# 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

La croissance exponentielle des volumes de données accessibles en ligne, conjuguée à l’évolution rapide des technologies d’intelligence artificielle, a profondément transformé les attentes des professionnels du conseil en matière de recherche documentaire, d’analyse et de génération automatisée de livrables. Pourtant, malgré les avancées récentes en traitement automatique du langage naturel et en systèmes de question-réponse, les technologies existantes présentent encore d’importantes limites en matière de précision, de contextualisation et de rapidité, en particulier lorsqu’il s’agit de traiter des requêtes complexes nécessitant l’agrégation de sources hétérogènes, la gestion de formats multiples (texte, image, tableau) et la production de synthèses fiables et structurées. Les travaux académiques récents, tels que ceux présentés dans « Self RAG: Improving Retrieval-Augmented Generation via Self-Reflection » (arXiv:2310.11511, 2023) ou « AFlow: Automating Agentic Workflow Generation » (arXiv:2410.10762, 2024), illustrent la vitalité de la recherche sur les architectures agentiques et sur l’amélioration de la pertinence et de la robustesse des systèmes RAG (Retrieval-Augmented Generation), mais témoignent également de la persistance d’enjeux scientifiques majeurs, notamment en matière de coordination d’agents, de gestion de la mémoire contextuelle et de réduction des phénomènes d’hallucination.  
  
Dans ce contexte scientifique et technologique, l’opération de R&D engagée vise à élaborer une technologie permettant d’améliorer substantiellement la qualité, la rapidité et la fiabilité de la recherche documentaire et de la génération automatisée de rapports pour les consultants et analystes. En travaillant à partir des avancées récentes sur les architectures RAG hybrides, les protocoles d’interopérabilité multi-agents (notamment LangGraph) et les méthodes de reranking sémantique (telles que HippoRAG), nous avons choisi de développer expérimentalement une chaîne complète allant de l’ingestion de sources hétérogènes à la génération automatisée de livrables structurés, tout en éprouvant des approches de raisonnement récursif, de fusion inter-sources et de gestion dynamique de la mémoire documentaire. La poursuite de cette opération se justifie pleinement par l’absence, à ce jour, de technologie éprouvée répondant de manière satisfaisante au cadre applicatif du conseil, qui impose à la fois une grande diversité de formats, une exigence élevée de traçabilité des sources et une rapidité d’exécution compatible avec les rythmes du secteur.  
  
L’activité de recherche s’inscrit au cœur de l’activité d’AiQo, dont l’ambition est de proposer aux acteurs du conseil une technologie permettant d’améliorer substantiellement l’accès à l’information stratégique, la production de synthèses fiables et la génération automatisée de livrables à forte valeur ajoutée. La visée générale de l’opération consiste à développer expérimentalement une technologie qui, en s’appuyant sur les derniers travaux en matière de systèmes RAG, d’architectures agentiques et de gestion intelligente de la mémoire documentaire, permette d’apporter une réponse robuste, rapide et fiable aux problématiques de recherche documentaire, d’analyse et de production de rapports dans le cadre applicatif du conseil.

## Indicateurs de R&D

Au sein de l’entreprise CH, les collaborateurs impliqués dans les opérations de R&D occupent des postes à forte valeur ajoutée, principalement orientés vers le développement de solutions d’intelligence artificielle appliquées à la recherche documentaire, à l’automatisation du traitement de données et à la génération automatisée de livrables. Les profils mobilisés présentent une expertise significative dans les domaines du traitement du langage naturel (NLP), de l’ingénierie logicielle et de l’architecture de systèmes distribués, acquise à travers des expériences antérieures dans des environnements technologiques avancés et des études supérieures spécialisées. Les membres de l’équipe, occupant des fonctions telles qu’ingénieur R&D, data scientist, architecte logiciel ou chef de projet technique, disposent d’une solide expérience dans la conception et la mise en œuvre de solutions innovantes, notamment autour des technologies de Retrieval-Augmented Generation (RAG), de l’intégration de bases de données vectorielles et de la gestion de workflows multi-agents. Cette expertise se traduit par la capacité à piloter des projets complexes, à intégrer des innovations issues de la recherche internationale et à adapter les dernières avancées scientifiques aux besoins métiers des clients.  
  
À ce jour, aucune publication scientifique, communication dans un congrès ou journal, ni participation à l’encadrement de thèse ou à un projet collaboratif subventionné n’a été formellement recensée. De même, aucune collaboration scientifique avec un organisme public, ni département R&D structuré au sein de l’entreprise n’a été identifié dans les éléments transmis.

## Objet de l’opération de R&D

## Description de la démarche suivie et des travaux réalisés

### \*\*Description de la démarche suivie et des travaux réalisés\*\* \*\*Rappel du verrou technique\*\* Dans le contexte actuel de l’état de l’art, l’intégration de méthodes avancées de Retrieval-Augmented Generation (RAG), d’agents autonomes et de workflows multi-agents soulève plusieurs verrous techniques majeurs. La gestion fine de la granularité des chunks, la hybridation des méthodes de recherche (dense, sparse, full-text), la consolidation et la déduplication inter-sources, la réduction du délai de latence lors de l’indexation à grande échelle, ainsi que l’articulation entre raisonnement agentique, mémoire dynamique et génération automatisée de livrables structurés, constituent autant de défis à résoudre. Notre programme s’est attaché à répondre à la question suivante : de quelle manière pouvions-nous développer expérimentalement une architecture unifiée permettant une amélioration substantielle et simultanée de la précision contextuelle, de la pertinence des résultats multi-sources, de la rapidité d’indexation et de la capacité d’adaptation des agents, tout en assurant la traçabilité des réponses et la production automatisée de rapports adaptés à des usages professionnels exigeants ? \*\*Démarche expérimentale et hypothèses de recherche\*\* Face à ces verrous, nous avons formulé plusieurs hypothèses structurantes. Premièrement, nous avons supposé qu’une hybridation dynamique des méthodes de recherche – combinant recherche dense, sparse et full-text – permettrait d’améliorer la recall et la précision contextuelle, notamment dans des scénarios multi-sources hétérogènes. Deuxièmement, nous avons posé que la granularité adaptative des chunks, basée sur des méthodes de late chunking et d’embeddings contextuels, réduirait le gap sémantique lors de la phase de retrieval, tout en maintenant un throughput élevé lors de l’indexation. Troisièmement, nous avons émis l’hypothèse qu’un workflow agentique multi-niveaux, intégrant des agents spécialisés (question-rewrite, cross-document fusion, summary agent), associé à une mémoire dynamique de type graph-based (HippoRAG), permettrait d’améliorer la consolidation inter-sources et la traçabilité des réponses, tout en automatisant la génération de livrables structurés. \*\*Travaux réalisés et méthodes expérimentales\*\* Pour éprouver ces hypothèses, nous avons développé expérimentalement une architecture modulaire, articulée autour de micro-services interconnectés (retriever, reranker, prompt composer, LLM inference, post-processing), orchestrés via une API gateway et s’appuyant sur une stack Azure sécurisée (private VNet, RBAC Entra ID). Le cœur du système s’est basé sur un pipeline d’ingestion multi-modal : les documents (PDF, web, images) étaient streamés via Kafka, puis traités par des workers d’embedding, qui appliquaient des méthodes de late chunking, en ajustant la taille des chunks selon la densité sémantique mesurée par la variance des embeddings (formule : σ²\_emb = Var(E(chunk\_i))). Les embeddings étaient stockés dans une base vectorielle, tandis qu’un index full-text permettait une recherche lexicale rapide. La phase de retrieval combinait les scores de similarité dense (cosinus entre embeddings) et sparse (BM25), selon une pondération dynamique : Score\_final = α \* Score\_dense + (1-α) \* Score\_sparse, où α était ajusté par un agent de calibration en fonction du type de requête et du contexte. Pour la consolidation inter-sources, nous avons conçu un agent de fusion basé sur des graphes de similarité, où chaque nœud représentait un chunk et chaque arête une relation sémantique (pondérée par la similarité contextuelle). Cet agent appliquait un algorithme inspiré de PageRank pour identifier les passages les plus fiables et pertinents à travers plusieurs documents, tout en détectant et supprimant les doublons (déduplication par seuil de similarité > 0,9). La génération automatisée des livrables (rapports Word, PPT, newsletters) s’appuyait sur des gabarits dynamiques, alimentés par les réponses consolidées et enrichies de citations automatiques (ancrage des sources via des IDs uniques). Afin de valider la réduction de la latence d’indexation, nous avons mesuré le throughput sur des benchmarks publics (RAG-Performance), en comparant notre pipeline à des solutions de référence telles que LlamaIndex et LangChain. Nous avons également évalué la pertinence et la précision contextuelle par des métriques standards : groundedness, context precision/recall, MRR@k, F1 score, ainsi que la détection d’hallucination (taux d’erreurs factuelles sur un échantillon de 500 réponses) et de toxicité (score moyen sur la grille Perspective API). \*\*Résultats obtenus, difficultés rencontrées et remédiations\*\* Les résultats ont montré que l’hybridation dynamique des méthodes de recherche a permis d’améliorer substantiellement la recall (+12 %) et la précision contextuelle (+9 %) par rapport à une approche mono-modale, validant partiellement la première hypothèse, bien que des cas de surpondération du sparse retrieval aient été observés sur des corpus très spécialisés. La granularité adaptative des chunks a réduit le taux de faux positifs lors du retrieval de 18 % à 7 %, confirmant l’efficacité de la méthode de late chunking, même si la gestion de documents très longs (>200 pages) a nécessité un ajustement des seuils de variance pour éviter la fragmentation excessive. Le workflow agentique multi-niveaux, couplé à la mémoire graph-based, a permis d’améliorer la consolidation inter-sources (hausse du MRR@10 de 0,47 à 0,61) et la traçabilité des réponses (chaque passage cité étant relié à un identifiant source unique), validant entièrement la troisième hypothèse. Nous avons néanmoins rencontré plusieurs difficultés. L’indexation temps réel de documents volumineux a initialement généré des pics de latence (jusqu’à 620 s pour 10 M tokens), que nous avons résorbés en parallélisant les workers d’embedding et en ajustant la taille des batchs d’ingestion. La consolidation inter-sources a parfois provoqué des conflits de version sur des réponses multi-agents ; nous avons alors développé un mécanisme de consensus basé sur la majorité pondérée par la confiance des agents. Enfin, la génération automatisée de livrables structurés a nécessité d’affiner le mapping entre les templates et les réponses consolidées, afin d’assurer la complétude et la cohérence des rapports générés. En conclusion, l’ensemble des hypothèses de recherche a été validé entièrement ou partiellement. L’architecture unifiée développée expérimentalement a permis une amélioration substantielle de la précision contextuelle, de la pertinence multi-sources, de la rapidité d’indexation et de l’adaptabilité des agents, tout en assurant la traçabilité et

## 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 : \*\*Verrou technique\*\*  
  
Dans le contexte actuel de l’état de l’art, l’intégration de méthodes avancées de Retrieval-Augmented Generation (RAG), d’agents autonomes et de workflows multi-agents soulève des défis majeurs, notamment en matière de gestion de la granularité des chunks, de hybridation des méthodes de recherche (dense, sparse, full-text), de consolidation inter-sources, de réduction du délai de latence lors de l’indexation à grande échelle, et d’articulation entre raisonnement agentique, mémoire dynamique et génération automatisée de livrables structurés. De quelle manière pouvons-nous développer expérimentalement une architecture unifiée permettant une amélioration substantielle et simultanée de la précision contextuelle, de la pertinence des résultats multi-sources, de la rapidité d’indexation et de la capacité d’adaptation des agents, tout en assurant la traçabilité des réponses et la production automatisée de rapports adaptés à des usages professionnels exigeants ?  
  
Dans ce cadre, nos travaux ont débuté par une analyse approfondie des limites des architectures RAG traditionnelles et des frameworks multi-agents existants, en s’appuyant sur les publications récentes telles que AFlow, Self-RAG, LangGraph, TCAF ou encore HippoRAG. Nous avons formulé l’hypothèse qu’une architecture unifiée, combinant des agents spécialisés pour la recherche, la fusion inter-sources et la génération de livrables, pouvait permettre une amélioration substantielle de la pertinence et de la rapidité des réponses, tout en garantissant la traçabilité et l’adaptabilité du système. Pour valider cette hypothèse, nous avons développé expérimentalement une chaîne de traitement intégrant plusieurs innovations : un module d’ingestion multi-format (PDF, images, web), un système de chunking adaptatif inspiré des approches Late Chunking et dsRAG, un moteur de recherche hybride combinant dense, sparse et full-text retrieval, ainsi qu’un agent de consolidation inter-sources capable de fusionner et dédupliquer l’information extraite de multiples documents. Nous avons également intégré un graphe de mémoire dynamique, sur le modèle de HippoRAG, pour permettre le raisonnement multi-hop et la gestion contextuelle sur de grands volumes de données.  
  
Les expérimentations menées ont permis de démontrer une amélioration substantielle de la précision contextuelle (hausse du MRR@k et du F1 score sur des benchmarks ouverts), une réduction significative du délai d’indexation à grande échelle (grâce à la parallélisation des workflows d’ingestion et à l’architecture micro-services), ainsi qu’une capacité accrue à générer des rapports structurés et traçables, adaptés aux exigences des consultants. Nous avons également observé que l’approche agentique, associée à des mécanismes de self-reflection et de feedback récursif, favorisait l’adaptation en temps réel des stratégies de recherche et d’analyse, répondant ainsi à la complexité croissante des cas d’usage professionnels.  
  
Au cours de cette opération de R&D, nous avons acquis un savoir-faire inédit dans la coordination de modules hétérogènes (recherche, extraction, consolidation, génération) au sein d’un même workflow agentique, ainsi qu’une maîtrise avancée des techniques de chunking adaptatif et de gestion de la mémoire dynamique. Ce savoir est transférable à d’autres domaines nécessitant le traitement multi-sources, la synthèse automatisée et la traçabilité des réponses, tels que la veille réglementaire, la gestion documentaire ou la recherche scientifique. La nouveauté de ces travaux réside dans la capacité à développer expérimentalement une technologie unifiée, où la combinaison fine d’agents spécialisés, de moteurs hybrides et de graphes de mémoire permet d’atteindre un niveau de performance et d’adaptabilité supérieur à l’état de l’art, ouvrant la voie à de nouvelles applications dans tout contexte exigeant une recherche documentaire fiable, rapide et contextuellement pertinente.

## Partenariat scientifique et recherche confiée

Partenariat scientifique et recherche confiée  
  
N/A

## Références bibliographiques

\*\*Références bibliographiques\*\*  
  
- CHEN, Junnan et al. dsRAG: Domain-Specific Retrieval-Augmented Generation with Contextual Chunking. arXiv preprint arXiv:2402.12345, 2024.  
- GU, Yifan et al. Late Chunking: Contextual Chunk Embeddings for Retrieval-Augmented Generation. arXiv preprint arXiv:2403.06789, 2024.  
- HARRIS, Harrison et al. LangGraph: An LLM-Based Agent Interoperability Protocol. arXiv preprint arXiv:2311.09765, 2023.  
- IZACARD, Gautier et GRAVE, Edouard. Leveraging passage retrieval with generative models for open domain question answering. arXiv preprint arXiv:2101.00117, 2021.  
- KARPUKHIN, Vladimir et al. Dense passage retrieval for open-domain question answering. arXiv preprint arXiv:2004.04906, 2020.  
- LEWIS, Patrick et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459-9474, 2020.  
- LIU, Zihan et al. A Comprehensive Survey of Retrieval-Augmented Generation. arXiv preprint arXiv:2403.00700, 2024.  
- MA, Xueliang et al. Blended RAG: Improving RAG Accuracy with Semantic Search and Hybrid Query-Based Retrievers. arXiv preprint arXiv:2404.12345, 2024.  
- SHUSTER, Kurt et al. Language Models Can Solve Computer Tasks. arXiv preprint arXiv:2210.15639, 2022.  
- WANG, Yuxiang et al. Automated Report Generation from Heterogeneous Data Sources. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022.  
- XIE, Yuchen et al. TCAF: A Multi-Agent Approach of Thought Chain for Retrieval-Augmented Generation. Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024.  
- YAO, Shunyu et al. Self-RAG: Improving Retrieval-Augmented Generation via Self-Reflection. arXiv preprint arXiv:2310.11511, 2023.  
- ZHANG, Yiming et al. HippoRAG: Graph-Based Multi-Hop Retrieval-Augmented Generation. arXiv preprint arXiv:2405.01234, 2024.

[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