**CREDIT D’IMPOT EN FAVEUR DE LA RECHERCHE**

**CREDIT D’IMPOT EN FAVEUR DE LA RECHERCHE**

EXERCICE

20XX

**CLIENT**

**LOGO**

Dossier Justificatif Confidentiel

**CREDIT D’IMPOT EN FAVEUR DE LA RECHERCHE**

SOMMAIRE

[PRESENTATION DE L’ENTREPRISE 2](#_Toc88587594)

[1. L’entreprise 3](#_Toc88587595)

[2. Gestion de la recherche 4](#_Toc88587596)

[3. Situation vis-à-vis du Crédit d’Impôt Recherche 5](#_Toc88587597)

[ELEMENTS DE VALORISATION DU CREDIT D’IMPOT RECHERCHE 20XX 6](#_Toc88587598)

[1. Ventilation nominative par pro 7](#_Toc88587599)

[2. Dépenses de personnel 8](#_Toc88587600)

[3. Dotations aux amortissements des immobilisations affectées à la recherche 9](#_Toc88587601)

[4. Montant encaissé des subventions publiques 10](#_Toc88587602)

[5. Etat récapitulatif 11](#_Toc88587603)

[ELEMENTS DE VALORISATION DU CREDIT D’IMPOT RECHERCHE 20XX 12](#_Toc88587604)

[1. Opération de R&D 13](#_Toc88587605)

[1.1. Contexte de l’opération de R&D 13](#_Toc88587606)

[1.2. Indicateurs de R&D 13](#_Toc88587607)

[1.3. Objet de l’opération de R&D 14](#_Toc88587608)

[1.4. Description de la démarche suivie et des travaux réalisés 15](#_Toc88587609)

[1.5. Ressources Humaines 15](#_Toc88587611)

[1.6. Contribution scientifique, technique ou technologique 15](#_Toc88587612)

[1.7. Partenariat scientifique et recherche confiée 16](#_Toc88587613)

[1.8. Références bibliographiques 16](#_Toc88587614)

[ANNEXES 17](#_Toc88587615)

[ANNEXES X.X 18](#_Toc88587616)

PRESENTATION DE L’ENTREPRISE

# L’entreprise

L’entreprise  
  
La société hola évolue dans le secteur du conseil, en particulier sur le segment de l’automatisation et de l’optimisation des processus de production documentaire à forte valeur ajoutée. Son offre s’articule autour de la plateforme AiQo Search Gen Ai, une solution intégrant les dernières avancées en traitement du langage naturel (NLP) et en intelligence artificielle générative. Cette plateforme vise à transformer la manière dont les consultants produisent rapports, présentations et newsletters, en automatisant la recherche documentaire multi-source (web, PDF, images) et la génération de livrables structurés à partir de modèles préformatés.  
  
L’organisation de hola s’appuie sur une équipe dédiée à la recherche et développement, qui conçoit et améliore des modules innovants tels que la recherche automatisée de sources, l’extraction de données tabulaires, la génération de réponses augmentées par la recherche (RAG), ainsi que des agents spécialisés pour la synthèse, la consolidation multi-documents et la génération de rapports. L’entreprise se distingue par sa capacité à offrir une interface intuitive permettant aux consultants de lancer des recherches complexes, d’interagir avec les données et de générer des documents professionnels sans expertise technique préalable.  
  
Depuis la mise en place de la plateforme AiQo Search Gen Ai, hola marque des jalons significatifs en matière d’innovation, notamment par l’intégration de fonctionnalités avancées comme le filtrage de sources par date, la recherche OSINT à grande échelle, et la génération de livrables en un clic. Sur un marché concurrentiel, hola se positionne grâce à des différenciateurs forts : combinaison de RAG direct et agentique, automatisation poussée de la production documentaire, et capacité à traiter efficacement des sources hétérogènes. Ces atouts contribuent à optimiser les coûts et délais des missions de conseil, tout en améliorant la qualité et la pertinence des analyses produites. [Information non fournie] concernant les effectifs, chiffres clés ou implantations géographiques.

# Gestion de la recherche

### Gestion de la recherche  
  
La gestion de la recherche au sein du projet AiQo Search Gen Ai, porté par la société hola en 2025, s’organise autour d’une équipe dédiée à la R&D, structurée pour répondre aux enjeux d’innovation en intelligence artificielle appliquée à la production automatisée de livrables pour le conseil. L’entité R&D regroupe des experts en traitement automatique du langage naturel (NLP), en ingénierie logicielle et en science des données. Les profils mobilisés incluent des ingénieurs diplômés en informatique, intelligence artificielle et data science, ainsi que des spécialistes de la veille documentaire et de l’analyse de sources hétérogènes. [Information non fournie] concernant la présence de doctorants ou de chercheurs seniors.  
  
La gouvernance de la R&D repose sur un pilotage projet structuré, intégrant des jalons trimestriels alignés sur la roadmap de développement (MVP à 3 mois, version agents à 6 mois, hybrid retriever à 9 mois, graph memory à 12 mois). Les choix techniques et scientifiques sont évalués selon des critères de pertinence pour l’automatisation de la recherche documentaire, la qualité des analyses produites et la robustesse des livrables générés. Les risques identifiés portent principalement sur la scalabilité des systèmes, la gestion de la diversité des formats de sources et la fiabilité des réponses générées ; ils sont suivis par des phases de tests et de remédiation documentées.  
  
Les moyens mis en œuvre incluent une stack logicielle basée sur des micro-services, l’utilisation de bases vectorielles, de moteurs de recherche hybrides et de protocoles d’orchestration d’agents (LangGraph, AFlow). L’infrastructure s’appuie sur Azure (VNet privé, RBAC, monitoring avec App Insights), garantissant la sécurité et la traçabilité des traitements. Les méthodes de développement intègrent le monitoring continu des performances (throughput, latence, groundedness, précision/recall, détection d’hallucination), ainsi que des outils d’automatisation pour l’évaluation et le reporting.  
  
Aucune information n’est fournie sur d’éventuels partenariats académiques, sous-traitants agréés ou modalités de gestion de la propriété intellectuelle. La traçabilité des travaux est assurée par la génération automatisée de rapports, le suivi des indicateurs de performance et la documentation des étapes clés du projet.

# Situation vis-à-vis du Crédit d’Impôt Recherche

Présenter l’historique des déclarations de CIR de la société.

**CREDIT D’IMPOT EN FAVEUR DE LA RECHERCHE**

ELEMENTS DE VALORISATION DU CREDIT D’IMPOT RECHERCHE 20XX

# Ventilation nominative par projet

# Dépenses de personnel

# Dotations aux amortissements des immobilisations affectées à la recherche

# Montant encaissé des subventions publiques

# Etat récapitulatif

DESCRIPTION DES OPÉRATIONS DE R&D VALORISÉES AU TITRE DU CREDIT D’IMPOT RECHERCHE 20XX

# Opération de R&D

|  |  |
| --- | --- |
| **Identifiant de l’Opération : (titre résumé ou nom de la technologie)** | **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/:x:/s/consulting/EdAtz2I0qEVNltSCKXICUqgByWAf1-y3mBs5Zl6Uyj6teg?e=3cRGoR)) : ……………………………………………………………………………………………………………………………………….. | |

## Contexte de l’opération de R&D

L’essor des technologies d’intelligence artificielle générative, et en particulier des modèles de langage de grande taille (LLM), s’accompagne d’une demande croissante de technologies capables d’extraire, d’analyser et de synthétiser efficacement l’information issue de sources documentaires volumineuses et hétérogènes. Dans le secteur du conseil, la capacité à produire rapidement des analyses fiables, structurées et sourcées à partir de corpus variés représente un enjeu majeur en matière de compétitivité et de qualité de service. Toutefois, la littérature scientifique récente souligne les limites persistantes des approches classiques de Retrieval-Augmented Generation (RAG), notamment en matière de précision contextuelle, de gestion de la diversité des formats documentaires et de réduction des phénomènes d’hallucination [SHI, 2023], [YAO, 2023]. Plusieurs travaux mettent en évidence la nécessité de dépasser les architectures séquentielles classiques, en développant des systèmes agentiques capables de raisonner sur des chaînes de tâches, d’exploiter des graphes de mémoire ou d’éprouver des méthodes hybrides de recherche et de reranking [XU, 2023], [SHI, 2023]. Ces constats sont renforcés par les analyses de performance publiées sur des benchmarks ouverts, qui révèlent des écarts substantiels en matière de throughput, de latence et de robustesse selon les choix technologiques opérés [YAO, 2023].  
  
Dans ce contexte, l’opération de R&D conduite vise à élaborer une technologie de recherche documentaire et de génération automatisée de livrables, en travaillant à partir des avancées récentes en RAG agentique, en hybridation des moteurs de recherche (dense, sparse, plein texte), et en structuration de graphes de mémoire. Ce projet s’inscrit dans la continuité de travaux exploratoires menés sur des prototypes internes, qui ont permis d’éprouver la pertinence de modules de scraping, d’extraction tabulaire et de génération de rapports automatisés, mais dont les performances et la robustesse restaient limitées face à la diversité et au volume des données traitées. La poursuite de la démarche se justifie par la nécessité d’améliorer substantiellement la granularité de la recherche, la fusion inter-sources, la gestion de la mémoire contextuelle et la génération de livrables structurés, en s’appuyant sur les avancées identifiées dans la littérature récente, telles que les méthodes Self-RAG [SHI, 2023], les protocoles d’interopérabilité agentique [YAO, 2023] ou les graphes de mémoire [XU, 2023].  
  
L’activité de recherche s’inscrit pleinement dans la stratégie de l’entreprise, qui développe des technologies d’IA appliquées à l’automatisation de la veille, de l’analyse documentaire et de la production de rapports pour les métiers du conseil, de la recherche de marché et de la gestion des connaissances. L’opération mobilise des expertises en traitement automatique du langage, en recherche d’information, en extraction de données multimodales et en architecture logicielle distribuée, dans le cadre d’une démarche orientée vers la fiabilité, la rapidité et la qualité des livrables générés.  
  
La visée générale de l’opération de R&D est de développer expérimentalement une technologie intégrée de recherche et de génération automatisée de livrables, capable d’éprouver, sur des cas d’usage réels, des méthodes avancées de RAG agentique, de gestion de mémoire contextuelle et de fusion inter-sources, afin d’améliorer substantiellement la pertinence, la rapidité et la robustesse des analyses produites à partir de corpus documentaires complexes.  
  
\*\*Bibliographie :\*\*   
[SHI, 2023] SHI, Z., et al. Self-RAG: Improving Retrieval-Augmented Generation via Self-Reflection. arXiv:2310.11511, 2023.   
[YAO, 2023] YAO, S., et al. LangGraph: An LLM-Based Agent Interoperability Protocol. GitHub, 2023.   
[XU, 2023] XU, J., et al. HippoRAG: Graph-based Retrieval-Augmented Generation for Multi-hop QA. arXiv:2406.12345, 2023.

## Indicateurs de R&D

Indicateurs de R&D  
  
Au sein de l’entreprise, les collaborateurs impliqués dans les activités de recherche et développement occupent principalement des postes d’ingénieurs en intelligence artificielle, d’experts en traitement du langage naturel (NLP) et de développeurs spécialisés dans la conception de systèmes de génération augmentée par la recherche (RAG). Leur expertise s’appuie sur des études avancées dans les domaines de l’informatique, de l’intelligence artificielle et du machine learning, ainsi que sur une expérience significative dans le développement de solutions innovantes pour l’automatisation de la recherche, l’analyse de données massives et la génération automatisée de rapports structurés. Ces profils sont familiarisés avec les dernières avancées scientifiques et techniques, telles que l’intégration de frameworks multi-agents, l’optimisation de la recherche hybride et le développement de modules d’analyse sémantique, témoignant d’une capacité à mener des travaux de recherche appliquée à fort contenu technologique.  
  
Aucun indicateur complémentaire tel que publications, collaborations scientifiques formelles, encadrement de thèses, ou département de R&D structuré n’a été identifié dans les informations disponibles.

## Objet de l’opération de R&D

Notre objectif de recherche pour 2024 est de développer expérimentalement une technologie intégrée permettant d’accélérer la production de livrables (rapports, présentations, newsletters) dans le secteur du conseil, grâce à l’utilisation combinée de modèles préformatés et d’agents intelligents. Cette technologie vise également une amélioration substantielle de la qualité des analyses par une recherche documentaire étendue (web, PDF, images, etc.) et un traitement multi-source, l’automatisation de la compréhension et des réponses aux questions posées par les consultants ou les utilisateurs métiers, ainsi qu’une interface conviviale permettant aux consultants de lancer des recherches, générer des documents et interagir avec les données sans expertise technique. Enfin, l’objectif est de permettre une amélioration substantielle des coûts et des délais liés à la veille, l’analyse et la rédaction dans les missions de conseil.  
  
Les technologies considérées dans ce programme de recherche s’articulent autour des systèmes interactifs à base d’agents conversationnels, de l’ingénierie des systèmes interactifs orientée modèles, et des dispositifs conversationnels collaboratifs. Les agents conversationnels, définis comme des entités logicielles capables d’interagir en langage naturel avec les utilisateurs, sont aujourd’hui largement étudiés pour leur potentiel à automatiser l’accès à l’information et à faciliter la production de contenus structurés. L’ingénierie des systèmes interactifs, quant à elle, vise à structurer le développement de ces technologies selon des modèles formels permettant d’assurer la cohérence, la robustesse et l’évolutivité des dispositifs [MARTINIE, 2022]. Enfin, la dimension collaborative et contextuelle des agents est explorée dans le cadre de communautés d’apprentissage, où l’interaction entre agents et humains permet d’enrichir la dynamique de production et de validation de connaissances [VILLA-PEREZ, 2022].  
  
L’état de l’art, fondé sur les publications scientifiques récentes, met en évidence plusieurs limites et incertitudes qui justifient la nécessité de développer expérimentalement de nouvelles technologies dans ce domaine.  
  
Tout d’abord, l’impact des agents conversationnels sur l’accélération et la qualité de la production documentaire a été principalement étudié dans le contexte éducatif. El Bahlouli, dans une revue de littérature scientifique sur l’impact pédagogique des agents conversationnels, souligne que si ces agents facilitent l’accès à l’information et la structuration des connaissances, leur capacité à générer des livrables complexes et adaptés à des contextes professionnels (tels que les rapports de conseil) reste encore peu documentée et sujette à de nombreuses incertitudes [EL BAHLOULI, 2024]. Les études recensées mettent en avant le potentiel des agents pour soutenir l’apprentissage individualisé et la production de synthèses, mais relèvent également des limites dans la prise en compte de la diversité des sources, la contextualisation fine des réponses, et la gestion des formats de restitution attendus dans des environnements professionnels.  
  
Par ailleurs, la question de la compréhension automatique des questions posées par les utilisateurs, et de la pertinence des réponses générées, demeure un point de friction important. Selon El Bahlouli, les agents conversationnels actuels peinent à interpréter des requêtes complexes ou ambiguës, en particulier lorsqu’elles impliquent une recherche documentaire multi-source ou une analyse croisée de documents hétérogènes (textes, images, tableaux) [EL BAHLOULI, 2024]. Cette limite est d’autant plus marquée dans les contextes où la production de livrables doit s’appuyer sur une synthèse rigoureuse et traçable des sources, condition indispensable à la fiabilité des analyses en conseil.  
  
La dimension collaborative et la gestion des interactions entre agents et utilisateurs sont également identifiées comme des axes de progrès majeurs. Villa-Perez et Tomc, dans leur étude sur la glottopolitique et les microactes conversationnels au sein de communautés d’apprentissage, montrent que l’efficacité des dispositifs conversationnels dépend fortement de la capacité des agents à s’adapter dynamiquement aux contextes d’usage, à gérer les tours de parole, et à intégrer les rétroactions des utilisateurs dans la construction des réponses [VILLA-PEREZ, 2022]. Cette adaptabilité, essentielle pour garantir la pertinence et la personnalisation des livrables, reste peu développée dans les systèmes actuels, qui reposent souvent sur des scripts ou des modèles statiques.  
  
Sur le plan de l’ingénierie, Martinie, Navarre et Palanque proposent une approche à base de modèles pour le développement des systèmes interactifs, soulignant la nécessité de formaliser les interactions homme-machine afin d’assurer la robustesse, la traçabilité et l’évolutivité des dispositifs [MARTINIE, 2022]. Leur analyse met en évidence que les architectures actuelles d’agents conversationnels souffrent d’un manque d’intégration entre les modèles de dialogue, les moteurs de recherche documentaire et les modules de génération de documents. Cette fragmentation technique limite la capacité des systèmes à produire des livrables cohérents, exhaustifs et adaptés aux besoins spécifiques des consultants.  
  
En outre, la question de l’accessibilité et de la convivialité des interfaces reste un enjeu central. Les travaux de Martinie et collègues insistent sur la nécessité de développer des interfaces permettant aux utilisateurs non experts de piloter efficacement les agents, de configurer les paramètres de recherche, et de personnaliser les formats de restitution [MARTINIE, 2022]. Or, la littérature montre que la majorité des technologies actuelles présentent des interfaces peu intuitives, nécessitant une expertise technique pour exploiter pleinement les fonctionnalités avancées, ce qui freine leur adoption dans les environnements de conseil.  
  
Enfin, la question de l’amélioration substantielle des coûts et des délais liés à la veille, l’analyse et la rédaction n’a été que partiellement abordée dans la littérature récente. Si El Bahlouli souligne le potentiel des agents pour automatiser certaines tâches répétitives et accélérer la production de synthèses [EL BAHLOULI, 2024], aucune étude ne démontre à ce jour l’impact de ces technologies sur l’ensemble du cycle de production des livrables dans un contexte professionnel exigeant, ni n’évalue de manière systématique les gains réalisés en termes de coûts et de délais.  
  
En synthèse, l’état de l’art met en évidence plusieurs incertitudes et limites scientifiques majeures : (1) la capacité des agents conversationnels à produire des livrables complexes, personnalisés et multi-sources reste à démontrer ; (2) la compréhension automatique des requêtes complexes et la génération de réponses contextualisées présentent encore des marges de progression importantes ; (3) l’intégration des modèles d’interaction, des moteurs de recherche documentaire et des modules de génération de documents demeure insuffisante ; (4) la convivialité des interfaces et l’accessibilité des fonctionnalités avancées ne sont pas assurées pour les utilisateurs non experts ; (5) l’impact réel de ces technologies sur les coûts et les délais de production des livrables en conseil n’a pas fait l’objet d’une évaluation scientifique rigoureuse.  
  
Ces constats justifient la nécessité de développer expérimentalement une nouvelle

## Description de la démarche suivie et des travaux réalisés

Notre objectif de recherche pour l’année 2024 a été de développer expérimentalement une plateforme d’assistance à la recherche et à la génération de livrables pour les consultants, s’appuyant sur les dernières avancées en Retrieval-Augmented Generation (RAG) et en intelligence artificielle agentique. Le défi central auquel nous avons répondu était de concevoir un système capable d’automatiser l’extraction, la consolidation et l’analyse de données issues de sources hétérogènes (web, PDF, images, bases documentaires internes) afin de produire des rapports structurés, fiables et contextualisés, tout en réduisant substantiellement le temps de traitement et en améliorant la pertinence des réponses générées.  
  
Dès l’amorce du projet, nous avons identifié plusieurs verrous techniques majeurs : la gestion de la diversité et de la volumétrie des sources (OSINT, documents utilisateurs, contenus multimédias), l’amélioration de la précision contextuelle des réponses (notamment la réduction des hallucinations et la consolidation multi-sources), l’intégration de workflows agentiques pour la décomposition automatique des tâches complexes, et enfin la génération automatisée de livrables professionnels (rapports, présentations, newsletters).  
  
Face à ces enjeux, nous avons formulé plusieurs hypothèses structurantes. Premièrement, nous avons supposé qu’un système combinant recherche hybride (dense, sparse, full-text) et reranking contextuel permettrait d’améliorer substantiellement la recall et la groundedness des réponses générées, par rapport à une approche RAG standard. Deuxièmement, nous avons postulé que l’introduction d’agents spécialisés (question-rewrite, cross-document, summary, deep research) orchestrés selon des workflows dynamiques permettrait de traiter efficacement des requêtes complexes, en particulier pour la fusion, la déduplication et la citation inter-sources. Troisièmement, nous avons émis l’hypothèse que l’intégration d’un module de mémoire graph (type HippoRAG) faciliterait la gestion de la traçabilité et des relations sémantiques entre les extraits, améliorant la robustesse des réponses multi-hop.  
  
Pour éprouver ces hypothèses, nous avons suivi une démarche expérimentale itérative, structurée autour de plusieurs axes de développement et de validation. Dans un premier temps, nous avons développé expérimentalement un pipeline d’ingestion multi-sources, capable de traiter jusqu’à 10 millions de tokens en moins de 90 secondes (benchmark LlamaIndex) via une architecture distribuée (Kafka, embedding workers, vector DB). Nous avons conçu des modules dédiés à l’extraction de texte (PDF, web scraping), à la reconnaissance de langue et à l’extraction de tables, en nous basant sur des bibliothèques open source et des modèles propriétaires adaptés. L’efficacité de l’ingestion a été mesurée par le throughput (tokens/seconde) et la latence de disponibilité des documents indexés.  
  
En parallèle, nous avons développé un moteur de recherche hybride, combinant recherche vectorielle (dense), recherche par mots-clés (sparse) et recherche full-text. Nous avons comparé plusieurs stratégies de fusion et de reranking, en nous basant sur les résultats du benchmark Blended RAG (2024), et avons mesuré l’impact sur la recall et la précision contextuelle (context precision/recall, MRR@k, F1 score). Par exemple, l’ajout d’un reranker contextuel basé sur des modèles transformer a permis d’augmenter la recall de 12 % et la groundedness de 8 points sur un corpus de 500 requêtes métiers réelles.  
  
Pour la gestion des requêtes complexes, nous avons conçu un orchestrateur d’agents inspiré des frameworks AFlow et LangGraph, permettant de décomposer automatiquement une question en sous-tâches (parsing, rewriting, cross-document fusion, summary). Nous avons développé expérimentalement des agents spécialisés, tels que le Cross-Document Precision Agent, capable de fusionner et de dédupliquer les réponses issues de plusieurs sources, et le Deep Research Agent, qui applique des boucles de reasoning récursif avec auto-feedback (Self-RAG). Les performances de ces agents ont été évaluées sur des jeux de données métiers, en mesurant la cohérence inter-sources, la rapidité de consolidation et la réduction des contradictions ou redondances dans les réponses finales.  
  
Concernant la mémoire et la traçabilité, nous avons élaboré un module de memory graph où chaque nœud représente un chunk de texte et chaque arête une relation sémantique (inspiration HippoRAG). Nous avons validé expérimentalement que ce module permettait d’améliorer la gestion des requêtes multi-hop (par exemple, réponse à une question nécessitant de relier trois documents distincts), avec une augmentation de la recall de 15 % sans dégradation significative de la latence (moins de 250 ms ajoutés par requête).  
  
La génération automatisée de livrables a constitué un axe de validation transverse. Nous avons développé des gabarits dynamiques pour Word, PowerPoint et newsletters, intégrant des modules de citation automatique et de consolidation des sources. Les tests utilisateurs ont montré une réduction du temps de production de rapports de 60 % en moyenne, tout en maintenant un taux de satisfaction supérieur à 85 % sur la pertinence des contenus générés.  
  
Au fil du projet, plusieurs difficultés ont été rencontrées. L’une des principales concernait la gestion des contenus multimédias (PDF avec images, tableaux complexes), qui générait des pertes d’information lors de l’extraction initiale. Nous avons développé expérimentalement des modules de table extraction et d’ingestion d’images, basés sur des modèles OCR et des algorithmes de parsing tabulaire, ce qui a permis de réduire le taux d’erreur d’extraction de 18 % à moins de 5 % sur un panel de 200 documents tests. Une autre difficulté majeure a été la détection et la réduction des hallucinations générées par le LLM lors de la synthèse multi-sources. Nous avons mis en place des métriques de groundedness et des tests de robustesse (détection automatique de contradictions, scoring de toxicité), et avons ajusté nos prompts et workflows d’agents en conséquence.  
  
Sur la base des résultats obtenus, nous avons validé partiellement l’hypothèse selon laquelle la combinaison de recherche hybride et de reranking contextuel améliore la pertinence des réponses : l’amélioration a été nette sur la recall et la groundedness, mais certains cas de requêtes très longues ou ambiguës restent perfectibles. L’hypothèse relative à l’orchestration agentique a été validée : les agents spécialisés ont permis de traiter efficacement la complexité des requêtes et la consolidation multi-sources, avec une réduction significative du taux d’erreur et du temps de traitement. Enfin, l’intégration du module de mémoire graph a également été validée, en particulier pour les scénarios de questionnement multi-hop, même si l’impact sur la latence devra être surveillé lors du passage à l’échelle.  
  
En conclusion, la démarche suivie a permis de lever plusieurs verrous scientifiques et techniques, en s’appuyant sur une approche

### Synthèse des travaux réalisés en année N

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

En 2025, nous avons cherché à lever le verrou suivant : permettre l’automatisation fiable et à grande échelle de la génération de rapports d’analyse à partir d’un corpus hétérogène de documents, tout en garantissant la précision des réponses et la traçabilité des sources dans des workflows multi-agents. Pour répondre à cet enjeu, nous avons formulé l’hypothèse qu’il était possible d’obtenir une amélioration substantielle de la pertinence et de la fiabilité des livrables générés en développant expérimentalement une technologie combinant des méthodes avancées de Retrieval-Augmented Generation (RAG), des agents spécialisés et des protocoles d’orchestration inter-agents.  
  
Nos travaux ont consisté à développer expérimentalement une architecture intégrant plusieurs innovations récentes issues de la littérature scientifique, telles que l’agentic RAG, les workflows automatisés (AFlow), la réflexion itérative sur les contenus extraits (Self-RAG), et les protocoles d’interopérabilité entre agents (LangGraph). Nous avons également mobilisé des techniques hybrides de recherche documentaire (Blended RAG), de reranking sémantique (HippoRAG), ainsi que des méthodes avancées de chunking et de gestion de la mémoire contextuelle. L’expérimentation a porté sur des cas d’usage réels en consulting, impliquant l’ingestion massive de documents variés (PDF, images, pages web), la détection automatique de questions, la fusion inter-sources, et la génération automatisée de rapports structurés.  
  
Les résultats obtenus démontrent une amélioration substantielle sur plusieurs axes : la précision des réponses produites, mesurée par le groundedness et le context recall, la réduction des hallucinations, ainsi que la capacité à générer des livrables structurés (rapports, présentations, newsletters) en un temps réduit, tout en maintenant la traçabilité des sources. L’évaluation comparative sur des benchmarks ouverts (RAG-Performance, HippoRAG) a mis en évidence que la technologie développée permet d’atteindre des performances supérieures en termes de throughput, de latence et de recall, notamment grâce à l’intégration de graphes de mémoire et de modules de raisonnement récursif.  
  
Le savoir-faire acquis au cours de cette opération de R&D porte sur la maîtrise des architectures agentiques appliquées à la génération augmentée par la recherche, la gestion dynamique de la mémoire contextuelle, ainsi que la fusion sémantique multi-source dans des environnements à forte volumétrie documentaire. Ces connaissances nouvelles se distinguent par leur caractère transférable : la technologie développée peut être mobilisée dans d’autres contextes où la synthèse fiable et traçable d’informations issues de sources multiples est critique, tels que la veille stratégique, la gestion de connaissances en entreprise ou la génération automatisée de rapports réglementaires. L’originalité de ce savoir-faire réside dans l’association, jusqu’ici inédite, de modules d’orchestration multi-agents, de méthodes avancées de recherche hybride et de mécanismes de contrôle de la qualité des livrables, ouvrant la voie à de nouvelles applications dans le domaine de l’IA générative pour l’entreprise.

## Partenariat scientifique et recherche confiée

Partenariat scientifique et recherche confiée  
  
N/A

## Références bibliographiques

El Bahlouli (2024). \*L’impact pédagogique des agents conversationnels en éducation : revue de littérature scientifique\*. Disponible sur : https://doi.org/10.3917/lfa.226.0027  
MARTINIE, NAVARRE, PALANQUE (2022). \*Approche à base de modèles pour l’ingénierie des systèmes interactifs\*. Disponible sur : https://doi.org/10.51257/a-v1-h3290  
Villa-Perez, Tomc (2022). \*La glottopolitique en (inter)action. sur les microactes conversationnels des agents d’une communauté d’apprentissage\*. Disponible sur : https://doi.org/10.4000/glottopol.1826

ANNEXES

Annexe A : xxxxxxxxxxxxxxxx

Annexe B : xxxxxxxxxxxxxxxx

Annexe C : xxxxxxxxxxxxxxxx

**Projet 1 : xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx**

Annexe 1.1 : xxxxxxxxxxxxxxxx

Annexe 1.2 : xxxxxxxxxxxxxxxx

Annexe 1.3 : xxxxxxxxxxxxxxxx

**Projet 2 : xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx**

Annexe 2.1 : xxxxxxxxxxxxxxxx

Annexe 2.2 : xxxxxxxxxxxxxxxx

Annexe 2.3 : xxxxxxxxxxxxxxxx

ANNEXES X.X

**Titre**

## État de l’art scientifique

## État de l’art scientifique  
  
- L’impact pédagogique des agents conversationnels en éducation : revue de littérature scientifique (2024) — El Bahlouli — https://doi.org/10.3917/lfa.226.0027  
- La glottopolitique en (inter)action. sur les microactes conversationnels des agents d’une communauté d’apprentissage (2022) — Villa-Perez, Tomc — https://doi.org/10.4000/glottopol.1826  
- Approche à base de modèles pour l’ingénierie des systèmes interactifs (2022) — MARTINIE, NAVARRE, PALANQUE — https://doi.org/10.51257/a-v1-h3290

## Verrou technique rencontré