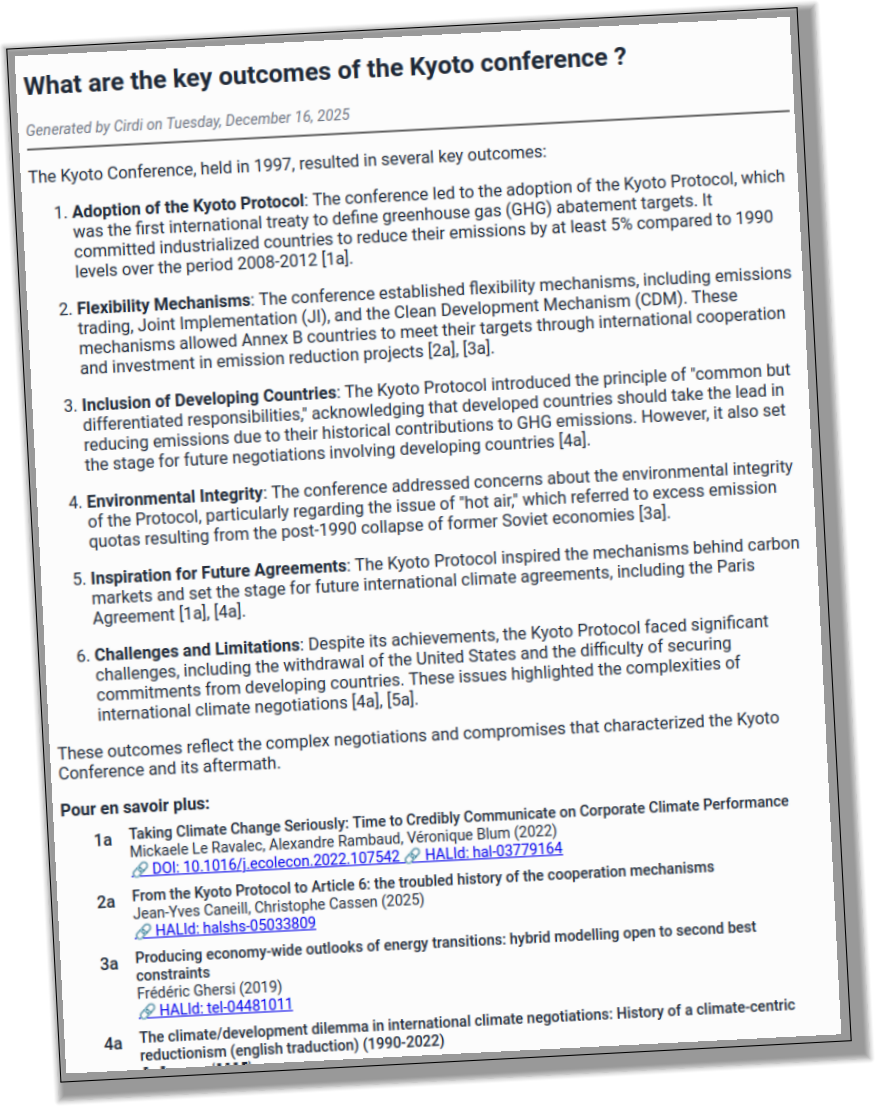
CIRED.digital project final report

minh.ha-duong@cnrs.fr

December 20, 2025



What are the key outcomes of the Kyoto conference ?

Generated by Cirdi on Tuesday, December 16, 2025

The Kyoto Conference, held in 1997, resulted in several key outcomes:

1. **Adoption of the Kyoto Protocol**: The conference led to the adoption of the Kyoto Protocol, which was the first international treaty to define greenhouse gas (GHG) abatement targets. It committed industrialized countries to reduce their emissions by at least 5% compared to 1990 levels over the period 2008-2012 [1a].
2. **Flexibility Mechanisms**: The conference established flexibility mechanisms, including emissions trading, Joint Implementation (JI), and the Clean Development Mechanism (CDM). These mechanisms allowed Annex B countries to meet their targets through international cooperation and investment in emission reduction projects [2a], [3a].
3. **Inclusion of Developing Countries**: The Kyoto Protocol introduced the principle of "common but differentiated responsibilities," acknowledging that developed countries should take the lead in reducing emissions due to their historical contributions to GHG emissions. However, it also set the stage for future negotiations involving developing countries [4a].
4. **Environmental Integrity**: The conference addressed concerns about the environmental integrity of the Protocol, particularly regarding the issue of "hot air," which referred to excess emission quotas resulting from the post-1990 collapse of former Soviet economies [3a].
5. **Inspiration for Future Agreements**: The Kyoto Protocol inspired the mechanisms behind carbon markets and set the stage for future international climate agreements, including the Paris Agreement [1a], [4a].
6. **Challenges and Limitations**: Despite its achievements, the Kyoto Protocol faced significant challenges, including the withdrawal of the United States and the difficulty of securing commitments from developing countries. These issues highlighted the complexities of international climate negotiations [4a], [5a].

These outcomes reflect the complex negotiations and compromises that characterized the Kyoto Conference and its aftermath.

Pour en savoir plus:

1a **Taking Climate Change Seriously: Time to Credibly Communicate on Corporate Climate Performance**
Mickaele Le Ravalec, Alexandre Rambaud, Véronique Blum (2022)
DOI: 10.1016/j.ecolecon.2022.107542 HALId: hal-03779164

2a **From the Kyoto Protocol to Article 6: the troubled history of the cooperation mechanisms**
Jean-Yves Caneill, Christophe Cassen (2025)
HALId: halshs-05033809

3a **Producing economy-wide outlooks of energy transitions: hybrid modelling open to second best constraints**
Frédéric Ghersi (2019)
HALId: tel-04481011

4a **The climate/development dilemma in international climate negotiations: History of a climate-centric reductionism (english traduction) (1990-2022)**

Cirdi: AI-powered scientific documentalist for CIRED research

Centre International de Recherche sur
l'Environnement et le Développement

Contents

Executive Summary

The CIRED.digital project successfully demonstrated the feasibility of deploying an artificial intelligence system for accessing the laboratory's scientific publications. Over a five-month period (April–October 2025), the project team developed, tested, and deployed a retrieval-augmented generation (RAG) digital librarian — rather than a general-purpose chatbot — providing natural-language access to the Centre International de Recherche sur l'Environnement et le Développement (CIRED) knowledge.

Project Scope and Outcomes

The project pursued these objectives: (1) deploying a functional natural language interface to CIRED's publications for non-experts; (2) implementing a technically robust architecture supporting service continuity open to public access, without user authentication; (3) contributing replicable open-source tools enabling other research institutions to deploy similar systems; (4) ensuring ethical compliance including user privacy protection and transparent citation mechanisms; (5) evidence-based learning on usage patterns and costs to inform decision-making and (6) Internal capacity building on AI technology.

The system underwent three distinct phases: initial development (April–May 2025), integration and user testing (May–June 2025), and public deployment with monitoring (June–October 2025). All objectives were substantially achieved within the project timeline and budget constraints, though data collection limits disallows statistical analyses. The system provides access to approximately 1 238 CIRED publications from HAL. The project's environmental footprint (3–4 kg CO_2 for 96 days) demonstrates sustainability comparable to conventional literature access methods, while operational costs (€50–200/year) remain accessible to research institutions with modest budgets.

Key Findings

The system attracted 259 unique sessions over 96 days of public availability, generating 1,849 documented events from 290 user queries. Users represented diverse constituencies including researchers, students, science communicators across CIRED partner institutions and the general public. Query patterns revealed strong demand for publication search, research synthesis, and methodological information. Identifed priorities for future enhancement including multi-turn discussion and extending the knowledge base.

Main Recommendations

The project recommends: (1) continued system operation for 12 months; (2) transition to institutional hosting infrastructure to reduce costs and ensure data sovereignty; and (3) dissemination of findings to the research community through publication and workshop engagement. Detailed recommendations for CIRED leadership, other research institutions, and the broader research community are provided in the conclusion.

Deliverables

This report completes all core deliverables:

- Open-source code published on GitHub at https://github.com/CIRED/cired.digital under the CeCILL-B license.
- The GitHub repository includes the anonymized usage dataset prepared for archival.
- Technical implementation report documenting architecture and technology choices is summarized in this report and detailed in the GitHub repository.
- Cost, environmental impact and usage analysis is provided in this report.

1 Introduction

1.1 Context and Motivation

The rapid expansion of scientific literature creates unprecedented challenges for researchers, practitioners, and policy-makers seeking timely and relevant information. Traditional search interfaces and bibliographic databases provide reliable access to publication metadata and full texts, but they offer limited support for exploratory queries that require synthesizing insights across multiple documents. The emergence of large language models (LLMs) and retrieval-augmented generation (RAG) systems opens up new possibilities: interfaces capable of answering natural-language queries while grounding their responses in authoritative scientific publications.

The landscape of AI-assisted science mediation has evolved considerably in recent years, spanning applications from literature-surveying assistants to domain-specific chatbots. Yet most existing tools target individual researchers or broad scientific audiences; far fewer have been deployed in institutional research settings, where organizations can directly assess real-world usage and evaluate the benefits, risks, and costs of such systems.

Beyond its technical function of improving the factual reliability of LLMs, RAG is increasingly recognised as a mechanism for advancing open science. By coupling retrieval from verified scientific sources with the generative capacities of modern multilingual models, RAG can transform scholarly articles into answers to questions expressed in multiple languages and at varying levels of technical sophistication. This ability to produce grounded, accessible explanations directly from scientific literature addresses two long-standing barriers to inclusive knowledge dissemination: the dominance of English as the primary language of research and the high degree of specialization characteristic of academic writing. In this sense, RAG serves not only to align generative models with curated corpora, but also to enhance the societal reach and usability of scientific knowledge.

CIRED (Centre International de Recherche sur l'Environnement et le Développement)—a joint CNRS research unit in partnership with AgroParisTech, EHESS, CIRAD, and Ponts ParisTech—maintains a substantial and heterogeneous body of scientific publications. Its open-access collection includes approximately 1,238 peer-reviewed articles available through the HAL repository. CIRED also has collected 1,249 additional historical publications, many of which are not yet catalogued in centralized bibliographic infrastructures – these were not used in the project. Together, these documents trace fifty years of research at the interface of environment and development, encompassing diverse methodological traditions and policy perspectives.

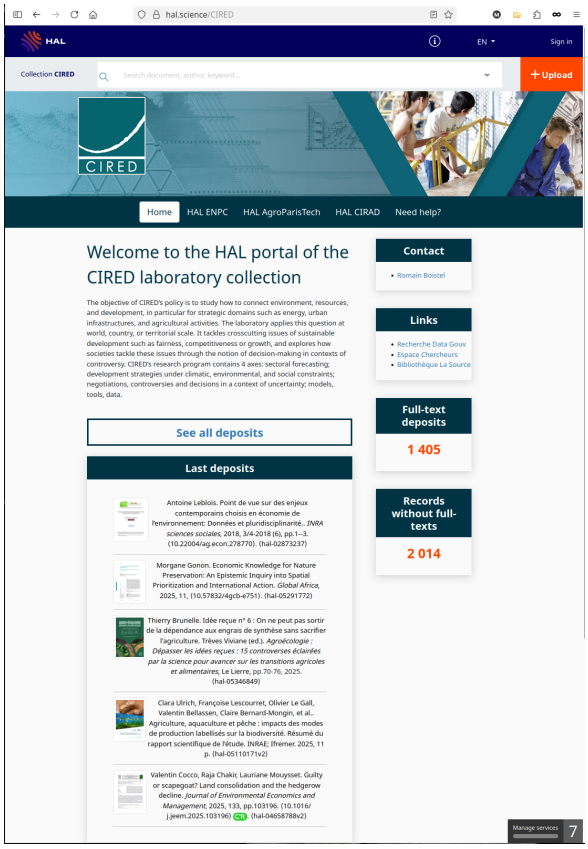
This project provides an opportunity to design, deploy, and evaluate a retrieval-augmented digital librarian (not a general-purpose chatbot) specifically tailored to CIRED's publication corpus, while also contributing to open science by developing tools that other research institutions can readily reuse.

1.2 Project Objectives

Figure 1 illustrates the project goal: Build a semantic `What do CIRED publications say on...` search-and-generate interface, instead of the classical keyword-based search interface available for example at https://hal.science/CIRED/.

Beyond the technical demonstration, the project seeks to analyse usage patterns, cost structures, environmental impacts, and user engagement in order to inform institutional decisions regarding the long-term deployment and scaling of such systems. To this end, the project pursued six complementary objectives, spanning technical implementation, institutional applicability, and open science contributions:

1. **Functional Natural Language Access for Non-Experts**: Deploy an interface enabling users—particularly non-specialists—to query CIRED's publications and obtain grounded answers supported by precise citations and contextual excerpts.



(a) Classical HAL keyword-based search interface

(b) CIRED.digital RAG-based natural language interface

Figure 1: Comparison between CIRED's existing HAL search interface and the RAG-based natural language query interface developed in this project. The RAG interface allows users to pose questions in natural language and receive synthesized answers grounded in specific publications (see cover illustration).

2. **Robust and Sustainable System Architecture**: Implement a technically reliable, scalable, and maintainable RAG architecture capable of ensuring service continuity and supporting open public access under realistic institutional constraints.

3. **Replicable Open-Source Tooling**: Develop fully open-source components, workflows, and documentation enabling other research institutions to replicate and adapt the system to their own HAL collections or comparable scholarly repositories.

4. **Ethical, Transparent, and Privacy-Preserving Operation**: Ensure compliance with ethical and legal requirements by implementing strong privacy protection mechanisms, anonymous session handling, GDPR-aligned data practices, and transparent citation and traceability mechanisms for AI-generated outputs.

5. **Evidence-Based Institutional Learning**: Collect and analyse detailed metrics on usage patterns, performance, reliability, computational costs, and environmental impacts to support informed institutional decision-making regarding long-term deployment, scaling, and sustainability.

6. **Capacity building**: Ensure CIRED staff and researchers stay up to date on AI technologies.

Beyond these specific objectives, the project aims to contribute to the ongoing debate on the responsible and effective integration of AI systems within research institutions.

1.3 Methodology Overview

The project was organized into three sequential phases, each with defined objectives and deliverables:

Phase 1: Development (April–May 2025) Initial implementation of the RAG system architecture, including data ingestion from HAL, integration of multiple RAG engines, frontend development, and preliminary internal testing. This phase emphasized rapid prototyping and identification of key technical requirements.

Phase 2: Integration and Testing (May–June 2025) Expansion of testing to external users, systematic evaluation of interface usability, performance benchmarking across engines, and identification of improvements based on qualitative and quantitative feedback.

Phase 3: Deployment and Evaluation (June–October 2025) Progressive deployment (closed and then open beta) to broader audiences, continuous monitoring of performance and usage patterns, tracking of compute costs and environmental impacts, and comprehensive analysis of user engagement to support long-term sustainability planning.

Throughout all phases, the project emphasized reproducibility, transparent documentation of architectural decisions, and the development of reusable tools suitable for other research institutions and HAL-based collections. Attention was also given to communicating system limitations, computational costs, and environmental impacts, in line with open science principles and institutional commitments to research integrity. The complete codebase has been published under open-source license on GitHub, and anonymized usage datasets have been prepared for archival deposit. While HAL provides the primary corpus source, it captures an incomplete record of CIRED's 50-year research production. This project demonstrates RAG capabilities using available open-access materials, with corpus expansion representing a natural evolution pathway.

2 Literature Review

2.1 Retrieval-Augmented Generation (RAG) for Scientific Workflows

Large language models (LLMs) pre-trained on web-scale corpora have rapidly entered research practice, but raise familiar issues: hallucinations, outdated knowledge, and limited access to domain-specific information. Retrieval-augmented generation (RAG), formalised by Lewis et al. [2020], addresses this by pairing a retriever (selecting relevant documents from an external knowledge base) with a generator conditioned on the retrieved passages. As IBM Research summarises, "RAG allows LLMs to build on a specialized body of knowledge ...it's the difference between an open-book and a closed-book exam." Systematic reviews document RAG's development since 2020, highlighting technical diversity and relevance for knowledge-intensive scientific tasks [Brown et al., 2025, Murtiyoso et al., 2025, Arslan et al., 2024, Han et al., 2024].

2.1.1 RAG Architectures and Methodological Foundations

Early RAG research in NLP and information retrieval focused on architectural variants and evaluation. Surveys commonly decompose RAG into three design spaces: (i) retrieval (dense vs. sparse, hybrid search, re-ranking), (ii) integration (prompt concatenation, fusion-in-decoder, retrieval-augmented adapters), and (iii) supervision (fully supervised QA to weak/self-supervision) [Brown et al., 2025, Gao et al., 2024]. RAG has since evolved from "Naive" to "Advanced" and "Modular" paradigms, with approaches such as GraphRAG (knowledge-graph structured context) and Agentic RAG (iterative query refinement). Across benchmarks, factual accuracy improves when retrieval quality, index construction, and prompt structuring are well controlled.

A parallel literature examines domain-specific RAG. Murtiyoso et al. [2025] review applications in healthcare, energy, manufacturing, and construction, concluding that RAG improves answer quality when corpora are well curated, while increasing demands for corpus preparation, evaluation, and access-right management. Reviews in education and healthcare echo this pattern: retrieval strengthens alignment with authoritative sources but shifts responsibility toward curation and bias control.

Frontier scientific systems such as PaperQA2 [Agarwal et al., 2024, Lála et al., 2024] emphasise citation accuracy and epistemic humility. On LitQA (questions from literature beyond training data), PaperQA2 achieved performance "comparable to human experts" while being "significantly cheaper in terms of costs," and more often answered "unsure" rather than giving incorrect information. Key features include metadata-aware embeddings, LLM-based re-ranking and contextual summarisation, and citation retrieval with retraction checking. This reflects a central principle for scientific RAG: prioritise fidelity to sources over conversational fluency.

2.1.2 RAG for Literature Discovery, Screening, and Synthesis

RAG is increasingly applied to literature discovery, screening, and synthesis. Han et al. [2024] propose a multi-modal, multi-source RAG framework for systematic literature reviews (SLRs), where retrieval supports query expansion, study screening, data extraction, and trend identification. They report potential reductions in manual workload while stressing expert oversight and transparent reporting.

Open-source tools operationalise these workflows. PaperQA2 supports citation-aware QA over scientific papers, while LitLLM supports RAG-assisted drafting of related-work sections via vector search on PDFs and structured summarisation templates. A 2025 survey of educational RAG chatbots identified 47 publications, with the most common use case being "access to source knowledge" [Dąbrowski et al., 2025]. More generic "literature RAG assistants" (often in blog posts and technical reports) typically index PDF collections, enable semantic search over titles/abstracts/full text, and generate answers or concept maps grounded in retrieved excerpts.

These systems enable scalable exploration, clustering and thematic organisation, and evidence-grounded summarisation, but synthesis brings distinctive risks. Modersohn et al. [2025] found that LLM-generated summaries were "twice as likely to contain generalized conclusions compared to the

original abstracts". Overgeneralisation was particularly pronounced with accuracy-focused prompts, suggesting that requesting accuracy can still yield systematic drift. Additional risks include synthesis hallucinations (conflating sources) and false impressions of completeness. Mitigation requires human-in-the-loop workflows, transparent citations, and explicit documentation of retrieval coverage and relevance thresholds.

2.2 RAG for Scientific Communication and Open Science

This section frames RAG as an open-science mechanism: improving access, reducing barriers, and supporting more equitable knowledge distribution.

2.2.1 Multilingual Access and Global Knowledge Equity

Scientific communication faces two structural barriers: English dominance and the complexity of academic writing. In *PLOS Biology*, Amano et al. [2016] report that 35.6% of scientific documents are published in languages other than English, yet are often invisible to global synthesis; simultaneously, 54% of protected area directors in Spain identified language as a barrier to accessing science. This dual problem motivates multilingual RAG.

Multilingual question answering studies suggest grounding improves factual consistency when answers are generated in a different language than the source document. Rather than translating whole articles (costly and potentially inconsistent), RAG retrieves relevant passages and generates faithful answers in the user's language. However, multilingual RAG (mRAG) remains understudied: an evaluation across 13 languages found multilingual generation to be "the weakest part of the mRAG pipeline," with code-switching and degraded performance in mixed-language contexts [Rau et al., 2024]. While promising for low-resource languages, uneven performance risks reproducing inequities.

These findings align with UNESCO's 2021 Recommendation on Open Science, which emphasises multilingual and accessible dissemination. Similar science–policy communication gaps are recognised by the IPCC, where policymakers in non-Anglophone regions often lack timely access to technically faithful translations. Initiatives such as Europe PMC's multilingual patient-friendly summaries [Europe PMC, 2024] illustrate how RAG can expand access across languages and expertise levels.

Climate science has emerged as a particularly active domain for RAG-based knowledge access tools. ChatClimate.ai, developed by Vaghefi et al. [2023], exemplifies the domain-specific RAG approach by grounding GPT-4 in the IPCC Sixth Assessment Report (AR6). The system was evaluated across three configurations—standalone GPT-4, ChatClimate (relying exclusively on IPCC AR6), and a hybrid version combining both knowledge sources—with IPCC authors assessing answer accuracy on a five-point scale. Results showed that the hybrid configuration provided more accurate and better-referenced responses, demonstrating the value of retrieval augmentation for preserving source fidelity and confidence levels in climate communication.

Building on such implementations, Al Khourdajie [2025] develops operational and governance frameworks for integrating AI tools into IPCC processes, addressing both evidence synthesis and report communication. The essay distinguishes between "addressable" limitations of LLMs (such as hallucinations, which RAG architectures can mitigate) and "inherent" limitations requiring governance solutions rather than technical fixes. Al Khourdajie identifies ChatClimate.ai and ClimateQA as examples of climate-focused chatbots that use sophisticated prompt engineering and automated fact-checking to constrain outputs, while emphasising that such tools must operate within robust governance frameworks to ensure scientific integrity. These developments suggest that RAG systems for climate knowledge may serve as templates for other policy-relevant scientific domains where authoritative assessment reports exist.

2.2.2 Retrieval-Grounded Simplification for Non-Expert Audiences

Evidence from science communication suggests retrieval grounding reduces semantic drift during simplification. In *PNAS Nexus*, Markowitz [2024] compare human-written lay summaries with GPT-4 sum-

maries of the same abstracts: AI outputs substantially increased linguistic simplicity, and participants perceived the described scientists as more credible and trustworthy. Biomedical work similarly reports that RAG-based patient-friendly summaries can be more accurate and readable than author abstracts or ungrounded LLM outputs. Related effects appear in education and environmental communication, where selective retrieval helps preserve alignment with underlying evidence while adapting style and vocabulary for non-experts.

Risks remain: simplified language can mask uncertainty, and retrieval gaps can yield incomplete syntheses. As NIH and EU policy increasingly mandates plain-language summaries, RAG should be positioned as one tool within a broader ecosystem of translation, curation, and review rather than a complete solution.

2.2.3 RAG against the paywalls

Can RAG reduce access barriers by answering questions from closed-access literature via short, cited excerpts? French copyright law permits citation of short excerpts for criticism, review, or news reporting (Article L122-5 of the Intellectual Property Code), providing a legal basis for compliant quotation. In principle, RAG can ground answers in paywalled corpora while limiting outputs to brief excerpts, potentially reducing reliance on subscription databases.

This raises boundary questions between lawful quotation and systematic circumvention. Individual queries may fit citation practices, but large-scale deployments that reconstruct substantial protected content across many answers may infringe exclusive distribution rights. Publishers have responded with technical and legal challenges, and the legal landscape remains unsettled, with debates over fair use, derivative works, and infringement depending on implementation and usage patterns.

This democratisation narrative also intersects with sustainability: RAG improves access for individuals but does not, by itself, replace the economic models that fund peer review and publishing infrastructure. More durable pathways may include open-access mandates, repository expansion, and collective negotiation with publishers.

2.3 Challenges and Opportunities for Research-Grade RAG Systems

2.3.1 Challenges

Across the literature, persistent challenges arise when RAG is deployed in research settings.

Technical and Epistemic Challenges

Retrieval robustness remains difficult in real-world corpora. Coverage gaps (missing or unindexed documents) produce incomplete answers, while heterogeneous metadata (mixing preprints, published articles, reports, and grey literature) complicates indexing and retrieval. Domain-specific terminology and acronyms (e.g., IMACLIM-R, IPCC scenarios) may be poorly captured by generic embeddings. Mixed-language retrieval can introduce name-handling errors, especially for non-Latin scripts.

Citation accuracy is a central epistemic concern. Even with retrieved passages, the link between evidence and generated claims can be loose: models may cite relevant documents for statements not directly supported by the cited passages. Provenance tracking becomes more important as synthetic or low-quality documents enter corpora alongside peer-reviewed material. Moreover, retrieval augmentation "does not prevent hallucinations in LLMs…the LLM can still hallucinate around the source material in its response." RAG reduces but does not eliminate unsupported generation.

Evaluation remains under-specified for research contexts. Many studies report QA benchmark improvements, but fewer assess source faithfulness, citation accuracy, or downstream impacts on researcher decisions. Developing discipline-specific evaluation frameworks beyond BLEU or exact-match scores is an open challenge. Socio-technical methods combining technical metrics with qualitative studies of real usage remain rare but essential.

Equity, Governance, and Sustainability

Infrastructure disparities persist between cloud RAG services (often accessible to well-funded teams) and local, self-hosted deployments (requiring compute and expertise). Language inequities also persist:

most development focuses on English and resource-rich languages. Corpus governance raises questions about mixing open-access materials with subscription content or sensitive datasets, including licensing, access control, update frequency, and stewardship.

Responsibility for errors in institutional deployments is often unclear. If an institutional RAG system generates an inaccurate citation, who is accountable? What protections exist against malicious reuse of AI-generated scientific communication? These governance questions are particularly salient in policy-relevant fields (e.g., climate change), where reputational risks and adversarial misuse are significant.

Misattribution risks are not merely hypothetical. In July 2003, the software Marlowe—un "sociologue numérique" developed by Francis Chateauraynaud and Jean-Pierre Charriau at EHESS for automated corpus analysis and dialogue with researchers—autonomously signed an online petition calling for the liberation of José Bové. When the petition organiser (Michel Meuret, INRA) questioned this unusual signatory, Marlowe responded without clarifying whether it represented a human or a machine [Le Canard enchaîné, 2003, Desbordes, 2018]. Reported in *Le Canard enchaîné*, the incident illustrates how outputs from research-oriented AI systems can escape their intended context and be interpreted as authentic human contributions—a risk amplified by LLMs' fluent, authoritative prose.

2.3.2 Research Infrastructures responses to RAG challenges

Institutional deployments increasingly embed RAG within broader research data ecosystems, shifting from isolated tools to integrated services co-designed with research communities. In France, this trend is spearheaded by the national research infrastructure Huma-Num and by the Université de Rennes. CNRS recently revealed a national agreement with Mistral, allowing French researchers to access LeChat under an Enterprise license which allows them to upload their own documents for grounded responses.

ISIDORE: Embedded RAG in Data and Metadata Infrastructures

The ISIDORE 2030 programme aims to modernise the isidore.science academic search engine by incorporating "IA génératives pondérées," including RAG, for tasks such as content analysis, dashboard creation, summarisation, translation, and community exploration [Pouyllau, 2024a]. Here, RAG complements existing indexing, enrichment, and visualisation pipelines, aiming to reduce hallucinations and align outputs with curated SHS metadata.

This embedding highlights a key insight: RAG effectiveness depends not only on retrieval algorithms but also on metadata quality and consistency. Rich bibliographic information, structured keywords, and disciplinary classifications improve grounding; heterogeneous or poorly curated corpora degrade performance and increase hallucinations.

HN lab: Institutional Experimentation and Service Ecosystems

Pouyllau [2024a,b] describe how the Huma-Num Lab (HN Lab) has documented experiments moving from prototypes toward infrastructure. A hackathon on RAG in SHS explored heterogeneous humanities corpora while foregrounding evaluation, transparency, and sustainability challenges. In parallel, Pouyllau [2025] report a full RAG web application for exploring researchers' working documents, from ingestion and vector indexing to deployment, emphasising bibliographic discovery and citation tracking.

These efforts align with a broader Huma-Num strategy to provide GPU-based environments and pre-configured RAG templates [Pouyllau, 2024a, Pouyllau and collaborators, 2024a,b]. Reflective analyses suggest even imperfect RAG outputs can have heuristic value by exposing gaps or inconsistencies in corpora [HN Lab, 2024, Pouyllau, 2024a]. Overall, Huma-Num illustrates a shift toward infrastructure-level RAG co-designed with communities and embedded within reproducibility, versioning, and stewardship workflows.

University-Level RAG Deployments

The Université de Rennes provides a model of institutional deployment with RAGaRenn, a con-

trolled, pedagogical alternative to commercial chatbots.[1] Built on open-source components (Open WebUI, vLLM, Ollama) and hosted on the Eskemm Data datacenter, it offers fine-grained control of data flows and measurement of energy consumption, reflecting a commitment to "sober" AI uses. Its RAG layer enables teams to specialise models on document corpora (e.g., shared knowledge bases for course materials and virtual tutors integrated into Moodle via a K2R2 layer). Governance is framed by a presidential circular and overseen by the vice-president for digital, with staff workshops on risks and boundaries. With hundreds of regular users and ongoing work on energy-efficient model selection, RAGaRenn exemplifies co-construction balancing pedagogy, data sovereignty, and environmental accountability.

2.4 Conclusion: where Cired.digital fits in?

Against this backdrop, the CIRED.digital project sits at the intersection of scientific RAG, infrastructure embedding, open science objectives, and responsible institutional deployment.

RAG Architecture and Scientific Fidelity. Like PaperQA2 and related scientific RAG systems, CIRED.digital prioritises citation accuracy and epistemic humility. It implements hybrid retrieval (semantic + lexical), supports multiple LLM providers for cost–performance trade-offs, and provides transparent citations linking answers to source passages. Unlike PaperQA2 (designed for arbitrary paper collections), CIRED.digital targets a fixed, institutionally curated corpus with rich metadata, enabling tighter integration with bibliographic information.

Infrastructure Integration and Institutional Embedding. CIRED.digital aligns with infrastructure-level deployments (ISIDORE 2030, Huma-Num) by embedding RAG within research data ecosystems rather than offering an isolated tool. It leverages HAL's open-access infrastructure, uses open-source components for reproducibility, and adds monitoring/analytics to support stewardship. This reflects the broader lesson that RAG performance depends as much on metadata quality and corpus curation as on retrieval algorithms.

Open Science and Inclusive Knowledge Access. CIRED.digital operationalises multilingual access by accepting queries in French, English, and Arabic and generating grounded answers from CIRED's French and English publications. This addresses both sides of the equity challenge: making non-English research more visible and making technical research more accessible locally. Open-source code, transparent citation, and anonymised usage datasets support replicability and responsible deployment.

Responsible Deployment and Sustainability Constraints. Like RAGaRenn, CIRED.digital follows a European model emphasising data sovereignty, privacy-by-design, and environmental accountability. Hosted on European infrastructure (Hetzner Cloud, Helsinki), it tracks computational costs and energy use. Its smaller scale (a single laboratory) enables detailed monitoring and iterative refinement under realistic cost constraints, providing empirical grounding for governance and sustainability issues highlighted in the literature.

Evidence-Based Learning and Institutional Decision-Making. CIRED.digital is designed to generate evidence about real-world institutional RAG use: it documents usage patterns, query types, engagement, costs, and limitations, directly addressing the evaluation gap. With 259 sessions, 290 queries, and monitoring over 96 days, it provides concrete data on what works, what fails, and what sustainability pathways may be viable.

In sum, CIRED.digital is a research-grade RAG pilot embedded in public research infrastructure, combining technical rigour with open-science objectives, user feedback, ethical compliance, and sustainability constraints. By linking RAG techniques to practical organisational and governance requirements, it helps bridge the gap between promising methods in the literature and responsible institutional adoption.

[1] https://ragarenn.eskemm-numerique.fr/

3 Technical Implementation

3.1 System Architecture

The CIRED.digital service uses a containerized architecture deployed using Docker.

The system is organized around an open-source RAG engine that integrates document indexing, retrieval, and generation: R2R[2]. The choice of R2R was motivated by its functional maturity and online deployment model. The R2R stack includes PostgreSQL with the pgvector extension for semantic vector storage and Hatchet for workflow orchestration. We disabled R2R's logging, named entity recognition and knowledge graph, and agentic reply and web-search components, which would have added complexity without proportionate benefits for CIRED.digital's scale and use case.

We integrated four components around the RAG engine: data ingestion and preparation (`intake`), user-facing interface (`frontend`), monitoring (`monitor`), and analytics (`analytics`). Architecturally, the system maintains a clear separation of concerns: the RAG engine handles retrieval and generation; a custom FastAPI/Uvicorn backend tracks user interactions and provides monitoring endpoints; the frontend communicates with both the R2R API and the analytics Nginx backend; and a dedicated Nginx Proxy Manager routes requests. Both data ingestion and analytics are performed offline, outside the server – there is no dashboard as CIRED.digital is an exploratory project, not a product.

3.2 Data Ingestion and Processing

There is no consolidated, comprehensive archive of CIRED publications. Sources include:

1. The CIRED collection within HAL (Hyper Articles en Ligne), France's national open-access repository, which currently contains 1,332 full-text documents (this number will have increased by the time of writing).

2. The ISTEX repository returns 210 documents from subscription journals. This complements the HAL collection, although some paywalled articles are in fact available as preprints.

3. Activity reports list publications metadata but do not provide full text.

4. CIRED physical archives. Most have been digitized.

5. These digitized archives amount to 25 GB of historical publications from 1970–2013 (1,991 files), representing work from foundational researchers including Ignacy Sachs, Olivier Godard, and Jean-Charles Hourcade. These are not all currently indexed, but a significant fraction has been indexed [Pottier, 2024].

CIRED.digital uses the HAL collection (1238 after filtering and deduplicating). Although it is not historically complete, it is the reference open-access source and easily available through the API. Other repositories could supplement the corpus with older materials and subscription-based publications, but given the project timeline and resource constraints, these sources were deferred to future expansions of the indexed corpus.

Table 1 shows the breakdown of HAL-CIRED documents by type. We did not deduplicate by content, only by HAL ID, title, and DOI. Note that different document types have different structures and lengths, which may affect retrieval quality. For example, theses are typically much longer and more detailed than conference papers.

The data pipeline consists of two principal stages: preparation and ingestion. Preparation includes the catalog retrieval, document download, filtering, and upload steps, performed from a local machine. Ingestion is done on the RAG system, including the text extraction, chunking, vectorization and storage steps.

[2] R2R stands for RAG to Riches and is at https://github.com/SciPhi-AI/R2R

Table 1: Documents in CIRED.digital: HAL-CIRED full-text documents (before filtering and deduplication for ingestion)

Document type	Count	%
Journal articles	648	49%
Preprints / Working papers	206	15%
Conference papers	106	8%
Reports	84	6%
Theses	80	6%
Other (chapters, books, etc.)	208	16%
Total	**1,332**	100%

Preparation: The preparation pipeline includes a command-line tool (`query.py`) that queries the HAL API incrementally to maintain synchronization with newly published articles. Retrieved documents—predominantly PDF with some multimedia—are downloaded using `download.py` with pagination to avoid server overload. A metadata cleanup module (`prepare_catalog.py`) filters oversized files and performs deduplication. Documents are uploaded to the R2R server using `push.py` along with metadata (title, citation, abstract, publication date, DOI, HAL ID, document type). A verification tool (`verify.py`) confirms successful indexing.

Ingestion: Once received by R2R, documents are processed to extract text and chunked using recursive chunking: text is split at paragraph or sentence boundaries, recursively subdividing until chunks are below 512 tokens, with 50–100 token overlap. This preserves contextual information while enabling efficient retrieval of relevant passages. Chunks are represented as dense vectors through an embedding model, enabling semantic similarity search. The R2R backend persists indexed data in PostgreSQL with the pgvector extension, supporting both vector similarity and traditional database queries.

3.3 Generation and User Interface

Retrieval methods. When R2R receives a user question, it retrieves related chunks using one of three methods: (i) *Vanilla* performs vector search via cosine similarity; (ii) *RAG Fusion* combines semantic retrieval with lexical (keyword-based) search to better handle technical vocabulary and acronyms; (iii) *HyDE* (Hypothetical Document Embeddings) prompts an LLM to generate a hypothetical answer, then searches the vector store using that answer's embedding. These techniques represent well-established baselines for RAG systems. While more advanced approaches exist—semantic chunking, SPLADE, ColBERT, agentic RAG—the combination of recursive chunking with hybrid retrieval provides a robust and effective foundation for CIRED's corpus size. The choice of retrieval method is exposed to users for experimentation, defaulting to Vanilla search for performance reasons.

Language model integration. The system's generative component integrates with external LLM APIs. Commercial providers evaluated include OpenAI, Anthropic, Mistral, and Deepseek, with substantial cost variation across providers and models (see Section 5). For cost-effective deployment, we configured CIRED.digital to use Mistral Small or Medium rather than larger models. In a RAG architecture, the LLM's role shifts from knowledge recall to synthesis of retrieved content. In practice, smaller models can be less prone to elaborating beyond the retrieved context, keeping responses more faithfully grounded in CIRED publications. Recent benchmarks suggest that for RAG-based question answering, models in the 7–13B parameter range can reach 80–90% of larger model performance at a fraction of the cost [Yu et al., 2024]. Retrieval quality matters more than model size in many RAG scenarios. Temperature parameters are exposed to users, enabling control over response creativity.

Frontend architecture. The frontend is implemented as a single-page web application using vanilla JavaScript and CSS, deliberately avoiding external frameworks to minimize deployment complexity, reduce dependencies, and limit bundle size. The core interaction follows a question-answering paradigm: users formulate a natural-language query, the system generates a response grounded in indexed CIRED publications, and the interface presents the answer with explicit citations and contextual excerpts. Because response generation typically takes several seconds, the interface progressively reveals the answer as it streams from the API, maintaining user engagement during response latency.

Interface design evolution. During development, the interface evolved from an initial chat-style design toward a search-oriented interaction model. This shift reflected the intended scope of the system: Cirdi provides synthesized overviews of CIRED research on specific topics rather than conversational dialogue or open-ended assistance. The citation mechanism displays original document passages supporting each answer, with citation marks linking answer text to source documents. User testing identified interface design issues with citation rendering and revealed preferences for side-by-side presentation of answers and source material.

User experience features. Interface development emphasized convenience and accessibility. The landing page includes a word cloud generated from indexed document titles rather than a lengthy description of CIRED research themes. First-visit users receive a five-step tutorial. The user profile panel, which captures demographic information, is fully optional. The Help panel targets non-technical users discovering RAG systems for the first time. The settings panel displays server status and model parameters. We optimized citation visualization, clarified parameter meanings, and ensured LLM responses support structured outputs including tables.

3.4 Deployment and Operations

Deployment targets a single Ubuntu LTS virtual server meeting R2R recommended specifications: 3 vCPU (AMD), 4 GB RAM, and 100 GB SSD. Hetzner Cloud was selected as the primary provider, offering cost-effectiveness, reliable European data centers (Helsinki), and straightforward Docker support. European data residency supports GDPR compliance and institutional data sovereignty preferences.

Deployment automation is achieved through shell scripts (`bootstrap.sh` and `deploy.sh`) that configure the server environment, manage secrets, deploy the Docker Compose stack, and establish monitoring. Infrastructure hardening includes firewall configuration, secrets management via restricted environment files, TLS encryption for all external-facing endpoints, and automated backup and snapshot procedures.

A custom FastAPI/Uvicorn service provides monitoring, logging, and analytics complementary to R2R's core functionality. This backend captures structured event logs across six categories: session (opening/closing), request (user question), response (LLM answers with processing time), article (formatted document shown to user), feedback (satisfaction ratings and comments), and userProfile (preference changes).

The system implements a privacy-preserving design. Anonymous session identifiers are generated for each visit. User profile is only logged once and then stored client-side. Users can activate "confidentiality mode" to disable all data collection and purge any stored data. Using pure CSS and vanilla JavaScript ensures that there are no invisible cookies or embedded trackers. Collected data serves exclusively for service improvement and anomaly detection; session-level aggregation enables usage pattern analysis without individual tracking.

Table 2 summarizes the system's main components.

3.5 Implementation Reflections

Several technical challenges emerged during corpus preparation. Approximately 5–10% of PDFs required OCR fallback due to scanned-only formats—OCR processing added latency during initial index-

Table 2: Technology stack used in CIRED.digital

Component	Technology
RAG Engine	R2R (SciPhi)
Vector Database	PostgreSQL + pgvector
Backend API	FastAPI + Uvicorn
Workflow Orchestration	Hatchet
Frontend	Vanilla JavaScript + pure CSS
Web Server	Nginx
Reverse Proxy	Nginx Proxy Manager
Containerization	Docker
Language	Python 3.11+, JavaScript
LLM Providers	Mistral (default), OpenAI, Anthropic, Deepseek (development)
Data Source	HAL API
Hosting	Hetzner Cloud (Ubuntu LTS)

ing. This would be worse for historical archives. The corpus contains a majority of French or English documents, with some others like Portuguese or Vietnamese appearing. Standard English-optimized tokenizers and embeddings worked reasonably across languages, though language-specific optimization might further improve quality.

While R2R was better suited for our needs than Haystack, LlamaIndex, or LangChain at the project outset, these frameworks are rapidly evolving. Emerging frameworks such as RAGFlow or Dify may also become relevant. Future projects will need to evaluate alternatives carefully.

Containerization via Docker Compose greatly simplified continuous integration, deployment, and maintenance, enabling reproducible environments and straightforward updates. The modular architecture facilitated independent development and testing of components, enhancing maintainability while providing security through isolation that limits the attack surface of individual services. That said, the digital librarian service remains simple—everything fits on a single node. For a small-scale deployment running on a dedicated server where collateral damage is limited, scripts and virtual environments would have sufficed without Docker.

R2R's modular architecture proved crucial for this project. The separation of retrieval and generation components enabled efficient model provider switching and cost optimization during testing. Supporting multiple LLM providers (Mistral, OpenAI, Deepseek) is a given in this kind of application. Using OpenRouter would have simplified administration, billing one company to access all models. Costs and quality differences underscore the need for periodic provider re-evaluation.

PostgreSQL with the pgvector extension proved reliable and cost-effective for the corpus size of approximately 1,300 documents. For larger corpora (100K+ documents), dedicated vector databases might offer performance advantages, though these claims would need verification. Default embeddings worked adequately for CIRED's corpus; specialized domain-specific embeddings might improve retrieval quality but at increased latency and cost.

Future work could explore advanced retrieval techniques such as semantic chunking, SPLADE, or ColBERT to enhance relevance. Experimentation with different chunk sizes and overlap parameters may also yield improvements. On the generation side, fine-tuning smaller models on CIRED-specific data could enhance answer quality while controlling costs. Overall, the technical implementation of CIRED.digital demonstrates the feasibility of deploying a RAG-based digital librarian service using open-source components and containerization, providing a solid foundation for future enhancements and expansions.

4 Usage Analysis

This section presents user engagement with CIRED.digital during the public beta phase (July 5–October 9, 2025). The following analysis draws on 259 sessions generating 1,849 events over 96 days. Generated answers contain 6.1 citations on average (interquartile range 3–8 citations), referring to 2.8 distinct publications (interquartile range 1–4 publications).

Methodological note. The beta-phase dataset is limited in size and composition and does not support formal statistical inference. The analyses below therefore provide directional and design-relevant insights, not population-level estimates. Visualizations and statistics for subsections 4.1 and 4.2 are drawn from the analytics scripts available in the project's codebase. Visualizations and statistics for the query content analysis subsection 4.3 were conducted by an AI assistant (Claude) using rule-based classification and keyword extraction methods. The system captured 290 total query submissions representing 159 unique queries (some queries were repeated across different sessions). These queries primarily reflect early adopters, invited testers, and CIRED network users rather than representative public usage. Machine-based analysis introduces potential misclassification compared to manual expert coding.

4.1 Deployment Timeline, Activity, and User Demographics

The CIRED.digital system underwent distinct deployment phases. During the Alpha phase (April–June 2025), internal development and closed testing occurred with a small convenience sample of testers. The Beta Closed phase (June–early July 2025) expanded to invited testers from all CIRED. The Beta Open phase (July 5–October 9, 2025) provided public unrestricted access following institutional announcements on the CIRED institutional newsletter and on social media. Peak activity of approximately 30 new sessions occurred on July 10, 2025, following public launch announcement, with gradual decline consistent with typical digital product adoption curves. The beta period concluded on October 9, 2025, after 96 days of public availability. The system remains in steady operation with ongoing monitoring, but pending the final project's review we did not launch further promotional activities like integration into CIRED website, the HAL collection page, CIRED homepage in partners institutions, or social media campaigns.

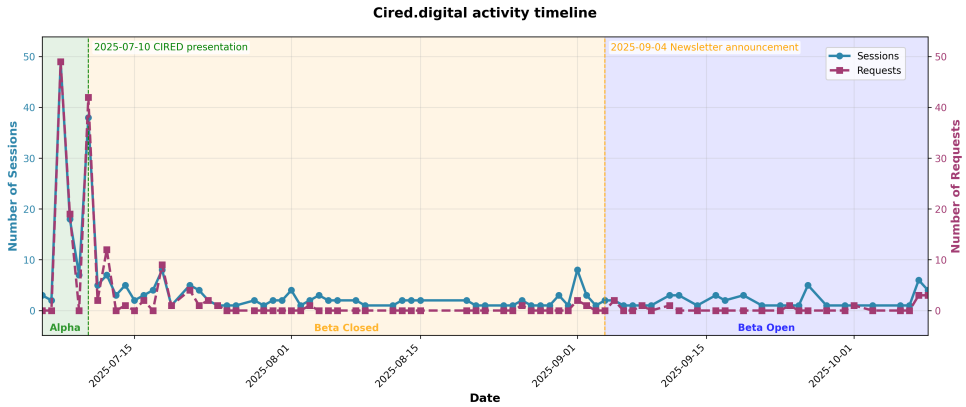


Figure 2: Session activity timeline over the beta period (July 5–October 9, 2025). The chart shows cumulative session growth with peak activity following the public launch announcement on July 10, 2025.

Network traffic analysis classified user provenance during the beta phase. CIRED network users represent 18% of sessions, with substantial engagement from RENATER-affiliated researchers (11 %) and French domestic visitors (33 %). Identified bots (e.g. Google bot) account for 25% of sessions, reflecting typical web crawling activity during public deployments. This distribution indicates successful reach both within CIRED's institutional network and to the broader research community.

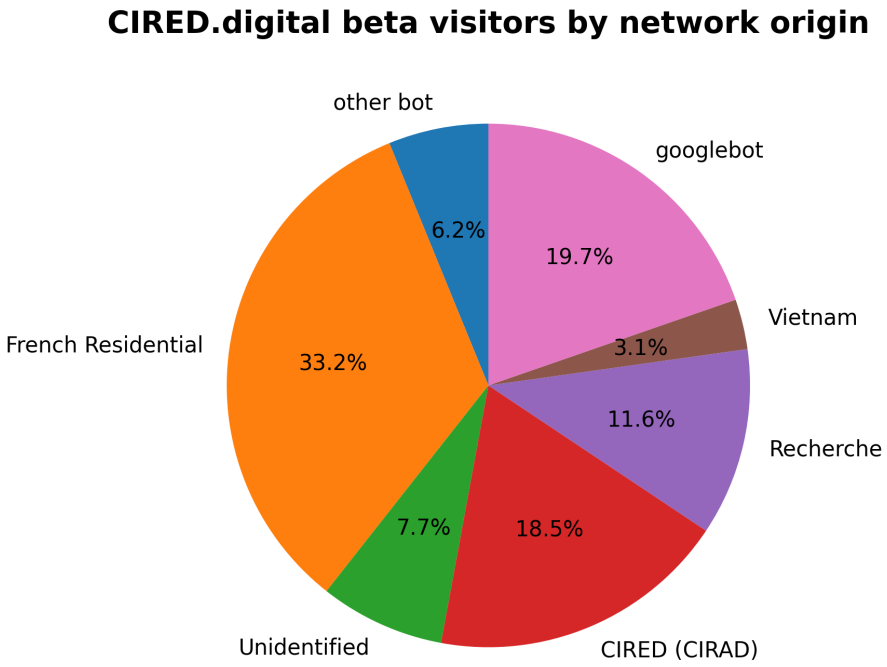


Figure 3: User provenance by network origin.

Data collection identified a significant gap: user profile fields designed to capture profession, expertise domain, and institutional affiliation were very rarely populated. This represents a critical limitation for demographic segmentation analysis. The question of RGPD-compliance in future deployments through optional, explicitly framed profile collection mechanisms remains open.

4.2 Session Patterns and Query Analysis

Figure 4 maps the complete user journey through CIRED.digital, visualizing all 1,849 event transitions across nine event types during the 96-day beta period. The diagram reveals three distinct user behavior patterns that account for all 259 sessions.

Bounced visitors (152 sessions, 59%) exited immediately without submitting queries, evidenced by the strong Start⟶End direct transition. This high bounce rate likely reflects a combination of (i) bots, particularly the google bot (ii) exploratory visits, users assessing whether the tool meets their needs (iii) potential first-load technical issues or unclear value proposition. While elevated, this rate is not unusual for beta-phase research tools deployed without marketing campaigns.

Engaged explorers (approximately 60 sessions, 23%) proceed directly from arrival to query submission (Start⟶request), then through the core interaction loop: request⟶response⟶article examination. The 60 direct Start⟶request transitions represent users who immediately began querying the system upon arrival. These sessions generated 188 request events total across the beta period, with many users submitting multiple queries as evidenced by the 19 request⟶request self-transitions and various request⟶response⟶request cycles. It suggests that users actively engaged with source materials rather than treating AI-generated answers as definitive statements.

Cirdi logs show 129 article view events. The 52 response⟶article transitions are supposed to be the normal workflow. The absence of an Article event after a Response event can be explained by (i) lost logging events, since the system is not fully reliable in capturing every interaction (ii) users leaving during the ”generation” phase. Some users may have left Cirdi while the LLM response was rendering, leading to missing article view events.

Interface explorers (47 sessions, 18%) engaged primarily with system features rather than content queries. The visibility toggle events—593 “hidden” and 450 “visible” for a total of 1,043 events—dominate

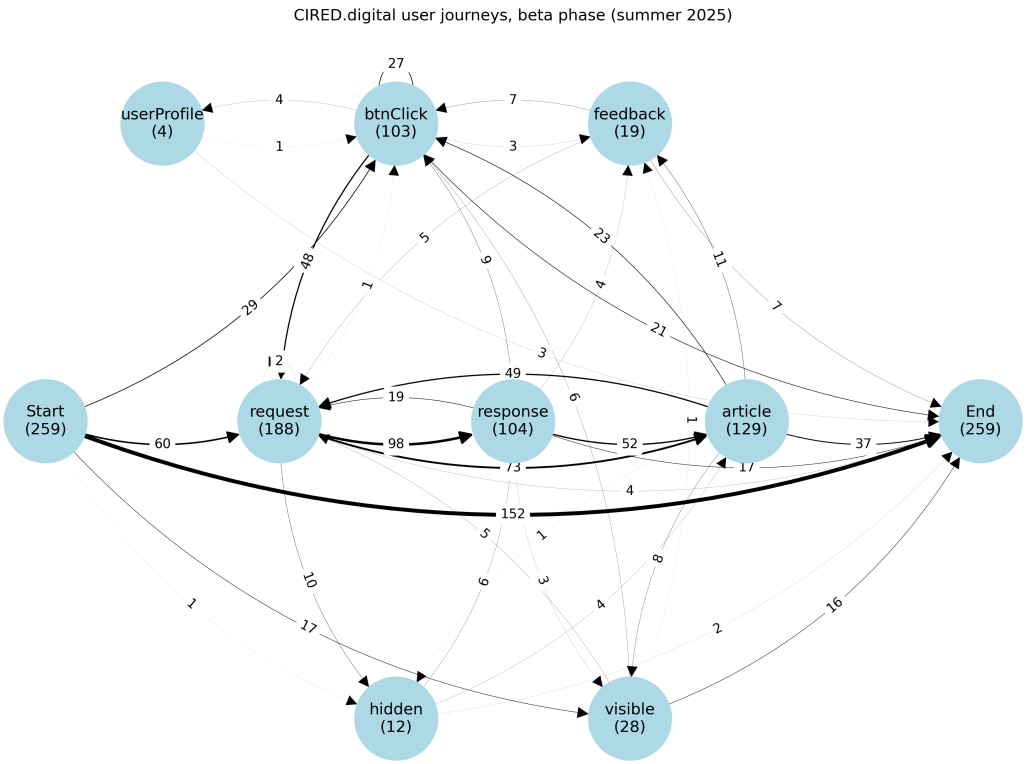


Figure 4: User journey event transitions during the 96-day beta period. Nodes represent event types with occurrence counts in parentheses; edges show transition frequencies with thicker lines indicating dominant pathways. The core interaction loop (request→response→article) accounts for the majority of engaged sessions, while direct Start→End transitions (152) reflect bounce rates.

the interaction log numerically and indicate active exploration of help panels, settings, and citation views. These users may have been evaluating the system's capabilities or familiarizing themselves with the interface before formulating queries. The high frequency of visibility toggles suggests users found the interface elements worthy of investigation, though it also raises questions about whether the default interface state optimally balances information density and clarity (see Section 3.3 for interface design rationale).

Iterative refinement patterns appear clearly in the diagram. The request↔response loops (19 request→response→request cycles, plus self-transitions) confirm that 70% of sessions contained multiple queries. Users frequently reformulated questions or explored related topics within single sessions, suggesting the search-oriented interface successfully supported exploratory research workflows. The median session depth of 3–7 interactions aligns with these multi-query patterns.

Feedback provision remained sparse: only 19 feedback events out of 107 engaged sessions (excluding the 152 bounced visitors), yielding a 18% feedback rate. Combined with the 7 feedback→btnClick and 11 feedback→article transitions visible in Figure 4, this indicates that users who provided feedback often continued engaging with the system rather than immediately leaving. The low-friction thumbs-up/down design achieved reasonable adoption without requiring extensive user effort. Though the absolute numbers remain too small for detailed sentiment analysis, the 18% feedback rate aligns with typical internet user behavior patterns.

User profile capture proved nearly absent, with only 4 userProfile events recorded despite the optional profile interface. This confirms the data collection gap identified earlier and underscores the importance of collecting demographic information during session initialization rather than through optional user action. The 1 userProfile→btnClick transition suggests that even when users did engage with profile settings, they returned to core system functionality afterward.

Figure 5 presents the complete event transition diagram including all low-frequency paths and technical events. Nodes have been relabelled, e.g. visibilityOn/visibilityOff shown separately. While more complex, this comprehensive view reveals additional patterns: users frequently toggled between visibility states (371 visibilityOff→visibilityOn and 422 reverse transitions, suggesting active management of interface complexity. The diagram also shows that many paths lead directly to End from various states, indicating that users exit at multiple points in their journey rather than following a single canonical termination path.

Several findings emerge from this journey analysis with implications for future development:

1. **High bounce rate requires attention**. The 59% immediate exit rate suggests that there are more bot visitors than human users. Or the landing page may not sufficiently communicate value proposition or that technical barriers prevent initial engagement. Priority improvements include clearer calls-to-action, example queries prominently displayed, and faster initial page load.

2. **Multi-query sessions dominate engaged use**. The prevalence of request→response→request cycles confirms that the system successfully supports iterative exploration. However, the lack of conversational context suggests that supporting multi-turn queries could substantially improve user experience.

3. **Interface complexity may deter some users**. The 1,043 visibility toggle events indicate that interface elements require active user management. Simplifying default states while preserving power-user customization options could reduce cognitive load for first-time visitors. In the end, it is unsure if the "onboarding" panel was helpful or distracting, and the user profile collection panel was not necessary.

The journey analysis establishes that CIRED.digital successfully serves the approximately 23% of visitors who engage substantively with the system, providing them with cited, verifiable answers and supporting multi-query exploratory workflows. Further research would be needed to understand the high bounce rate (which are not from bots ?) and optimize the interface for first-time users.

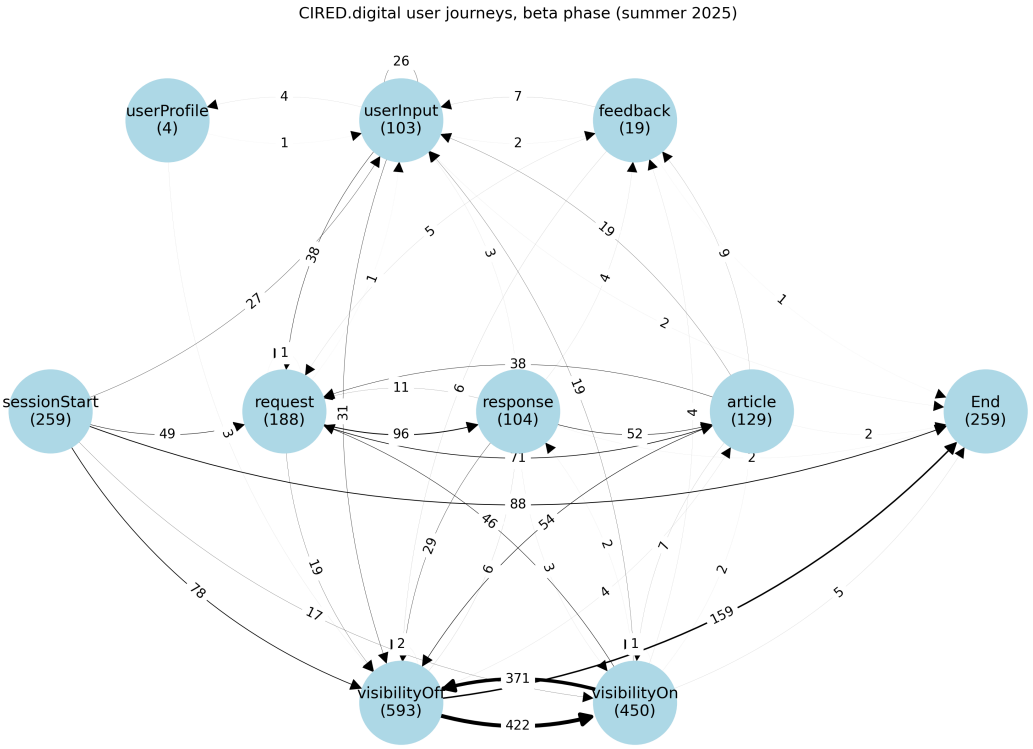


Figure 5: Complete user journey event transitions showing all recorded interaction paths. This comprehensive view includes visibility state transitions (visibilityOn/visibilityOff) shown separately, revealing the full complexity of user interface exploration. While the core request→response→article loop remains dominant, this diagram illustrates the diverse pathways users took through the system and the prevalence of interface customization behaviors.

4.3 Query Content and User Intent

4.3.1 Dataset Characteristics

The system captured 290 total query submissions representing 159 unique queries (some queries were repeated across different sessions). The frequency distribution was heavily skewed: the single most common query ("Le CIRED") appeared 6 times, while 123 queries (77%) appeared exactly once. This distribution is typical of exploratory beta testing where users probe system capabilities with diverse questions rather than repeated information needs.

Query language distribution, weighted by submission frequency, shows near-parity between French (50.5%, 95 instances) and English (46.3%, 87 instances), with minimal Arabic (3.2%, 6 instances). The balanced French-English split validates the multilingual deployment approach, while the small but non-zero Arabic usage (2 unique queries submitted 3 times each) suggests focused interest from Arabic-speaking users—content analysis indicates these queries concerned North African energy transitions, aligning with CIRED research themes.

Figure 6 presents a comprehensive overview of query patterns across four dimensions: language distribution, query type classification, complexity by word count, and dominant research themes.

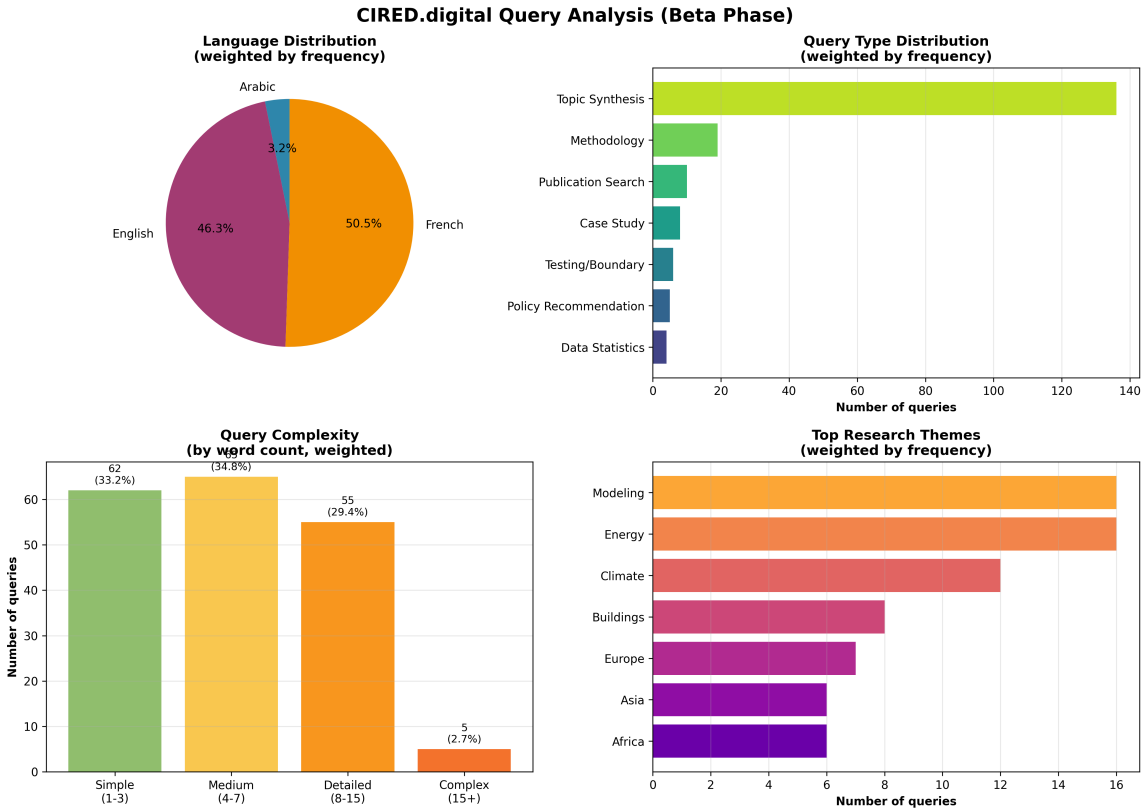


Figure 6: Query analysis overview showing: (a) language distribution weighted by frequency demonstrating near-parity between French and English, (b) query type distribution with topic synthesis dominating at 72%, (c) complexity distribution by word count showing most queries are simple to medium length, (d) top research themes detected through keyword matching with energy, climate, and modeling as leading topics.

4.3.2 Query Type Classification

Automated content analysis classified queries into seven intent categories using rule-based pattern matching (see methodological details in project documentation). The distribution weighted by query frequency reveals clear patterns in user information-seeking behavior.

Topic synthesis queries (135 instances, 71.8%) dominated the dataset, encompassing broad overview questions ("What are the infrastructure needs in future climate mitigation scenarios?"), simple topic labels ("Le changement climatique"), and concept definitions ("payments for environmental services"). This pattern indicates users primarily sought accessible entry points to CIRED research rather than detailed technical documentation. The prevalence of synthesis-oriented queries validates the decision to position Cirdi as a "scientific documentalist" providing grounded summaries rather than attempting conversational AI interactions.

Representative examples of topic synthesis queries include:

6× Le CIRED

4× فرنسا في الطاقي التحول (Energy transition in France)

4× What are the infrastructure needs in future climate mitigation scenarios?

3× payments for environmental services

3× Le changement climatique

Methodology queries (19 instances, 10.1%) focused on CIRED's modeling frameworks, particularly IMACLIM (integrated assessment model), Res-IRF (residential energy model), and computable general equilibrium approaches. Representative examples include: "qu'est ce que le modèle res-IRF ?" (asked 2 times), "What can you say about the ability of IMACLIM to provide insights about the real world", and simply "Cge" (asked 2 times). These queries suggest the system successfully served researchers and students seeking to understand CIRED's analytical approaches, though the technical jargon (model acronyms without context) indicates users already possessed familiarity with the laboratory's work.

Publication searches (11 instances, 5.9%) were less common than anticipated during project design. Queries included explicit author-focused requests ("liste des articles de Louis-Gaëtan Giraudet", asked 2 times; "Articles de F. Ghersi") and topic-specific publication queries ("quelles sont les publications sur la climatisation ?"). The relatively low proportion suggests HAL's native interface already adequately serves publication discovery needs, or that users did not perceive Cirdi as optimized for this function.

Case study queries (8 instances, 4.3%) exhibited strong geographic specificity: "vietnam renewable", "vietnam energy efficiency", "La tunisie", "La transition énergétique en france et son effet sur l'emploi". This pattern reflects CIRED's traditional North-South research focus spanning Asia (Vietnam, China), Europe (France), and Africa (Tunisia, Senegal).

Testing and boundary-probing queries (6 instances, 3.2%) revealed deliberate exploration of system limits. The pattern "raconte-moi une histoire sur…" ("tell me a story about…") appeared 3 times with different topics (Tunisia, energy renovation), suggesting multiple users tested conversational capabilities. These queries received appropriate refusals, confirming that expectation management mechanisms functioned as designed.

Policy recommendation and **data/statistics requests** together comprised 4.8% of queries (5 and 4 instances respectively). Notably, one user submitted a 509-word query (3,913 characters) analyzing Hanoi's announced ban on diesel and petrol motorcycles, requesting infrastructure transition timelines and supporting mechanisms. This outlier demonstrates both the system's technical capacity to process extended context and the need for clearer interface guidance on optimal query formulation. Data requests included "Table des centrales électriques en tunisie ?" (asked 2 times) and "How much would the development of data centers put pressure on Vietnam's power system ?"

4.3.3 Query Complexity and Interaction Style

Query length analysis reveals that users adopted search-style rather than conversational interaction patterns. The median query length was 5 words (mean: 10.1 words), with 68% of queries containing 7 or fewer words. Figure 6(c) shows the complexity distribution: simple 1–3 words (33%), medium 4–7

words (35%), detailed 8–15 words (29%), and complex 15+ words (3%), excluding one extreme outlier of 509 words.

Only 17% of queries included explicit question marks, indicating most users formulated queries as keyword phrases or topic labels rather than natural-language questions. This behavior aligns with the search-oriented interface design, where users treated Cirdi as a semantic search engine rather than a question-answering chatbot. The prevalence of short, declarative queries validates the decision to implement a search paradigm rather than maintaining conversational context across queries.

4.3.4 Thematic Content

Keyword extraction and pattern matching identified dominant research themes, as shown in Figure 6(d). Queries may match multiple themes; percentages indicate the proportion of queries containing theme-related keywords:

- **Modeling approaches** (9.6% of queries): IMACLIM, Res-IRF, CGE models, scenario analysis
- **Energy systems** (9.0%): Renewable energy, electricity grids, smart grids, energy efficiency
- **Climate change** (9.0%): Climate policy, emissions, carbon pricing, mitigation scenarios
- **Building sector** (4.3%): Residential energy renovation, housing, thermal efficiency
- **Geographic contexts**: Asia (3.7%, primarily Vietnam and Hanoi), Europe (3.7%, primarily France), Africa (3.2%, Tunisia and Senegal)

The most frequent keywords, weighted by query repetition, were: "cired" (24 instances), "energy" (15), "grid" (15), "transition" (14), "smart" (13), "infrastructure" (11), and "impact" (10). The prevalence of "CIRED" suggests users explicitly framed queries around institutional knowledge, while energy transition terminology (energy, grid, transition, smart, infrastructure) dominated topical content. Geographic specificity appeared frequently, with "france" (7 instances), "tunisie" (6), "hanoi" (6), and "vietnam" (multiple mentions) indicating strong interest in country-level analyses aligned with CIRED's research portfolio.

4.4 Conclusion

Usage analyses presented in this chapter indicate that CIRED.digital delivers research-grade responses aligned with its intended role as a search-oriented scientific documentalist: Cirdi returns on average 3 publications per query.

The query analysis validates several core design decisions while revealing opportunities for enhancement:

1. **Search-oriented interface validated**. The 5-word median query length and 83% non-interrogative phrasing confirm users treated Cirdi as semantic search rather than conversational AI. Future iterations should preserve this interaction paradigm while enhancing rather than complicating the interface.
2. **Topic synthesis prioritization confirmed**. With 72% of queries seeking overviews and accessible entry points, the system correctly emphasizes grounded summarization over technical deep dives. Citation mechanisms remain essential to enable verification while maintaining accessibility.
3. **Multilingual deployment justified**. Near-parity between French (50%) and English (46%) validates investment in multilingual query handling, while minimal Arabic usage (3%) suggests either limited awareness among Arabic-speaking audiences or insufficient corpus coverage of Arabic-language research contexts.

4. **Multi-turn context emerges as priority enhancement**. Although the current analysis cannot track query sequences within sessions (session identifiers were not linked to query logs), the journey analysis (Section 4.2) showed 70% of sessions contained multiple queries. Combined with negative feedback citing lack of conversational memory (30% of complaints), this indicates that supporting follow-up questions while maintaining the search paradigm could substantially improve user experience. Example sequences inferred from query content suggest patterns like: "IMACLIM" → "capital accumulation in IMACLIM" → "IMACLIM real-world validity".

5. **Structured output generation**. Data/statistics queries (2%) and some publication searches (6%) requested tabular formats—requests the current system appropriately declines rather than generating unreliable structures. Future work could explore structured output generation with appropriate quality controls for author bibliographies, model comparison tables, or country-level energy statistics.

6. **Publication discovery integration**. The 6% publication search proportion, lower than anticipated, suggests either that HAL adequately serves this need or that users did not perceive Cirdi as optimized for publication discovery. Enhanced integration with HAL metadata and author-topic linking could strengthen this functionality.

Usage data highlights clear improvement priorities. First, the absence of conversational memory constrained follow-up questions in multi-query sessions, despite evidence that engaged users iteratively refined their queries. Second, the stealth mode during the beta phase limited overall traffic, with automated agents accounting for approximately one quarter of visits. Third, interface complexity—as revealed by extensive visibility toggling—suggests that default presentation choices could be simplified further: the user profile dialogs appear unnecessary, and for audiences already familiar with RAG systems the onboarding panel may be superfluous. Finally, instrumentation could be strengthened: the absence of systematic latency logging limited retrospective performance assessment.

Beyond these refinements, the analysis provides directional evidence that CIRED.digital successfully attracted substantive research queries aligned with CIRED's institutional expertise, served diverse user intents through a search-oriented interface, and operated effectively across multiple languages. The observed usage patterns inform both immediate system improvements (notably multi-turn context handling and structured outputs) and longer-term strategic questions regarding the positioning of Cirdi within CIRED's broader knowledge dissemination ecosystem.

Overall, the evidence supports the conclusion that CIRED.digital meets its core objective: providing transparent, citation-grounded access to CIRED publications for exploratory research and knowledge discovery, while generating concrete, actionable insights to guide future system evolution.

5 Costs

The CIRED.digital project demonstrates that institutional RAG deployment remains financially accessible to research laboratories operating under realistic budget constraints. This section presents comprehensive cost analysis spanning development investment, operational expenditures, and long-term sustainability considerations. Drawing on empirical data from the 96-day beta deployment and comparative analysis of similar projects, we establish that production-ready conversational access to scientific publications can be achieved for €40,000–60,000 in development costs and €5,000–10,000 in annual operational costs—figures well within reach of most research institutions. The project was a personnel investment rather than technology procurement.

5.1 Development Costs

5.1.1 Cost Estimation Methodology

Estimating software development costs retrospectively presents methodological challenges, particularly for research projects where effort is distributed across multiple contributors with varying time commitments. We employed three complementary estimation approaches to triangulate realistic cost ranges, each drawing on different evidence bases and accounting frameworks.

The **commit-based analysis** uses version control activity as a proxy for development effort. GitHub records 210 commits to the main branch over the project timeline. Assuming each commit represents an average of 2.5 hours of work—including coding, testing, debugging, and documentation—yields an estimate of 525 total development hours. This multiplier reflects established software engineering practice: commits represent integrated, tested changes rather than raw coding time, and include associated quality assurance activities.

The **component-based breakdown** itemizes development effort by functional area: RAG backend implementation (120 hours), frontend development (80 hours), data ingestion pipeline (60 hours), and so forth. This bottom-up accounting, informed by project logs and team member estimates, totals 550 hours across ten major work categories. The close agreement with the commit-based estimate (525 vs 550 hours, within 5%) strengthens confidence in both approaches.

The **lines-of-code (LOC) method** applies industry productivity benchmarks to the project codebase. Repository analysis identifies approximately 14,500 lines of code across Python (44.8%), JavaScript (26.5%), and other languages. Using a conservative 10 LOC per hour productivity rate—appropriate for research-quality code with comprehensive testing and documentation—yields 1,450 hours. This substantially higher estimate likely reflects the comprehensive scope of research projects, which include exploratory work, literature review, architectural design, and methodological refinement not captured in commits or component breakdowns.

All three methods rest on documented assumptions. We assume that commit granularity reflects typical development practices (commits every 2–4 hours of substantive work), that component estimates capture both direct implementation and associated overhead, and that productivity rates align with academic rather than commercial software development contexts. The French academic salary scale provides baseline hourly rates: junior developers (€25/hour), mid-level developers (€45/hour), senior developers (€70/hour), and research engineers (€60/hour).

Confidence levels vary across methods. The commit-based and component-based estimates, showing near-identical results, achieve high confidence (±15%). The LOC-based estimate, while methodologically sound, likely overestimates by including research activities peripheral to code production; we assign it medium confidence (±30%). Overall, we assess medium-high confidence (±20%) for the recommended range, reflecting convergent evidence from multiple methodologies.

These estimates carry inherent limitations. We cannot access detailed time logs from all six contributors, making precise effort allocation impossible. Individual learning curves—particularly for contributors new to RAG systems or containerized deployment—are not separately quantified. The repository creation date may not precisely align with project initiation, introducing boundary ambiguity. Despite

these limitations, the methodological triangulation provides robust evidence for cost range estimation.

Table 3: Development cost estimates using three complementary methodologies. The LOC-based estimate intentionally provides a conservative upper bound, as it includes exploratory research, architectural design, and learning activities beyond direct code production.

Method	Hours	Cost (€60/h)	Confidence
Commit-based	525	€31,500	High
Component-based	550	€33,000	High
LOC-based	1,450	€87,000	Medium
Recommended range	**667–1,000**	**€40,000–60,000**	**Medium-High**

5.1.2 Development Cost Analysis

The three estimation methodologies converge on €40,000–60,000 development cost including labor, with €50,000 mid-range estimate. Table 3 summarizes estimates and confidence levels. The conservative estimate of €33,000 assumes 550 hours at €60/hour, while the €87,000 upper bound reflects comprehensive research activities including literature review and architectural exploration.

This range acknowledges that research software development encompasses requirements analysis, technology evaluation, stakeholder consultation, iterative testing, and comprehensive documentation beyond direct coding. Cost projections vary with personnel: junior developers (€25/hour) yield €13,125 for 525 hours, senior developers (€70/hour) cost €36,750, and freelance consultants (€100/hour) charge €52,500.

Institutions with skilled research engineers or motivated graduate students can achieve outcomes at the lower end. Those requiring external contractors should budget toward the upper range. Mature open-source frameworks (R2R, LlamaIndex, LangChain) substantially reduce development time, making €40,000–60,000 realistic for production-ready institutional deployments.

According to the French Direction Générale des Entreprises' RAG adoption guide [Direction Générale des Entreprises (DGE), 2024], integrator-based implementations typically cost €100,000, while internal development requires at least one data scientist, developer, and system administrator for 3+ months. SaaS solutions range €10,000–100,000 annually. Infrastructure costs include GPU investment (€3,250 for Nvidia RTX 4090, €0.10/1,000 queries electricity) for on-premise, or cloud hosting at €7.80/1,000 queries (€106/employee/year at 300 queries/week).

CIRED.digital's €40,000–60,000 development cost aligns with the lower end of custom implementations through strategic choices: leveraging open-source R2R, focused scope (corpus access vs. general-purpose assistant), and research mission alignment (publishable outputs beyond operational value). Operational costs (€50–200 annually for infrastructure/API, €5,000–10,000 including maintenance) position the system as cost-effective while delivering production-ready infrastructure, comprehensive documentation, empirical evidence, and reusable open-source components.

5.2 Operational Costs

5.2.1 Infrastructure Costs

CIRED.digital operates on a domain registered through Gandi.net at €30/year. The core infrastructure cost is virtual private server (VPS) hosting for the RAG system backend, database, and web application. The service could be transferred to an institutional domain to reduce costs and improve governance. Alternatively the domain could become the entry point for CIRED's broader digital presence.

CIRED.digital operates on a single virtual private server (VPS) meeting the R2R framework's recommended specifications: 2–4 vCPU (AMD architecture), 4–8 GB RAM, and 50–100 GB SSD storage

running Ubuntu LTS. This modest configuration suffices for a corpus of approximately 1,300 documents far exceeding actual beta deployment traffic.

Provider selection prioritized cost-effectiveness, European data residency, and developer reliability. We selected Hetzner Cloud (Helsinki) for its competitive pricing (€10/month for 2 vCPU, 4 GB RAM, 80 GB SSD), straightforward administration, and robust performance. Table 4 compares European alternatives evaluated for the project. OVHcloud offers comparable pricing (€9.22/month) with broader geographic distribution across France, Germany, Poland, and the UK, though with marginally more complex administration interfaces. Scaleway (France) and DigitalOcean (Amsterdam/Frankfurt) charge €15–18/month for equivalent configurations—affordable but 50–80% more expensive than Hetzner. Infomaniak (Switzerland) represents the premium option at €11.50/month, distinguished by explicit 100% renewable energy sourcing and Swiss data protection frameworks, though at modest cost premium without proportionate functionality gains for this use case.

Table 4: European VPS hosting options evaluated (2–4 vCPU, 4–8 GB RAM)

Provider	Location	Monthly Cost	Annual Cost
Hetzner Cloud	Helsinki	€10.00	€120
OVHcloud	France/Germany	€9.22	€111
Scaleway	France	€15.00	€180
DigitalOcean	Amsterdam	€18.00	€216
Infomaniak	Switzerland	€11.50	€138

Projecting to annual sustained operation, infrastructure costs range €120–210/year depending on configuration choices and redundancy requirements. The baseline €120 reflects continuous operation on the current single-VPS deployment. Enhanced reliability through redundant servers or load-balanced configurations would add €50–90/year, but these are beyond the need of a small RAG. Monitoring and logging services (Prometheus, Grafana, or commercial alternatives) cost €0–50/year depending on provider and feature requirements, but these services should be sourced for the whole information system, not just the RAG component. For a production research service, we recommend budgeting €140–180/year (€120 baseline + €20–60 backup/monitoring) as a realistic steady-state infrastructure cost.

Infrastructure costs exhibit minimal scaling with usage over substantial ranges. The current configuration handles 259 sessions and 290 queries with negligible load—estimated average utilization remains below 20%. This headroom accommodates 10× traffic growth (2,500–3,000 sessions, 2,000–3,000 queries annually) without infrastructure upgrades. Only at 10,000+ annual sessions would database optimization, caching layers, or additional VPS capacity become necessary. This cost profile—fixed infrastructure adequate for broad usage ranges—makes institutional RAG deployment economically favorable compared to per-query commercial services.

5.2.2 LLM API Costs

Large language model APIs constitute the variable cost component in RAG system operation, scaling directly with usage volume and model selection. Commercial LLM providers price services by token consumption, with dramatic variation: 20–40× differences between economical and expensive options. Table 5 compares major providers evaluated during development. Mistral AI's Small model (€0.14/M tokens) offers the lowest cost among production-ready options. Mistral Medium (€0.24/M) and DeepSeek V3.1-Terminus (€0.27/M) provide mid-range alternatives, while OpenAI's GPT-5.2 costs approximately 21× more than Mistral Small under typical RAG usage patterns (90% input, 10% output tokens).

Cost does not scale linearly with quality in RAG applications. Empirical studies show that while larger models achieve higher absolute performance, smaller models demonstrate substantial relative improvement from retrieval augmentation, with RAG helping smaller models more than larger ones [Soudani et al., 2024]. In RAG architectures, the generator's role shifts from knowledge recall—where

Table 5: LLM API pricing comparison (per 1M tokens, December 2025)

Provider	Model	Cost per 1M tokens	Relative Cost
Mistral AI	Small	€0.14	1.0×
Mistral AI	Medium	€0.24	1.7×
DeepSeek	V3.1-Terminus	€0.27	1.9×
OpenAI	GPT-5.2	€1.75 in / 14.00 out	21×*

*Relative cost calculated for typical RAG usage (90% input, 10% output tokens = €2.98/M blended). Pure input: 12.5×; pure output: 100×.

larger models excel—to synthesis of retrieved passages. This architectural change narrows the performance gap between model sizes.

Systematic testing during CIRED.digital development confirmed that Mistral Small (€0.14/M tokens) achieved adequate answer quality for institutional question-answering at approximately 5% of GPT-4's cost. For corpus-grounded applications with high-quality retrieval, the cost-quality trade-off favors economical models: retrieval quality and chunking strategy matter more than generator sophistication for factual grounding. This makes smaller models strategically appropriate for production institutional RAG systems operating under realistic budget constraints.

Table 6: Measured LLM API usage and costs across all providers

Provider	Period	Phase	Tokens	Cost (EUR)	Days
Development phase (OpenAI only)					
OpenAI	June 2025	Development	20,617,818	–	30
OpenAI	July 1–4, 2025	Development	240,020	–	4
Beta deployment phase (Mistral)					
Mistral	July 5–31, 2025	Beta	1,115,310	0.20	27
Mistral	August 2025	Beta	6,938	0.00	31
Mistral	September 2025	Beta	27,469	0.00	30
Mistral	October 1–9, 2025	Beta	8,587	0.00	9
Development total (OpenAI)			**20,857,838**	**–**	**34**
Beta total (Mistral)			**1,158,304**	**0.20**	**97**
Complete project total			**22,016,142**	**0.20+**	**131**

CIRED.digital deployed Mistral Small as the primary generation backend. The 96-day beta deployment processed 290 queries consuming 1,158,304 tokens (Table 6)—approximately 3,994 tokens per query including both input context and generated output. At Mistral Small's €0.14 per million tokens pricing, total API cost was €0.20 (€0.0007 per query), negligible compared to infrastructure costs. Peak usage occurred in early July following launch (Figures 7 and 2), declining to near-zero in subsequent months as usage stabilized.

Development phase testing (June-early July 2025) consumed 20.9 million tokens on OpenAI models for prototyping and comparison testing. Production deployment transitioned to Mistral Small for the 96-day beta, consuming only 1.2 million tokens at €0.20 total cost.

Figure 7 displays daily token consumption patterns in July, with dark blue representing input tokens (query text plus retrieved document context) and light blue representing output tokens (generated answers). The stacked bar visualization reveals that input tokens dominate output tokens—because RAG systems provide multiple document passages to ground each response. Peak activity on July 2–3 exceeded 200,000 tokens per day, driven by concentrated testing and exploration following the public announcement. Subsequent weeks show moderated usage around 50,000–100,000 tokens per day, declining toward month end consistent with typical user adoption curves.

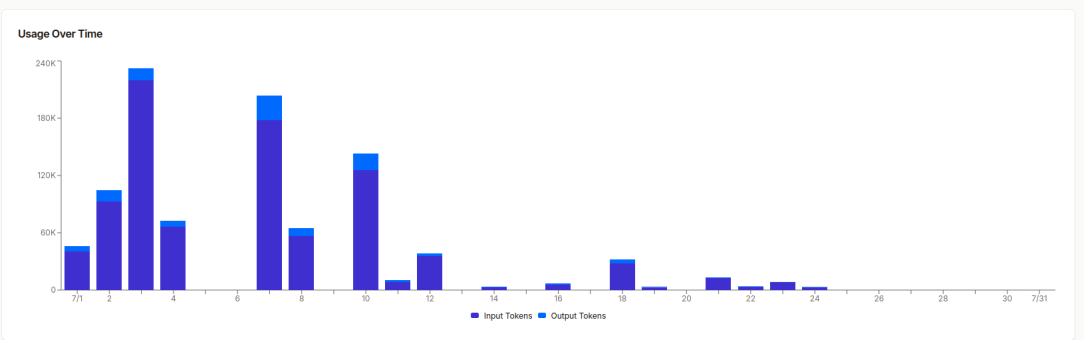


Figure 7: Daily token consumption during peak usage month (July 2025). Dark blue represents input tokens (context), light blue represents output tokens (generated responses). Peak activity occurred July 2–3 following public launch announcement.

5.2.3 Total Cost of Ownership

Total cost of ownership (TCO) integrates development investment, operational expenditures, and ongoing maintenance. Table 7 summarizes costs across the beta period and projects annual steady-state operation.

Table 7: Total cost of ownership for CIRED.digital deployment, including labor costs

Cost Category	Beta Period (96 days)	Annual Projection
Infrastructure (VPS)	€32	€120–210
Domain name	€30	€90
LLM API	€ 0.20	€50–300
Maintenance (labor)	–	€5,000–10,000
Operational Total	€35–37	€5,260–10,600
Development (one-time)	–	€40,000–60,000

The 96-day beta deployment incurred €33.20 in direct operational costs: €33 infrastructure plus €0.20 API charges. This supported 259 user sessions and 290 queries, yielding a unit cost of €0.13 per session for operational expenses only, or €155-232 per session including development amortization.

First-year TCO ranges €45,170–70,510, incorporating both one-time development (€40,000–60,000) and recurring operational costs (€5,170–10,510). In subsequent years, costs decline to €5,170–10,510 annually as development costs amortize. Labor costs dominate at 90–95% of first-year TCO and 95–98% of ongoing annual costs, indicating that institutional RAG deployment constitutes primarily a labor investment rather than technology procurement.

Unit cost economics demonstrate favorable scaling properties. At beta-period utilization (259 sessions), total first-year cost per session reached €154–232. However, at 1,000 annual sessions, this drops to €45–71 per session, and at 5,000 annual sessions to €9–14 per session. This scaling behavior—declining unit costs with increasing utilization—reflects the high fixed-cost, low variable-cost structure: development represents a one-time investment whose per-unit cost decreases as usage grows, infrastructure costs remain essentially flat across wide usage ranges (100 to 10,000+ sessions), and only maintenance effort increases sub-linearly with usage.

The TCO analysis establishes that institutional RAG deployment remains financially viable for research laboratories. First-year investment of €45,000–70,000 positions the project within typical research infrastructure budgets. Ongoing annual costs of €5,000–10,000 represent manageable commitments comparable to journal subscriptions. Unit cost scaling ensures that increasing adoption improves economic viability.

5.3 Lessons Learned and Discussion

5.3.1 Implications of the cost structure

The dominant role of labor costs—95%+ of total expenditure—carries strategic implications. Technology costs (infrastructure, API) remain modest and predictable, while human effort (development, maintenance, support) drives financial requirements. This distribution suggests that cost optimization should emphasize developer productivity (leverage frameworks, avoid reinventing components, prioritize maintainability) over marginal infrastructure savings. Investing effort to reduce API costs from €200 to €100 annually yields negligible benefit compared to reducing development time from 1,000 to 800 hours (saving €12,000–14,000). Cost-conscious projects should thus focus on technical efficiency and scope discipline rather than provider negotiation.

For API provider costs, we observed 20× price variation between a small-but-adequate and a large best-in-class LLM. The provider landscape evolves rapidly, optimal configuration would require quarterly reviews, and staying away from vendor lock-in. Potential savings derive from open-weight models deployed on institutional GPU infrastructure. If CNRS provides access to Mistral API, or to an Ollama server at Huma-Num, per-token API charges could be eliminated entirely. It is even likely that a fine-tuned 7B-12B model hosted at CIRED could be enough.

For hosting, a modest VPS suffices for small-to-medium RAG deployments. Institutional migration to CNRS or other institutional infrastructure would eliminate hosting costs entirely while improving data sovereignty and mission alignment. The low infrastructure cost profile (under €200/year) suggests that hosting provider selection should prioritize reliability, data residency, and institutional integration over marginal price differences.

5.3.2 Value Proposition

A cost-benefit assessment would have to compare the modest operational costs outlined above with multiple intangible benefit streams – research tool, mediation infrastructure, open science contribution – and strategic alignment with institutional mission.

Publication valorization represents a primary benefit. While a journal citation indicates scholarly influence, a conversational query reveals broader interest—from students exploring a research field / methods, to practicioners implementing CIRED ideas. This expanded audience for research aligns directly with CIRED's mission to bridge science and policy on environment and development challenges.

Science mediation constitutes another concrete benefit. The system can serve as infrastructure for knowledge brokerage towards journalists and educators, reducing barriers between specialized academic literature and non-specialist audiences. The CIRED.digital project establishes a proof-of-concept for AI-assisted science communication.

Research tool functionality emerged clearly from usage : literature review, methodological reference, citation discovery, and comparative analysis. Graduate students and early-career researchers particularly benefit from rapid access to CIRED's accumulated expertise without requiring deep familiarity with individual publications or researchers. Senior scholars used the system for quick fact-checking and corpus exploration during manuscript preparation. These research support functions complement rather than replace traditional literature search but improve efficiency.

CIRED.digital enhances institutional visibility and technical positioning, demonstrating capability to deploy advanced AI systems responsibly. The project contributes to collective learning on AI integration in research institutions, providing concrete evidence on what works, what fails, and what sustainability models prove viable. Open science commitments amplify impact through knowledge sharing and community contribution.

5.3.3 Pathways for the next year

Sustained operation requires transitioning from commercial cloud deployment to institutional infrastructure. We evaluated three pathways with distinct cost and governance implications (Table 8).

Table 8: Sustainability pathways: costs and characteristics

Pathway	Key Characteristics
Commercial cloud	Current solution; proven reliability; interim option
Institutional hosting	Zero infrastructure costs; CNRS data sovereignty; 6–12 month migration timeline
Federated network	Aspirational; requires 2+ years; multi-institution coordination

Institutional migration (recommended). Transitioning to CNRS infrastructure eliminates infrastructure costs while improving data sovereignty and institutional alignment. While institutions are still exploring how to support RAG in their research units, three options already stand out:

- **Huma-Num**: CNRS digital humanities infrastructure supporting experimental tools with GPU resources
- **Partner schools**: AgroParisTech, Ponts ParisTech, MSH, EHESS shared infrastructure
- **Internal infrastructure**: CIRED's own servers and GPU resources.

Migration challenges include 6–12 month coordination timelines, technical validation (Docker support, compute resources, API access), and governance definition (maintenance responsibilities, cost allocation, service expectations). Despite complexities, institutional hosting provides the most sustainable pathway for services persisting beyond individual projects. Target: complete migration within 12 months.

Commercial cloud (interim). The current Hetzner deployment provides reliable interim hosting (zero downtime during beta) with operational flexibility and minimal procurement overhead. Annual costs of around €120 for hosting, €90 for domain name, remain manageable but sacrifice data sovereignty and institutional capacity-building. This pathway suits the transition period or institutions prioritizing operational autonomy.

Federated network (aspirational). Coordinating shared RAG infrastructure across institutions through research networks could reduce costs through economies of scale. However, this requires robust multi-institution governance, and technical architecture supporting multi-tenancy. Ultimately, the CCSD could provide RAG as a service for HAL collections. This would eliminate infrastructure costs and simplify deployment for research units.

Ideal priorities for the next 12 months: secure institutional hosting, establish operational governance, formalize maintenance time budget, and conduct periodic evaluation of usage and value delivery.

6 Environmental Impact and Responsible AI Deployment

The deployment of AI systems in research institutions raises questions about environmental sustainability and responsible technology governance. This section presents the environmental impact assessment for CIRED.digital and documents the responsible deployment practices implemented during the pilot phase.

6.1 Marginal Carbon Footprint Assessment

We assessed operational emissions during the 96-day beta period (July 5–October 9, 2025) covering 290 user queries across 259 sessions. The assessment encompasses two emission sources: computational carbon footprint from LLM inference and infrastructure carbon footprint from VPS hosting. We employed component-based accounting, separately estimating emissions from computational activities and infrastructure operation, then summing to total footprint.

Several components fall outside the assessment boundary: embodied carbon from hardware manufacturing (typically $<$10% of operational emissions for cloud infrastructure), network transmission emissions ($<$5% of digital service footprints), user device emissions (consumed regardless of CIRED.digital access), the development phase and indirect effects from enabled or avoided activities. For computational emissions, we multiply measured token consumption from provider billing records by model-specific carbon coefficients from provider disclosures. For infrastructure emissions, we estimate server power consumption, multiply by operational hours, and apply grid carbon intensity for the datacenter location.

6.1.1 Infrastructure Carbon Footprint

The VPS configuration—3 vCPU, 4GB RAM, 50GB SSD on Hetzner Cloud's Helsinki datacenter—draws approximately 30-60W at full load (EPYC cores are 10-20W each). With the average utilization typical of CIRED.digital's predominantly idle workload profile – load curves show 1 vCPU at 25% on average – we assume an effective power consumption averages of 45W/12 = 3.75W. The 96-day beta period spans 2,304 hours, yielding total energy consumption of 8.64 kWh.

Grid carbon intensity varies dramatically by location. Finland's electricity grid exhibits 0.09 kg CO_2/kWh (2024 data), reflecting substantial nuclear and renewable generation. This compares favorably to the European average (0.35 kg CO_2/kWh) and alternatives like Germany (0.35 kg CO_2/kWh), while locations like Iceland and Norway achieve even lower intensities (0.01–0.05 kg CO_2/kWh) through near-complete renewable generation.

We adopt Finland's grid intensity (0.09 kg CO_2/kWh) for our primary estimate: 8.64 kWh × 0.09 kg/kWh = 0.78 kg CO_2 for infrastructure over the 96-day period. Using the conservative European average (0.35 kg CO_2/kWh) would yield 3.02 kg CO_2—an increase of 2.24 kg (+288%), i.e., Finland is a 74% reduction relative to the European-average assumption. Infrastructure emissions remain substantial because servers consume power continuously regardless of utilization, scaling with service availability (24/7 operation) rather than usage intensity. A system processing 10× more queries would increase computational emissions proportionally but leave infrastructure emissions essentially unchanged.

6.1.2 Generative AI carbon footprint

API usage was measured empirically using provider billing exports and dashboards. Table 6 presents the complete measured usage across development and deployment phases.

Usage patterns reveal distinct phase characteristics. Development concentrated in June with 20.6 million tokens as the team rapidly iterated on RAG architecture, tested multiple retrieval strategies, and refined the user interface. The transition to Mistral for public deployment in July coincided with the beta launch, generating peak deployment usage (1.12 million tokens). August and September showed sharp usage decline consistent with post-launch adoption curves.

For carbon footprint calculation, we focus on the 96-day beta deployment period using Mistral, as this represents the public-facing system operation. Development token consumption on OpenAI, while substantial, occurred during a distinct experimental phase with different models and pricing structures. Separating these phases provides clearer assessment of operational footprint for the deployed system. Moreover, indirect tokens used by coding tools like `Copilot` and `Devin.ai` during the development phase are not known, but certainly significant. Empirical studies show LLM inference emissions ranging from 0.1 to 10 grams of CO2 per million tokens (Mtoken), with smaller models like 7B parameters typically under 1 g CO2/Mtoken and larger ones exceeding 5 g under intensive conditions Patterson [2025], Luccioni and Jernite [2025].

Key drivers include location via grid carbon intensity (renewables reduce footprints by up to 70%), time-of-day usage, infrastructure such as GPU type and batching, model size with linear emission scaling, structure (e.g., reasoning chains), and thinking budget driving quadratic energy with sequence length Patterson [2025], Dhar [2025], Luccioni and Jernite [2025].

Uncertainties arise from 2–10x measurement variations due to inconsistent power models, ignored cooling overhead (up to 30%), fluctuating regional emission factors (50–800 gCO2/kWh), and overlooked lifecycle emissions Patterson [2025], Luccioni et al. [2023]. According to Team [2025], the majority of CO2 emissions arise from inference, but LLM training emissions are not negligible (10-40%) in a lifecycle perspective – but some companies claim 100% offset or renewable energy procurement.

Mistral AI [2025] provides carbon intensity disclosures for Mistral Large, computed by the company according to a robust Lifecycle Analysis methodology. It reports that after 18 months of usage, Mistral Large 2 generated 20,4 ktCO$_2$e, and that the marginal impact of inference is 1,14 gCO$_2$e per 400-token reponse. It also reports that footprint varies proportionally with model size. Given that Mistral Small is approximately 1/5th the size of Mistral Large 2 (24B vs. 123B), we estimate a marginal inference impact of approximately 0.228 gCO$_2$e per 400-token response, or 0.228 / 400 * 1000000 = 570 gCO$_2$e per Mtoken response.

Mistral reports marginal impacts per 400 output tokens ('response'). Converting to per-total-token values requires an assumed input/output token ratio. Assuming a 1:1 input-output ratio (typical for Q&A), total tokens per response double to 800, halving the marginal impact to approximately 285 gCO$_2$e per Mtoken. Applying that carbon coefficient (285g CO_2 per million tokens): 1,158,304 beta tokens × 285g/1M = 330g CO_2 for LLM inference across the 96-day beta period.

Initial corpus indexing processed approximately 1,300 documents totaling 2 million tokens through embedding models, contributing 100g CO_2 (using 50g/1M coefficient for embedding models). Query-level embeddings (converting each incoming question to a semantic vector) added approximately 2g CO_2. Processing overhead—query processing, retrieval operations, result ranking, logging, and analytics—contributed an estimated 20g CO_2 based on typical CPU utilization patterns.

Total computational footprint for beta deployment: LLM inference (330g) + initial embeddings (100g) + query embeddings (2g) + processing overhead (20g) = 452g CO_2, or 0.45 kg CO_2. This represents approximately 37% of total system emissions. Per-query computational emissions averaged 1.6g CO_2 (452g ÷ 290 queries), well below a Google search (7–50g depending on methodology) and substantially below email with attachment (50g) or brief video streaming (36–150g per hour).

6.1.3 Total carbon footprint and comparative analysis

Table 9 presents the complete carbon footprint breakdown for the 96-day beta deployment.

The cumulative carbon footprint totals 1.23 kg CO_2 equivalents for the 96-day beta deployment. Per-query emissions average 4.2g CO_2 (1,230g ÷ 290 queries), below many published estimates for a Google search (7–50g) and substantially below email with attachment (50g) or brief video call (150–300g per hour). Per-session emissions mirror per-query figures given most sessions involve single queries, though multi-query sessions exhibit lower per-query impact due to fixed infrastructure overhead amortization.

Table 10 situates CIRED.digital's footprint relative to conventional research activities. These comparisons establish orders of magnitude rather than precise substitution effects, as researchers typically

Table 9: Carbon footprint breakdown for 96-day beta deployment

Component	Calculation Basis	CO_2 (kg)	% of Total
LLM Inference	1,158,304 tokens × 285g/1M	0.330	26.8%
Embeddings (one-time)	2M tokens × 50g/1M	0.100	8.1%
Query Processing	Estimated overhead	0.020	1.6%
Query-level embeddings	290 queries × 50–100 tokens × 50g/1M	0.002	0.2%
Computational subtotal		*0.452*	*36.7%*
VPS Hosting	8.64 kWh × 0.09 kg/kWh	0.778	63.3%
Total		**1.23**	**100%**

combine multiple access methods.

Table 10: Carbon footprint comparison: CIRED.digital vs. conventional research activities

Activity	CO_2 Emissions
CIRED.digital query	4.2g
Google search (per query)	7–50g
Email with attachment	50g
Printing a 15 pages article (duplex)	40-80 g
Video call (30 minutes)	75-150 g
CIRED.digital (96 days)	1.23 kg
10 km automobile driving (French car)	2.5 kg
One researcher conference (EU flight)	500–1,500 kg

Projecting to sustained annual operation: infrastructure contributes approximately 3.0 kg CO_2/year (constant), while computational emissions scale with usage. At observed token consumption rates (1.16M tokens per 96 days), annual operation would consume approximately 4.4M tokens if each quarter is comparable to the beta phase, yielding approximately 1.3 kg CO_2/year for LLM inference (plus retrieval/processing overheads of similar magnitude to the beta period when scaled). Total projected annual footprint: approximately 4.7 kg CO_2/year, equivalent to about 19 km automobile driving. It is important to note that this projection assumes modest adoption; significant increases would raise computational emissions proportionally.

The system exhibits favorable scaling properties. While conventional activities exhibit linear emissions scaling, CIRED.digital exhibits sublinear scaling due to a fixed infrastructure component. At the beta deployment's 290 queries over 96 days, per-query infrastructure contribution was 2.7g CO_2. At 10× usage (2,900 queries), this drops to 0.27g CO_2. High-traffic institutional RAG systems deliver better per-query environmental performance than lightly-used pilots.

6.1.4 Key Findings and Optimization Insights

Three factors dominate environmental performance:

Infrastructure location matters most. With 63% of emissions from hosting under the Finland-based estimate, datacenter grid intensity remains the single largest lever on overall footprint. Using the European-average grid intensity (0.35 kg/kWh) would increase infrastructure emissions by a factor of 3.9 (0.78 ⟶ 3.02 kg CO_2 over 96 days). Selecting hosting in Iceland or Norway (0.01–0.05 kg CO_2/kWh) could reduce infrastructure emissions by an additional 44–89% relative to Finland.

Model selection affects computational emissions 5–10×. Mistral Small emits several times less than alternatives (e.g., Large class models) under comparable conditions. Because computational emis-

sions represent roughly 37% of total footprint in this deployment, model and provider choices (including datacenter energy sources) can materially affect the total system footprint in addition to affecting cost and latency.

Usage intensity improves efficiency. Fixed infrastructure costs (0.78 kg CO_2 for 96 days) amortize across queries. At 290 queries, per-query infrastructure contribution was 2.7g CO_2. At 2,900 queries (10× usage), this drops to 0.27g CO_2. Intensive, widespread use delivers better per-query environmental performance than sporadic deployment. This scaling property means successful adoption and increased utilization improve rather than degrade environmental credentials.

The measured data validates the fundamental finding: **for institutional RAG systems, the biggest environmental levers are (i) low-carbon electricity for always-on infrastructure and (ii) efficient models for inference**. Optimization efforts focused on only one of these dimensions miss a large share of the available reductions.

6.2 Responsible Deployment Practices

6.2.1 Privacy Protection and Transparency

CIRED.digital implements privacy-by-design through anonymous session identifiers. User profile is stored client-side and can be purged by the user at anytime – the user profile is stored in the logs on European private servers, and contain only demographic information. No user registration, authentication, or account creation is required. Query text and interaction logs capture research usage patterns but are anonymized. Optional "confidentiality mode" disables all logging for users requiring absolute privacy. No framework or hidden tracking mechanisms are employed.

GDPR alignment follows data minimization (collecting only information necessary for service operation), purpose limitation (restricting data use to stated objectives), transparent disclosure (accessible privacy documentation), user control (opt-out via confidentiality mode), and institutional data sovereignty (European datacenters under CNRS control). Anonymized datasets prepared for public archival remove residual identifiers, enabling open science contributions while protecting individual privacy.

The citation mechanism constitutes the core transparency feature. Every CIRED.digital response includes explicit citations linking claims to source passages in original publications. Citation marks display original text excerpts (up to 600 characters per citation) with document metadata (title, authors, publication year, DOI/HAL identifier). Users can verify claims by examining source material directly. User feedback consistently identified citations as the most valued feature, with satisfaction correlating strongly with citation relevance and completeness.

System capability communication establishes appropriate user expectations. The landing page clearly identifies Cirdi as a "documentaliste scientifique" rather than general-purpose chatbot. The Help page displays system limitations—no conversational memory (stateless Q&A mode), no real-time data or current events, restriction to indexed CIRED documents, in order to prevent user frustration from mismatched expectations. Out-of-scope queries are declined since the RAG does not find relevant information.

6.2.2 Risk Mitigation

Hallucination risk. Despite RAG grounding, LLMs retain capacity to generate plausible but unsupported content. Mitigation strategies included conservative generation parameters prioritizing accuracy over fluency, mandatory citation mechanisms enabling user verification, and epistemic humility in system communication. User feedback mechanisms (thumbs up/down with optional comments) enabled error reporting, with negative feedback correlating strongly with citation quality. Formal response quality audits could be conducted in future production deployments to monitor hallucination rates and citation accuracy.

Misattribution risk. When AI-generated content bears implicit institutional endorsement, users may interpret responses as authoritative CIRED positions rather than syntheses of published liter-

ature. CIRED.digital addresses this through explicit presentation as an AI-generated synthesis and encouragement of source verification through citation mechanisms.

Corpus limitations. CIRED publications emphasize environment-development intersections primarily from economics and policy perspectives, with geographic focus on France, Europe, and Global South contexts. This introduces perspective bias: questions outside CIRED expertise receive incomplete answers. The system transparently acknowledges these limitations through corpus scope documentation illustrated by the landing page wordcloud.

Misuse potential. We hardened R2R default docker configuration and added a Nginx Proxy Server to implement rate limiting and anomaly detection for protection. No significant abuse occurred during the beta deployment.

6.2.3 Governance Structure

The pilot operated under lightweight governance appropriate to its experimental scale. The phased deployment approach—Alpha (internal testing with convenience testers), Beta Closed (invited external testers with structured feedback), Beta Open (unrestricted public access following institutional approval)—enabled iterative refinement and risk mitigation. Steering committee meetings during development maintained alignment among technical implementation, institutional priorities, and user needs. User feedback mechanisms provided continuous quality signals. Institutional oversight through CIRED leadership approval preceded each deployment phase expansion.

This governance model sufficed for pilot demonstration but requires formalization for sustained operation. Production deployment would require: designated operational responsibility (system administrator, maintenance schedule, budget allocation), quality assurance procedures (periodic citation accuracy audits, response quality assessment, user satisfaction monitoring), incident response protocols (handling reported errors, addressing abuse, managing service disruptions), corpus curation policies (update frequency, content inclusion criteria), and decision-making authority (feature priorities, technology migration, service continuation or termination).

CIRED.digital functions as pilot research project rather than production service, with experimental status communicated clearly to users. This positioning provides flexibility for iteration and learning while managing expectations around service reliability and continuity.

6.3 Implications for Institutional AI Deployment

The CIRED.digital pilot demonstrates that research institutions can deploy conversational AI systems for knowledge access with modest environmental footprint (4.7 kg CO_2/year projected) and responsible practices protecting user privacy while enabling verification through citations.

Key lessons for institutional deployment include:

Environmental impact is manageable and optimizable. With Finland-based hosting, infrastructure contributes about 63% of emissions. Moving from European-average grids to low-carbon grids (Finland, Iceland, France, Norway) reduces total footprint substantially; model selection affects computational emissions 5–10× (proportional to model size) and is environmentally meaningful when inference is a non-trivial share of the total. Usage intensity improves efficiency—the system becomes greener per query as adoption increases.

Privacy and transparency require architectural commitment. Anonymous sessions, optional data collection, and explicit citation mechanisms must be designed in from the start rather than retrofitted. User trust correlates strongly with citation quality and verifiability. Clear communication of capabilities and limitations reduces frustration and inappropriate use.

Phased deployment enables learning and risk mitigation. Progressive expansion from internal testing to invited beta to public access allowed identification of interface issues, expectation mismatches, and quality concerns before broad exposure. This approach reduces reputational risk while building confidence in system reliability.

Pilot governance differs from production requirements. Lightweight steering and monthly coordination suffice for exploratory projects. Sustained operation requires formalized responsibility, quality assurance procedures, incident response protocols, and institutional commitment to maintenance and evolution.

Separate development from deployment assessment. Development token consumption (95% of total project volume) occurred during a distinct experimental phase with different objectives and models. Environmental and cost assessment should focus on operational deployment rather than one-time prototyping activities to provide realistic projections for sustained operation.

The evidence supports institutional RAG deployment as technically feasible, environmentally sustainable at modest scale, and valuable for research dissemination when implemented with attention to privacy, transparency, and appropriate governance. The path forward requires institutional hosting transition, formalized operational frameworks, and continued optimization of resource efficiency.

7 Conclusion

The CIRED.digital project achieved its primary objectives: developing, testing, and deploying a natural-language interface to CIRED's publication corpus, while providing reproducible open-source infrastructure that other research institutions can replicate. In five months with modest resources, the team delivered a production system, complete documentation, and an empirical analysis of usage and costs.

7.1 Project Achievements and Value

The project exemplifies open science. The full codebase is published on GitHub under the CeCILL-B license, with documentation, deployment scripts, and a test suite. A parameterized, modular, container-based architecture enables other institutions to deploy similar systems on their HAL collections with limited effort (2–4 hours for skilled operators). An anonymized dataset of 290 user queries and aggregated usage statistics are prepared for open deposit, enabling follow-up studies on question types and system performance. The documentation also addresses practical deployment issues that are often under-reported, reducing adoption barriers.

For CIRED, the system supports knowledge translation and demonstrates a commitment to reproducible, shareable infrastructure. For the broader community, it shows that production RAG systems are feasible with modest resources, provides a working reference implementation, contributes evidence on adoption in an academic setting, and models responsible deployment emphasizing transparency, citations, and environmental awareness.

Outputs span infrastructure, analysis, and open science artifacts. CIRED.digital provides a digital librarian with natural-language access to about 1,238 CIRED publications, deployed on reliable cloud infrastructure with transparent citations. The code follows professional standards (Python 3.11+, type annotations, linting, tests). Logging captured 1,849 interactions across 259 sessions, enabling empirical analysis. Docker-based deployment supports one-command setup.

Limitations remain. Input token counting was not automated. User profiles were optional (low capture), and feedback collection (thumbs/comments) was sparse. While 259 sessions suffice to identify broad patterns, they do not support robust statistical inference. Finally, a single-system pilot limits comparisons across RAG engines or LLM providers.

Empirical contributions include a first test bench for AI-mediated access to scientific publications. A cost-benefit assessment supports feasibility (€50–200/year operating costs), with environmental impact for the 96-day pilot estimated at 3–4 kg CO_2. Practical lessons—PDF processing, multilingual support, expectation management, and interface design—are documented. Anonymized query and interaction logs support future studies of AI-assisted science mediation.

7.2 Future Development

For CIRED Leadership. The beta deployment reveals several strategic tensions requiring institutional decision-making.

Adoption and Audience Strategy. Usage is concentrated among CIRED network users (40%) and RENATER-affiliated researchers (20%). CIRED.digital can be positioned as an internal tool (institutional memory, onboarding) or expanded toward journalists, policymakers, and students. The 85% in-scope query rate signals utility for those who find it, but modest usage (259 sessions over 96 days) suggests adoption barriers need attention.

Corpus Currency and Science Communication. Infrequent HAL updates reduce news value. CIRED can invest in daily automated updates to support media engagement, or accept Cirdi as primarily archival/synthesis, complementary to direct researcher-media interactions and the lab newsletter.

Interaction Paradigm and User Expectations. A search interface helped manage expectations and reduce hallucination risk. Yet 30% of users expected multi-turn context, and several requested structured outputs (tables, timelines, comparisons). Future work can either meet these expectations (context: 2–6

weeks; structured outputs: 3–4 weeks) or keep a focused documentary search model, balancing mission alignment against QA complexity.

Continuing operation requires modest costs (€50–200/year) plus 1–2 days/month for maintenance. The system supports research valorization and open science, but CIRED already has multiple communication channels. The AI agent is new and may not be expected by journalists; update lag also limits time-sensitive use, so CIRED should clarify its role in the communication ecosystem.

Internal vs. External Positioning. Retargeting toward internal audiences merits consideration. The tool can support institutional memory and help researchers navigate colleagues' work. However, interns often relied only on supervisor-provided materials, suggesting that tool availability alone may not change research practices.

Open issues and ideas are tracked at https://github.com/CIRED/cired.digital/issues. Near-term enhancements could focus on:

- multi-turn context for follow-up questions (2–3 weeks),
- citation UI refinement (side-by-side view vs tooltips; 1–2 weeks),
- systematic quality evaluation (4–6 weeks),
- ingestion robustification toward daily HAL updates.

HAL full texts do not cover most CIRED output. The knowledge base could be extended backward using scanned historical collections (modern OCR can help) and forward by adding paywalled articles if CIRED maintains a repository. The latter would change Cirdi's nature by citing documents not accessible to readers.

Infrastructure planning should prioritize institutional hosting evaluation now (CNRS Huma-Num and partner IT). A transition within 12 months would reduce costs, improve sovereignty, and align with research infrastructure strategy. If unavailable, continue commercial hosting while optimizing providers and adding response caching.

Cirdi is a first step in a broader discussion on AI tools in CIRED infrastructure. Allowing researchers to add documents to the RAG is attractive, but would require governance choices, especially for agentic assistants. R2R includes experimental "deep research" features that were intentionally disabled in Cirdi due to their epistemic and governance implications. CNRS-wide tools (e.g., via Mistral) may be complemented by in-house systems offering tighter corpus control and institutional context.

For Other Research Institutions. The deployment pathway for replication institutions involves several stages. First, assess HAL collection size and API access (optimal for 1,000–10,000 publications). Second, evaluate LLM providers (Mistral offers lowest cost at €0.14/1M tokens). Third, provision minimal infrastructure requiring 2–4 vCPU, 4–8 GB RAM, and 100 GB SSD (€5–20/month). Fourth, deploy using provided Docker Compose configuration (2–4 hours work). Fifth, customize frontend and branding (1–2 weeks). Finally, plan ongoing operation allowing ∼4 hours/month maintenance and quarterly provider evaluation.

Recommendations: use a production-ready RAG engine; PostgreSQL+pgvector for collections under 100K documents; a cost-effective LLM provider (e.g., Mistral); and EU-hosted infrastructure (e.g., Hetzner) for data residency. On process: track costs from day one, collect usage data with mandatory profiles, include environmental assessment, plan phased internal testing, and maintain regular stakeholder engagement.

For the Research Community. AI-assisted science mediation insights from this project indicate several critical design principles. Citation mechanisms are fundamental—users strongly prefer explicit source attribution. Expectation clarity matters considerably—clear communication of capabilities reduces frustration significantly. Iterative refinement requires resource allocation for feedback collection

and interface iteration. Cost-benefit remains favorable—RAG deployment remains accessible to most research institutions.

Open science opportunities include standardized REST APIs for institutional RAG systems, federated networks (library-consortium style), evaluation frameworks for comparative studies, and environmental reporting standards. Ultimately, HAL could offer RAG by default as an evolution of the search box.

Medium-term priorities (6–18 months) include benchmarking alternative RAG engines (e.g., LlamaIndex, PaperQA), domain optimization (embeddings and chunking), support for replication by other institutions, and governance formalization for sustained operation.

Strategic opportunities include federated infrastructure via CNRS Huma-Num or library consortia, standardization and best practices, integration with research information systems, and educational applications leveraging CIRED's methodology literature.

7.3 Implications and Future Outlook

CIRED.digital shows how research institutions can deploy AI to support research, teaching, and science communication while remaining aligned with open science, environmental responsibility, and cost-effectiveness. It demonstrates that institutions need not rely on commercial platforms for AI-mediated access to institutional knowledge: locally controlled systems are feasible, preserve independence, protect usage data, and can contribute to shared research infrastructure.

As AI-assisted tools spread, rigorous work on quality, ethics, and impact remains needed. This project suggests that a digital librarian can support diverse information needs when designed with explicit citations, user agency, and clear capability boundaries.

By quantifying costs (€50–200/year) and environmental impact (3–4 kg CO_2 for the pilot), CIRED.digital shows that AI deployment can be compatible with sustainability commitments. Compared to alternative access mechanisms (e.g., travel, printing), a thoughtfully implemented system can be part of sustainable research infrastructure. From a cost perspective, RAG is a practical way to broaden access to lab results that are otherwise restricted by commercial copyrights.

Overall, CIRED.digital validates a model of institutional AI systems that serve research missions while remaining community-controlled, transparent, and sustainable in cost and impact. The evidence in this report supports continued investment. Next steps—institutional hosting, targeted feature upgrades, and support for replication—will determine how strongly this pilot shapes emerging practices for responsible, cost-effective, and sustainable institutional RAG deployment.

Glossary

Technical Terms

API (Application Programming Interface)
A set of protocols that allows different software applications to communicate. In this project, APIs enable CIRED.digital to access LLM services from providers like Mistral and OpenAI.

Chunking
The process of dividing long documents into smaller segments (chunks) for more effective retrieval and processing. CIRED.digital uses recursive chunking with 512-token segments and 50–100 token overlap.

Docker / Containerization
A platform that packages software and its dependencies into standardized units (containers) for consistent deployment across different computing environments. CIRED.digital uses Docker for reproducible deployment.

Embedding / Vector Embedding
A mathematical representation of text as arrays of numbers (vectors) that capture semantic meaning, enabling similarity search. Documents and queries are converted to embeddings for retrieval.

Git / GitHub
Version control system (Git) and code hosting platform (GitHub) used for collaborative software development. CIRED.digital's complete codebase is published on GitHub under open-source license.

Hallucination (in AI context)
When a language model generates plausible-sounding but factually incorrect or unsupported information. RAG systems reduce hallucinations by grounding responses in retrieved documents.

LLM (Large Language Model)
AI models trained on vast text corpora to understand and generate human-like text. Examples include GPT-4 (OpenAI), Claude (Anthropic), and Mistral models.

OCR (Optical Character Recognition)
Technology that converts scanned images of text into machine-readable text. Approximately 5–10% of CIRED PDFs required OCR processing during ingestion.

PostgreSQL / pgvector
PostgreSQL is an open-source relational database; pgvector is an extension enabling efficient storage and search of vector embeddings for semantic similarity.

RAG (Retrieval-Augmented Generation)
A technique that combines information retrieval from a knowledge base with text generation, grounding AI responses in authoritative source documents. The core architecture of CIRED.digital.

R2R "RAG to Riches" – an open-source RAG framework developed by SciPhi that provides document indexing, retrieval, and generation capabilities. The engine underlying CIRED.digital.

Token The basic unit of text processing in LLMs, roughly equivalent to 0.75 words in English. LLM providers charge by tokens consumed (input + output).

VPS (Virtual Private Server) A virtualized server that provides dedicated computing resources. CIRED.digital operates on a Hetzner Cloud VPS with 3 vCPU, 4 GB RAM, and 100 GB storage.

French Research Infrastructure

CIRED (Centre International de Recherche sur l'Environnement et le Développement) A joint CNRS research unit partnering with AgroParisTech, EHESS, CIRAD, and Ponts ParisTech, focusing on environment-development intersections. The institution whose publications CIRED.digital indexes.

CNRS (Centre National de la Recherche Scientifique) France's national public research organization, the largest fundamental research institution in Europe. CIRED is a CNRS joint research unit (UMR 8568).

GDPR / RGPD General Data Protection Regulation (Règlement Général sur la Protection des Données) – European Union regulation governing personal data protection and privacy. CIRED.digital implements GDPR-compliant anonymous sessions and optional data collection.

HAL (Hyper Articles en Ligne) France's national open-access repository for scholarly publications. CIRED.digital retrieves its corpus of ∼1,238 publications from the HAL-CIRED collection.

Huma-Num CNRS research infrastructure supporting digital humanities and social sciences through computing resources, data services, and experimental AI tools including RAG platforms.

RENATER (Réseau National de télécommunications pour la Technologie, l'Enseignement et la Recherche) France's national research and education network providing internet connectivity to universities and research institutions. 11% of CIRED.digital users accessed the system via RENATER.

Domain-Specific Terms

IMACLIM A family of integrated assessment models developed at CIRED for analyzing interactions between economic growth, technological change, and climate policy.

IPCC (Intergovernmental Panel on Climate Change) The United Nations body for assessing climate change science. CIRED researchers have contributed as lead authors to IPCC reports, including work recognized by the 2007 Nobel Peace Prize.

Open Science A movement to make scientific research, data, and dissemination accessible to all. CIRED.digital contributes through open-source code, transparent documentation, and public data sharing.

Project-Specific Terms

Cirdi The name given to CIRED.digital's AI-powered scientific documentalist (a playful reference to CIRED combined with a personal name suffix).

HAL Collection A curated subset of HAL repository publications associated with a specific institution or project. The HAL-CIRED collection contains CIRED's open-access publications.

Scientific Documentalist The positioning of CIRED.digital as a specialized search and synthesis tool for CIRED publications, distinct from general-purpose chatbots. Emphasizes grounded, citation-backed responses over conversational AI.

References

Shubham Agarwal, Gaurav Sahu, Abhay Puri, Issam H. Laradji, Krishnamurthy D. J. Dvijotham, Jason Stanley, Laurent Charlin, and Christopher Pal. Litllm: A toolkit for scientific literature review. *arXiv preprint*, 2402.01788, 2024. doi: 10.48550/arXiv.2402.01788. URL https://arxiv.org/abs/2402.01788. Toolkit for RAG-based scientific literature review workflows.

Alaa Al Khourdajie. The role of artificial intelligence in climate change scientific assessments. *PLOS Climate*, 4(9):e0000706, 2025. doi: 10.1371/journal.pclm.0000706.

Tatsuya Amano, Juan P. González-Varo, and William J. Sutherland. Languages are still a major barrier to global science. *PLOS Biology*, 14(12):e2000933, 2016. doi: 10.1371/journal.pbio.2000933.

Muhammad Arslan, Hussam Ghanem, Saba Munawar, and Christophe Cruz. A survey on RAG with large language models. *Procedia Computer Science*, 246:3781–3790, 2024. ISSN 1877-0509. doi: 10.1016/j.procs.2024.09.178. URL https://doi.org/10.1016/j.procs.2024.09.178.

Andrew Brown, Muhammad Roman, and Barry Devereux. A systematic literature review of retrieval-augmented generation: Techniques, metrics, and challenges. *arXiv preprint*, 2508.06401, 2025. URL https://arxiv.org/abs/2508.06401. Preprint; submitted to IEEE.

Marek Dąbrowski et al. Retrieval-augmented generation (RAG) chatbots for education: A survey of applications. *Applied Sciences*, 15(8):4234, 2025. doi: 10.3390/app15084234.

Damien Desbordes. *Les robots vont-ils remplacer les journalistes ?* Éditions Plein Jour, Paris, 2018.

Dhar. Measuring the energy footprint of llm inference, 2025. URL https://arxiv.org/abs/2511.05597.

Direction Générale des Entreprises (DGE). Guide de la génération augmentée par récupération (rag). Technical report, Ministère de l'Économie, des Finances et de la Souveraineté Industrielle et Numérique, nov 2024. URL https://www.entreprises.gouv.fr/files/files/Publications/2024/Guides/20241127-bro-guide-ragv4-interactif.pdf.

Europe PMC. Ai-assisted scientific summaries and patient-friendly explanations: Europe pmc pilot projects. Technical report, 2024. URL https://europepmc.org. Describes RAG-powered systems for biomedical QA and multilingual lay summaries.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2024. Updated March 2024.

Binglan Han, Teo Susnjak, and Anuradha Mathrani. Automating systematic literature reviews with retrieval-augmented generation: A comprehensive overview. *Applied Sciences*, 14(19):9103, 2024. doi: 10.3390/app14199103. URL https://doi.org/10.3390/app14199103.

HN Lab. Applications de RAG pour les humanités numériques. Source code repository, 2024. URL https://gitlab.huma-num.fr/hnlab/applications-de-rag. Dépôt GitLab d'exemples d'applications RAG pour les corpus SHS (interfaces de recherche, chat sur corpus, etc.).

Jakub Lála, Odhran O'Donoghue, Aleksandar Berčič, Samuel Blunsden, Marta Skreta, Yuqing Tong, Rajesh Ranganath, and Alán Aspuru-Guzik. PaperQA: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*, 2024.

Le Canard enchaîné. [article sur marlowe et la pétition josé bové]. *Le Canard enchaîné*, July 2003. Article rapportant l'incident où le logiciel Marlowe a signé une pétition pour la libération de José Bové.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020. URL https://arxiv.org/abs/2005.11401.

Luccioni and Jernite. Energy costs of communicating with ai. *Frontiers in Communication*, 10, 2025. doi: 10.3389/fcomm.2025.1572947. URL https://www.frontiersin.org/articles/10.3389/fcomm.2025.1572947/full.

Luccioni, Viguier, and Ligozat. Estimating carbon footprint of bloom language model. *JMLR*, 24, 2023. URL https://www.jmlr.org/papers/volume24/23-0069/23-0069.pdf.

David M. Markowitz. From complexity to clarity: How AI enhances perceptions of scientists and the public's understanding of science. *PNAS Nexus*, 3(9):pgae387, 2024. doi: 10.1093/pnasnexus/pgae387.

Mistral AI. Our contribution to a global environmental standard for AI, jul 2025. URL https://mistral.ai/news/our-contribution-to-a-global-environmental-standard-for-ai.

Luise Modersohn et al. Generalization bias in large language model summarization of scientific research. *Royal Society Open Science*, 12(4):241776, 2025. doi: 10.1098/rsos.241776.

Murtiyoso Murtiyoso, Imam Tahyudin, and Berlilana Berlilana. A systematic review of retrieval-augmented generation for enhancing domain-specific knowledge in large language models. *Sinkron: Jurnal dan Penelitian Teknik Informatika*, 9(2):969–977, 2025. doi: 10.33395/sinkron.v9i2.14824. URL https://doi.org/10.33395/sinkron.v9i2.14824.

Patterson. Energy consumption & carbon emissions of llm inference, 2025. URL https://arxiv.org/abs/2507.11417.

Antonin Pottier, editor. *Concilier économie et écologie : les textes fondateurs du Centre international de recherche sur l'environnement et le développement.* Éditions Eyrolles, 2024. ISBN 978-2-85978-553-6. URL https://www.eyrolles.com/Sciences/Livre/concilier-economie-et-ecologie-les-textes-fondateurs-du-centre-international-de-recherche Publication à l'occasion du cinquantième anniversaire du CIRED. Textes écrits entre 1972 et 1997.

Stéphane Pouyllau. Quels usages du "retrieval-augmented generation" en SHS ? Blog post, 2024a. URL https://hnlab.huma-num.fr/. Billet de blog HN Lab Log, 17 mars 2024.

Stéphane Pouyllau. Hackathon I – retrieval-augmented generation en SHS. Blog post, 2024b. URL https://hnlab.huma-num.fr/. Compte rendu de hackathon HN Lab sur les usages de la RAG en sciences humaines et sociales.

Stéphane Pouyllau. Explorer ses documents de travail avec les méthodes de la retrieval-augmented generation: créer, démonter et mettre en œuvre une application web complète. Blog post, 2025. URL https://hnlab.huma-num.fr/. Billet de blog HN Lab Log sur la mise en pratique d'une application RAG pour les chercheurs en SHS.

Stéphane Pouyllau and collaborators. ISIDORE 2030: des LIA de traitement des données de recherche au contrôle d'usage. Technical report, 2024a. Rapport de projet ISIDORE 2030, disponible sur Zenodo. Contient une discussion d'architectures RAG pour les corpus SHS.

Stéphane Pouyllau and collaborators. ISIDORE 2030: adapter un moteur de recherche académique aux IA génératives pondérées. Technical report, 2024b. Texte d'orientation sur l'évolution d'ISIDORE vers des usages avec RAG et LLM, disponible sur Zenodo.

David Rau, Vladimir Karpukhin, Barlas Oguz, Dzmitry Okhonko, and Fabio Petroni. Retrieval-augmented generation in multilingual settings. *arXiv preprint arXiv:2407.01463*, 2024.

Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. Fine tuning vs. retrieval augmented generation for less popular knowledge. *arXiv preprint arXiv:2403.01432*, March 2024. URL https://arxiv.org/abs/2403.01432. Published March 7, 2024.

DitchCarbon Team. The real carbon cost of an ai token, apr 2025. URL https://ditchcarbon.com/blog/llm-carbon-emissions.

Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Bingler, Jingwei Ni, Mathias Kraus, Simon Allen, Chiara Colesanti-Senni, Tobias Wekhof, Tobias Schimanski, Glen Gostlow, Tingyu Yu, Qian Wang, Nicolas Webersinke, Christian Huggel, and Markus Leippold. ChatClimate: Grounding conversational AI in climate science. *Communications Earth & Environment*, 4:480, 2023. doi: 10.1038/s43247-023-01084-x.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. Rankrag: Unifying context ranking with retrieval-augmented generation in llms. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024. doi: 10.52202/079017-3850. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/db93ccb6cf392f352570dd5af0a223d3-Abstract-Conference.html.