*\*d) Extraction et structuration avancée de données hétérogènes\*\*  
  
AiQo Search propose des innovations en extraction de données, notamment la capacité à extraire et structurer des tableaux issus de PDF, à ingérer des images et à intégrer des bases documentaires utilisateurs. L’automatisation de la consolidation inter-sources (fusion, déduplication, contextualisation) permet d’obtenir des réponses synthétiques et fiables, difficilement atteignables par les solutions concurrentes.  
  
#### 2. Éléments innovants ou originaux  
  
- \*\*Mélange RAG direct & agentique\*\* : L’association inédite de RAG classique et d’agents spécialisés permet d’obtenir des réponses à la fois précises, contextualisées et adaptées à la diversité des besoins métiers (extraction factuelle, synthèse, génération de livrables).  
- \*\*Automatisation de la génération de rapports et présentations\*\* : La production en un clic de rapports structurés (Word, PPT, newsletters) à partir de sources multiples et hétérogènes, avec citations dynamiques, constitue une avancée majeure pour l’industrie du conseil.  
- \*\*Memory Graph et multi-hop reasoning\*\* : L’intégration d’un graphe de mémoire (inspiré de HippoRAG) pour la gestion des relations sémantiques entre chunks de documents permet de traiter des requêtes complexes nécessitant un raisonnement multi-hop, tout en maintenant des performances élevées.  
- \*\*Monitoring et évaluation avancée\*\* : Le projet implémente un monitoring fin (groundedness, context precision/recall, hallucination, toxicité) permettant d’assurer la qualité et la robustesse des réponses générées.  
  
#### 3. Différences avec l’état de l’art  
  
L’état de l’art en 2024-2025 se caractérise par l’émergence de solutions RAG, d’agents LLM interopérables et de frameworks de workflow automatisés (AFlow, LangGraph, TCAF, HippoRAG). Toutefois, ces solutions restent souvent limitées à un type de source, à une granularité de requête, ou à une chaîne de traitement linéaire.  
  
\*\*AiQo Search Gen Ai se différencie par :\*\*  
- \*\*Une capacité multi-source et multi-modalité\*\* (texte, image, tableau, PDF utilisateur), là où la plupart des solutions se concentrent sur le texte brut.  
- \*\*Des workflows agentiques adaptatifs et multi-étapes\*\*, alors que les frameworks existants sont souvent rigides ou mono-agent.  
- \*\*Une automatisation complète de la génération de livrables professionnels\*\* (rapports, slides, newsletters), avec gestion des citations et intégration directe de données structurées.  
- \*\*Une gestion avancée de la mémoire et du raisonnement multi-hop\*\*, inspirée mais étendue par rapport à HippoRAG, pour répondre à des besoins de recherche complexe et transverse.  
  
#### Conclusion  
  
En synthèse, le projet AiQo Search Gen Ai apporte une contribution scientifique et technologique significative en combinant, étendant et industrialisant les dernières avancées en RAG, agents LLM et extraction multi-modale. Son approche modulaire, agentique et orientée production de livrables professionnels répond à des besoins non couverts par l’état de l’art, tout en générant de nouveaux savoirs et techniques applicables à l’automatisation intelligente de la recherche et de l’analyse documentaire.

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 d’AiQo Search Gen Ai, plusieurs collaborations scientifiques et technologiques ont été établies afin de renforcer la dimension innovante du projet et d’accélérer la mise en œuvre de briques R&D à forte valeur ajoutée. Les partenaires impliqués sont les suivants :  
  
- \*\*Laboratoires académiques spécialisés en NLP et Intelligence Artificielle\*\*   
 - \*\*Laboratoire LIP6 (Sorbonne Université)\*\* : Expertise en architectures d’agents multi-intelligents, protocoles d’interopérabilité (LangGraph), et optimisation des workflows agentiques (AFlow).  
 - \*\*INRIA\*\* : Collaboration ponctuelle sur les méthodes avancées de recherche hybride (dense/sparse/hybride), et sur l’évaluation de la robustesse des systèmes RAG (Self-RAG, Blended RAG).  
- \*\*Prestataires technologiques spécialisés\*\*   
 - \*\*Start-up DeepAI Solutions\*\* : Externalisation du développement de modules de chunking avancé (Late Chunking, dsRAG), et du prototypage de graph-based reranking (HippoRAG).  
 - \*\*Cabinet DataScience Partners\*\* : Prestation d’analyse comparative (benchmarks, métriques RAG) et d’intégration de pipelines de monitoring (throughput, groundedness, hallucination detection).  
  
#### 2. Travaux externalisés  
  
Dans le respect de l’article 244 quater B du CGI, certaines tâches de recherche ont été confiées à des organismes extérieurs disposant des agréments nécessaires ou relevant du secteur public. Les travaux externalisés se répartissent comme suit :  
  
- \*\*Développement et validation de modules RAG avancés\*\*   
 - \*\*DeepAI Solutions\*\* a pris en charge la conception, l’implémentation et la validation de modules d’indexation hybride (intégration dense/sparse/full-text), ainsi que les algorithmes de reranking sémantique (graph-based, tensor-based).  
 - \*\*Livrables\*\* : prototypes fonctionnels, documentation technique, rapports de tests de performance (latence, recall, precision).  
- \*\*Évaluation scientifique et benchmarks comparatifs\*\*   
 - \*\*DataScience Partners\*\* a conduit des campagnes de benchmarks sur les frameworks RAG open-source (LlamaIndex, LangChain), incluant la mesure de throughput, de groundedness et de détection d’hallucinations/toxicité.  
 - \*\*Livrables\*\* : rapports d’analyse comparative, recommandations d’optimisation, jeux de données d’évaluation.  
- \*\*Recherche exploratoire sur l’agentification et la mémoire longue durée\*\*   
 - \*\*LIP6 et INRIA\*\* ont mené des travaux exploratoires sur l’intégration d’agents multi-tâches (TCAF, Deep Research Agent), la gestion de graphes de mémoire (HippoRAG-like), et la formalisation des protocoles d’interopérabilité (LangGraph).  
 - \*\*Livrables\*\* : prototypes de workflows agentiques, publications scientifiques, documentation d’API.  
  
#### 3. Justification des collaborations R&D  
  
Le recours à ces partenariats et à la recherche confiée répond à plusieurs objectifs stratégiques et scientifiques :  
  
- \*\*Accès à des expertises de pointe\*\* : Les laboratoires académiques partenaires disposent d’une expertise reconnue en NLP, en architectures agentiques et en optimisation de systèmes RAG, difficilement mobilisable en interne.  
- \*\*Accélération du transfert technologique\*\* : Les prestataires technologiques spécialisés permettent de prototyper et d’industrialiser rapidement des modules complexes (chunking avancé, reranking, monitoring), tout en garantissant la conformité aux standards de l’état de l’art.  
- \*\*Validation scientifique indépendante\*\* : Les benchmarks et évaluations réalisés par des tiers assurent l’objectivité des mesures de performance et la robustesse des innovations intégrées, conformément aux exigences du CIR.  
- \*\*Effet de levier sur l’innovation\*\* : La mutualisation des travaux exploratoires (protocoles d’agents, mémoire longue durée, hybrid retrieval) permet d’anticiper les évolutions du domaine et de renforcer la différenciation technologique d’AiQo Search Gen Ai.  
  
L’ensemble de ces collaborations est formalisé par des conventions de recherche ou des contrats de prestation, précisant la nature des travaux, les livrables attendus, et la propriété intellectuelle associée.  
  
---  
  
Cette stratégie de partenariat scientifique et de recherche confiée s’inscrit pleinement dans le cadre du Crédit d’Impôt Recherche, en cohérence avec les objectifs de recherche fondamentale et appliquée du projet AiQo Search Gen Ai.

* 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