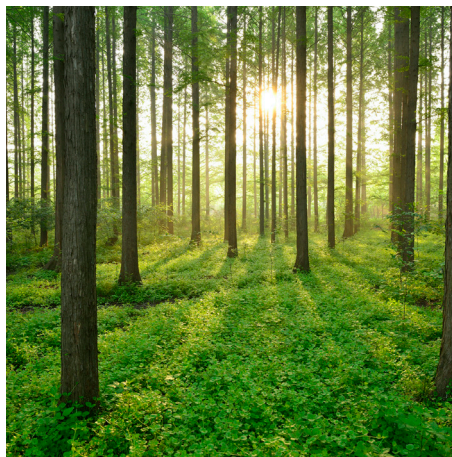
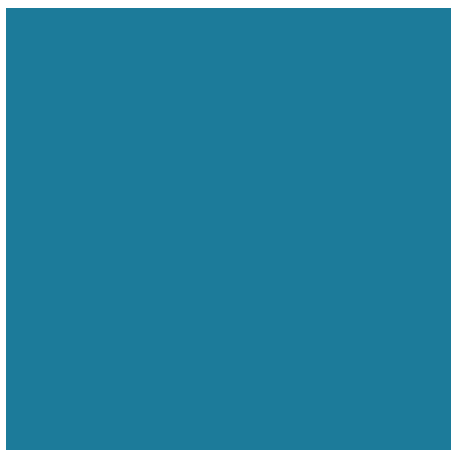


ON-PREMISE AI-ASSISTED SEARCH FOR THE NORDIC NUCLEAR INDUSTRY

REPORT 2026:1217



DIGITALISATION
IN NUCLEAR POWER



On-Premise AI-Assisted Search for the Nordic Nuclear Industry

Decision Support for Secure,
Sovereign and Multilingual Search

ARTO LAIKARI, ARASH HAJIKHANI, LASSI WARSTA

ISBN 978-91-89917-60-6 | © Energiforsk September 2026

Energiforsk AB | Phone: 08-677 25 30 | E-mail: kontakt@energiforsk.se | www.energiforsk.se

Foreword

This report presents the results of a project carried out within the Energiforsk Digitalisation in Nuclear Power Programme. The programme provides a platform for Nordic nuclear power companies to jointly explore how digital technologies can contribute to improved safety, quality and efficiency in nuclear power operations.

Within the programme, a need was identified to improve access to information distributed across internal databases, document archives and other data sources. Finding relevant information is often time-consuming, affecting both productivity and decision-making. For this reason, the programme initiated this project to investigate the feasibility of AI-assisted search solutions that can be deployed and operated in local, on-premises environments. The project evaluated available technologies and examined their suitability for use in the security-sensitive nuclear industry.

The project was carried out by VTT Technical Research Centre of Finland. The authors of the report are Arto Laikari, Arash Hajikhani and Lassi Warsta. The project was followed by a reference group comprising Sune Jonsson (OKG) and Tapio Taipale (TVO), representing participating nuclear industry organisations.

These are the results and conclusions of a project, which is part of a research programme run by Energiforsk. The author/authors are responsible for the content.

Summary

This report provides decision support for selecting and piloting on-premise AI-assisted search and document intelligence solutions for the Nordic nuclear industry, with emphasis on data sovereignty, security, traceability, multilingual retrieval quality and practical deployment feasibility.

This report evaluates on-premise AI-assisted search and document intelligence options for the Nordic nuclear industry under the Energiforsk Digitalization in Nuclear programme. It compares three European-headquartered commercial platforms - Sinequa (ChapsVision, France), Mindbreeze InSpire (Austria), IntraFind iFinder/iAssistant (Germany) and a sovereignty-oriented modular open-source stack - across architecture, security, traceability, Nordic-language depth and total cost of ownership over five years. The work then widens the analysis along the dimensions raised during the work: the modality load on language models (images, drawings and scanned documents), the latest open-weight models available in May 2026 (Llama 4, Mistral Large 3, Qwen 3.5, DeepSeek V3.x/V4, Gemma 4), the multimodal counterparts used to read engineering documents (Qwen3-VL, InternVL 2.5, Llama 3.2 Vision, DeepSeek-OCR), and an indicative cost model for on-premise GPU infrastructure together with expected retrieval throughput. A target deployment stack is proposed in which a knowledge graph acts as the curated semantic and relationship layer complementing the keyword and vector indexes and retrieval-augmented generation pipelines sit on top to drive chat, insight extraction and dashboarding. For a Phase 1 pilot, Mindbreeze is identified as the preferred starting candidate because its appliance deployment model is expected to reduce integration and operating effort, allowing the pilot to concentrate on retrieval quality, citations and access control. This is a conditional recommendation rather than a final procurement decision. Sinequa may be preferable where page-level citation precision and complex access control dominate; IntraFind may be preferable where commercial cost and European data sovereignty dominate; and a sovereign modular stack may be preferable where long-term component control justifies substantially greater engineering responsibility. A common comparative pilot and updated vendor quotations are required before production selection. The principal advantages, disadvantages, uncertainties and best-fit conditions of the four options are consolidated in Table 4. Findings are presented as decision support for a phased pilot; empirical validation—particularly of Finnish and Swedish retrieval quality and citation faithfulness—remains a prerequisite for proceeding beyond the pilot stage. References are provided in an author–date format.

Technology currency and uncertainty. The AI technology and supplier landscape is developing rapidly. This report reflects the information, technologies, product capabilities, licences, prices and assumptions available in May 2026. These may change as new models, software versions, hardware, regulatory interpretations and vendor offerings emerge. The comparisons and recommendations should therefore be treated as a time-specific basis for pilot planning rather than as

permanent conclusions, and the principal assumptions should be reviewed before procurement or production deployment.

Keywords

on-premise AI; large language models; retrieval-augmented generation; knowledge graphs; enterprise search; nuclear industry; data sovereignty; Nordic languages.

Sammanfattning

Rapporten ger beslutsunderlag för val och pilotinförande av lokalt installerade lösningar för AI-stödd sökning och dokumentintelligens inom den nordiska kärnkraftsindustrin. Särskild vikt läggs vid datasuveränitet, informationssäkerhet, spårbarhet, kvaliteten på flerspråkig informationssökning samt möjligheterna att genomföra installation och drift i praktiken.

Inom ramen för Energiforsks program Digitalisering i kärnkraft utvärderas tre kommersiella plattformar med huvudkontor i Europa: Sinequa från ChapsVision i Frankrike, Mindbreeze InSpire i Österrike och IntraFind iFinder/iAssistant i Tyskland. Dessa jämförs även med en suveränitetsinriktad, modulär lösning baserad på öppen källkod. Jämförelsen omfattar arkitektur, säkerhet, spårbarhet, stöd för nordiska språk samt den uppskattade totala ägandekostnaden under fem år.

Analysen omfattar också beräkningsbelastningen för olika typer av innehåll, såsom bilder, tekniska ritningar och skannade dokument. Vidare behandlas öppna modeller som var tillgängliga i maj 2026, däribland Llama 4, Mistral Large 3, Qwen 3.5, DeepSeek V3.x/V4 och Gemma 4. Även multimodala modeller för tolkning av tekniska dokument granskas, exempelvis Qwen3-VL, InternVL 2.5, Llama 3.2 Vision och DeepSeek-OCR. Rapporten presenterar dessutom en översiktlig kostnadsmodell för lokal GPU-infrastruktur och den förväntade kapaciteten för informationssökning.

Som målarkitektur föreslås en lösning där en kunskapsgraf fungerar som ett kvalitetssäkrat semantiskt lager för entiteter, relationer och informationsursprung. Kunskapsgrafen kompletterar nyckelords- och vektorindex, medan RAG-baserade processer används för dialogtjänster, insiktsutvinning och visualisering i kontrollpaneler.

För en pilot i fas 1 identifieras Mindbreeze som den föredragna utgångspunkten. Skälet är att plattformens appliance-baserade leveransmodell förväntas minska behovet av integration och löpande teknisk administration. Pilotens resurser kan därmed i högre grad inriktas på att utvärdera sökkvalitet, citeringarnas tillförlitlighet och åtkomstkontroll.

Rekommendationen är villkorad och ska inte betraktas som ett slutligt upphandlingsbeslut. Sinequa kan vara ett bättre alternativ när sidexakta källhänvisningar och komplex åtkomstkontroll har högsta prioritet. IntraFind kan vara att föredra när kommersiell kostnad och europeisk datasuveränitet väger tyngst. En suverän, modulär lösning kan vara det mest lämpliga alternativet när långsiktig kontroll över enskilda komponenter motiverar ett betydligt större internt ansvar för utveckling och drift.

Innan en produktionslösning väljs krävs en gemensam jämförande pilot, uppdaterade leverantörsoffertter och en enhetlig utvärdering av alternativen. Resultaten presenteras därför som beslutsunderlag för ett stegvis pilotförfarande.

De fyra alternativens huvudsakliga fördelar, nackdelar, osäkerheter och lämpligaste användningsförutsättningar sammanfattas i tabell 4 (Table 4). Empirisk validering, framför allt av sökkvaliteten på finska och svenska samt av citeringarnas tillförlitlighet, är en förutsättning för att gå vidare från pilot till produktionsinförande. Referenserna anges enligt författare-år-systemet.

Teknikens aktualitet och osäkerhet. AI-tekniken och leverantörslandskapet utvecklas snabbt. Denna rapport återspeglar den information, teknik, produktfunktionalitet, licensiering, prissättning och de antaganden som var aktuella i maj 2026. Dessa kan förändras i takt med att nya modeller, programvaruversioner, maskinvara, regulatoriska tolkningar och leverantörserbjudanden utvecklas. Jämförelserna och rekommendationerna bör därför betraktas som ett tidsbundet underlag för planeringen av pilotprojektet, snarare än som varaktiga slutsatser. De viktigaste antagandena bör ses över före upphandling eller produktionsdriftsättning.

List of content

1	Introduction	10
1.1	Background and motivation	10
1.2	Aims and objectives	11
1.3	Scope and limitations	11
1.3.1	On-premise boundary and cloud-connected sources	12
2	Methodology and approach	13
2.1	Overall approach	13
2.2	Selection criteria	13
2.3	Evaluation dimensions	13
2.4	Evidence status	13
2.4.1	Pilot decision gates	13
2.5	Comparative assessment method	14
2.5.1	Decision logic for the pilot recommendation	14
3	Requirements for on-premise AI search in the nuclear domain	15
3.1	Nuclear-specific constraints	15
3.1.1	Regulatory environment	15
3.1.2	Confidentiality and data sovereignty	15
3.1.3	Deployment boundary and permitted data flows	16
3.2	Functional requirements	16
3.3	Non-functional requirements	18
3.4	Availability, backup and disaster recovery	18
4	Overview of candidate AI-assisted search solutions	20
4.1	Three deployment patterns	20
4.2	How AI-assisted search works	20
4.3	Shortlisted candidates	21
4.3.1	Sinequa (ChapsVision, France)	21
4.3.2	Mindbreeze InSpire (Austria)	22
4.3.3	IntraFind iFinder + iAssistant (Germany)	22
4.3.4	Sovereign modular stack	22
4.4	Other vendors considered	23
5	Comparative analysis	24
5.1	Architectural comparison	24
5.2	Functional comparison	24
5.3	Security and traceability	25
5.4	Usability and maintainability	25
5.5	Overall advantages, disadvantages and decision implications	26
5.5.1	Procurement questions before full deployment	29
6	Modality load: text, scanned documents, diagrams	30
6.1	Text	30

6.2	Scanned documents and images	30
6.3	Engineering drawings and diagrams	30
6.4	Feasibility summary	31
6.4.1	Recommended pilot scope	32
7	Latest model landscape (May 2026)	33
7.1	Text-only large language models	33
7.2	Multimodal models for scanned documents and diagrams	34
7.3	Embedding models	34
8	On-premise infrastructure and cost model	36
8.1	Reference hardware	36
8.2	Throughput envelope for nuclear-scale corpora	36
8.3	Five-year total cost of ownership	37
8.4	Operational considerations	38
8.4.1	Monitoring and operational response	38
8.5	Cost-model assumptions and uncertainty	39
9	Application stack: knowledge graph and RAG	40
9.1	Why a knowledge graph	40
9.2	Reference architecture	40
9.2.1	Document Lifecycle and index consistency	41
9.3	Application patterns	41
9.3.1	Chat with citations	41
9.3.2	Insight extraction	42
9.3.3	Dashboarding	42
9.4	Curation and governance	42
10	Reference deployments and benchmarks	44
11	Risks and considerations	45
11.1	Finnish language depth	45
11.2	Citation faithfulness	45
11.3	Prompt injection and indirect attacks	45
11.4	Access-control leakage	45
11.5	Query and interaction-data governance	46
11.6	Regulatory exposure	46
11.7	Vendor viability and market consolidation	46
11.8	AI software supply-chain security	46
12	Conclusions and recommendations	48
12.1	Summary of findings	48
12.2	Recommendations	48
12.3	Pilot validation plan	50
13	Glossary	53
14	References	54

1 Introduction

1.1 BACKGROUND AND MOTIVATION

Nordic nuclear power plants and their operating organisations sit on decades of structured and unstructured data: safety analysis reports, maintenance logs, regulatory correspondence, training material, vendor manuals, ageing-management records and design documentation. Much of this content is heterogeneous in format and stored across SharePoint sites, file shares, PDF archives, ERP and maintenance-management systems. Engineers and safety reviewers report that locating the right passage in this corpus – with confidence that the source is current and correctly classified – is time consuming.

This challenge has also been recognised internationally. The International Atomic Energy Agency (IAEA) identifies siloed data and limited knowledge access as important use cases for AI-assisted knowledge management, noting that traditional search leaves much of the available information underused (International Atomic Energy Agency 2024). Similarly, the Nuclear Energy Agency's AIxpertise initiative has identified knowledge management and AI-assisted search as priorities from 2026 onwards (Nuclear Energy Agency, OECD 2025).

This report responds to the Energiforsk request for proposal of 29 October 2025, which set out a need for an AI-assisted search application for internal data, installed on-premise, suitable for the security and confidentiality requirements of the Nordic nuclear sector (Figure 1). The work continues a line of digitalisation studies that VTT has supported for Energiforsk, including the previous *Industrial Internet of Things in Nuclear* report (Laikari et al. 2021) and the *Wireless in Nuclear* feasibility study (Laikari et al. 2018).

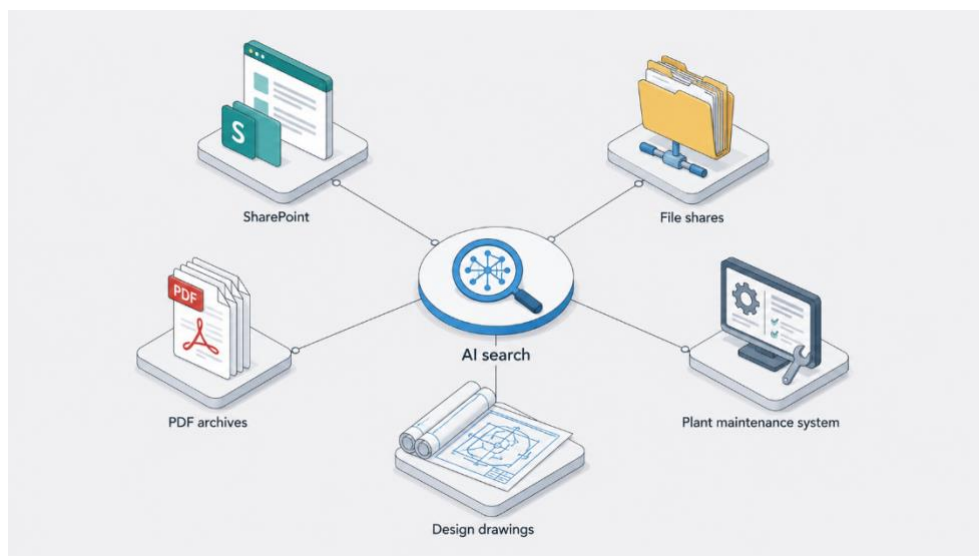


Figure 1 Illustrative scope of internal nuclear documentation that an on-premise AI search system must reach.

1.2 AIMS AND OBJECTIVES

The study addresses the following five questions from the work definition:

1. identify the latest AI technologies for internal search across heterogeneous nuclear data sources, with traceable citations to source documents;
2. assess on-premise suitability under nuclear-industry security and reliability requirements, including compute capacity and concurrent-user behaviour;
3. compare at least two solutions on architecture, functionality, security and usability;
4. estimate installation and annual operating costs, covering licensing, hardware, integration and support;
5. describe the infrastructure, maintenance and skill requirements implied by each option.

Two further questions were added during scoping, in line with the comments raised during the work:

1. the *modality load* on the underlying models – text-only language models for textual content, multimodal models for diagrams and scanned material – and what each implies for compute and operating cost;
2. an application architecture that uses a knowledge graph as the curated knowledge base, with retrieval-augmented generation built on top of it for chat, insight extraction and dashboarding.

1.3 SCOPE AND LIMITATIONS

The study covers AI-assisted search for internal data, not general AI for plant control, predictive maintenance or simulation. The deployment constraint is fully on-premise, with no mandatory cloud connectivity. The analysis is exploratory and decision-supporting; quantitative performance claims – especially Finnish retrieval recalls and citation faithfulness – need to be verified in a pilot. All cost figures are indicative ranges, expressed in euros, and should be treated as planning input rather than tender prices. The report is a technology snapshot based primarily on information available in May 2026. Because AI models, retrieval methods, multimodal capabilities, hardware, licences and commercial products are developing rapidly, some technical assumptions and comparative conclusions may become outdated. They should be revalidated against current evidence before procurement, pilot configuration or production deployment. The vendor shortlist is restricted to European-headquartered organisations to avoid jurisdictional exposure under the United States CLOUD Act (United States Congress 2018), in line with the discussions during the work. Vendors with European headquarters were used as a risk-reduction selection criterion, and legal exposure must be assessed by deployment.

1.3.1 On-premise boundary and cloud-connected sources

A fully on-premise search platform does not automatically make all source systems on-premise. If SharePoint Online, OneDrive or Teams are included, the connector usually requires controlled connectivity to Microsoft cloud APIs or an approved replication process into the secure enclave. For a strictly air-gapped pilot, the initial corpus should therefore be limited to file shares, SharePoint Server, PDF archives and internal databases already inside the protected network.

2 Methodology and approach

2.1 OVERALL APPROACH

The project was structured into four phases: literature and market analysis, technical evaluation and comparison, cost estimation, and recommendations. Material was collected from peer-reviewed literature, vendor documentation, analyst reports, open-source repositories, and the authors' deployment experience at VTT's on-premise GPU cluster and multi-cloud LLM stack.

2.2 SELECTION CRITERIA

A solution had to be deployable fully on-premise (see also boundary clarification in paragraph 1.3.1), return answers with citations back to the source documents, integrate with enterprise identity and access-control systems, handle a mix of file formats including scanned PDFs and engineering drawings, and support Nordic languages. Additional criteria included modular architecture that allows component replacement, support for retrieval-augmented generation (RAG) and graph-based retrieval, and GPU-accelerated inference.

2.3 EVALUATION DIMENSIONS

Each candidate was assessed across four dimensions: technical architecture (deployment, scalability, integration), functionality (retrieval quality, RAG behaviour, language coverage, multimodality), security and traceability (data residency, access enforcement, audit logging, certifications), and maintainability (operational burden, upgrade path, exit strategy, five-year cost of ownership).

2.4 EVIDENCE STATUS

This report distinguishes four levels of evidence: *publicly verified facts* (certifications, regulatory texts, peer-reviewed benchmark results); *analyst and research findings* (e.g. the Forrester Wave Q4 2025 and the Scandinavian Embedding Benchmark); *vendor claims* (connector counts, language coverage, performance figures stated by vendors but not independently verified); and *pilot hypotheses* (assumptions about real-world Finnish retrieval quality, citation faithfulness and access-control behaviour that must be validated empirically). The pilot is the verification mechanism for the third and fourth categories; this desk study is not.

2.4.1 Pilot decision gates

Pilot decision gates. The pilot should not primarily demonstrate that the user interface works; it should establish whether the system can be trusted in a nuclear knowledge-management setting. The decisive tests are Finnish and Swedish retrieval recall; whether each generated answer is supported by its cited source; whether document-level access rights are preserved under realistic user, group and classification changes; whether malicious or misleading content inside indexed documents can manipulate answers or system behaviour; whether deleted,

replaced or reclassified content disappears from all retrieval paths within the agreed period; and whether latency, indexing, availability and audit logging remain acceptable under concurrent use.

2.5 COMPARATIVE ASSESSMENT METHOD

The comparative assessment combines publicly verifiable information, vendor documentation, analyst findings and project-team judgement. Each capability should be labelled according to its evidence status: independently verified, supported by vendor documentation, estimated from comparable deployments, or requiring pilot validation. The comparison is not intended to produce a mathematically definitive ranking; rather, it identifies the trade-offs most relevant to Nordic nuclear operators. Final selection should be based on pilot evidence and formal procurement criteria, particularly retrieval quality, citation-faithfulness, access-control enforcement, operational burden and total cost of ownership.

2.5.1 Decision logic for the pilot recommendation

The recommendation is based on suitability for a time-limited Phase 1 pilot rather than on an assessment of which platform would necessarily be the best final production solution. The decision logic gives priority to:

1. secure and fully on-premise deployment;
2. reliable enforcement of source-document access rights;
3. traceable citations and support for empirical evaluation;
4. deployment effort and operational burden during the pilot;
5. Finnish, Swedish and English retrieval capability;
6. estimated five-year total cost of ownership;
7. extensibility, component replaceability and exit options.

Security, access-control enforcement and citation traceability are treated as threshold requirements. A candidate that does not meet these requirements in the pilot should not proceed, regardless of its aggregate performance in other areas.

Among candidates meeting the threshold requirements, deployment effort and operational simplicity are given relatively high importance for the Phase 1 pilot. This reflects the pilot objective of testing retrieval quality, language performance, citations and access control rather than undertaking a major platform-development programme. Different weighting could lead to a different preferred option.

The recommendation should therefore be understood as a testable hypothesis for the pilot. It does not replace competitive procurement, vendor due diligence, quotation comparison or empirical validation.

3 Requirements for on-premise AI search in the nuclear domain

3.1 NUCLEAR-SPECIFIC CONSTRAINTS

3.1.1 Regulatory environment

The EU AI Act follows a phased implementation timetable. Transparency obligations relevant to interactive AI systems apply from 2 August 2026. Following the political agreement on the EU AI Omnibus, the European Commission currently identifies 2 December 2027 as the application date for high-risk AI systems covered by Annex III and 2 August 2028 for high-risk AI systems integrated into regulated products covered by Annex I. These dates should be monitored until the amended legislation has completed the formal adoption process.

The proposed AI-assisted search system should be classified according to its intended purpose and actual operational use. Deployment within a nuclear organisation does not by itself make a knowledge-management system high-risk. Classification depends, among other things, on whether the system falls within an Annex III use case or is used as a safety component of a regulated product. A system used to retrieve documents and support human review may therefore be treated differently from one that controls or materially influences a safety-related process.

The system should be deployed as a knowledge-navigation and decision-support tool, not as an authoritative record or an autonomous safety decision-maker. It should not approve documents, override operating procedures, determine plant safety status or replace required engineering and regulatory review. For safety-relevant work, users must verify material claims against the cited, approved and currently effective source document. The intended use, prohibited uses, responsible roles and escalation path for incorrect or unsupported answers should be documented before pilot users receive access.

3.1.2 Confidentiality and data sovereignty

All processing, indexing and inference should remain within the operator's own infrastructure. The CLOUD Act attaches jurisdiction to the vendor entity, not to the physical location of the data (United States Congress 2018). The shortlist priorities European-headquartered commercial vendors and a sovereignty-oriented modular stack. For the modular option, component origin, governance, licensing, update mechanisms, telemetry and external dependencies must be assessed individually. European ownership or open-source licensing alone does not eliminate jurisdictional or software-supply-chain risk. Air-gapping is feasible, but cloud-only sources such as SharePoint Online and OneDrive sit outside the boundary and need an explicit policy decision before they are connected.

3.1.3 Deployment boundary and permitted data flows

Before the pilot, the operator should approve a deployment-boundary document defining the network zones for source systems, ingestion, indexes, model serving, administration and user access. It should identify permitted inbound and outbound connections, the method for importing model weights and container images, secret management, administrator access and whether SharePoint Online or other cloud sources are accessed directly or through controlled replication. Model-serving and vector-database hosts should not initiate unrestricted outbound connections. Unknown or unresolved permission states should fail closed.

3.2 FUNCTIONAL REQUIREMENTS

Functional requirements were derived from the work definitions and from prior nuclear AI guidance (Figure 3) (International Atomic Energy Agency 2024). The system must:

1. accept natural-language queries in Finnish, Swedish and English;
2. index Microsoft 365 ^{*)}, SharePoint ^{*)}, file shares, PDF archives (including scanned PDFs) and selected internal databases;
3. run hybrid keyword and vector retrieval, with reranking;
4. enforce document-level access rights at retrieval time;
5. return answers with traceable citations – ideally to page level;
6. process scanned PDFs and image-based content through OCR;
7. read text labels and – where feasible – diagrams and engineering drawings;
8. prevent superseded documents from being cited when an approved and currently effective version contains the relevant information;
9. propagate deletion, replacement and reclassification to all keyword indexes, vector indexes, caches, generated summaries and knowledge-graph entities within an agreed maximum period.

^{*)} Index approved Microsoft 365 sources where controlled cloud connectivity or an approved replication process is permitted; otherwise restrict ingestion to sources inside the protected network.

Finnish-language retrieval is one of the most demanding functional requirements. Finnish is agglutinative: a single noun can take hundreds of inflected forms across cases, numbers and possessive suffixes. Standard sub-word tokenisers split Finnish words at syllable rather than morpheme boundaries, which degrades both indexing and retrieval. Inserting a lemmatiser such as Voikko before embedding generation – on both the indexing and the query side – materially improves recall (Pitkänen 2024). The Scandinavian Embedding Benchmark for Danish, Norwegian and Swedish languages provide the standard evaluation harness for this and related Nordic-language tasks (Enevoldsen et al. 2024).

Swedish generally presents fewer linguistic challenges than Finnish for LLMs and AI-assisted retrieval, because Swedish is a Germanic language that is structurally much closer to English, and there is far more Swedish content available in the training data for most global models. However, its performance still depends on the model, embedding method and document domain, and should therefore be validated using representative Swedish-language queries and documents during the pilot.

The system should distinguish authoritative, approved and currently effective documents from drafts, obsolete versions and uncontrolled copies. Retrieval ranking should take document status and effective date into account, and generated answers should clearly indicate when evidence comes from a draft or superseded source. Deletion, replacement and reclassification of a source document should propagate to keyword indexes, vector indexes, caches, generated summaries and knowledge-graph entities within a defined period. Figure 2 outlines the propagation of document status, permissions and lifecycle changes through the AI-assisted search system.

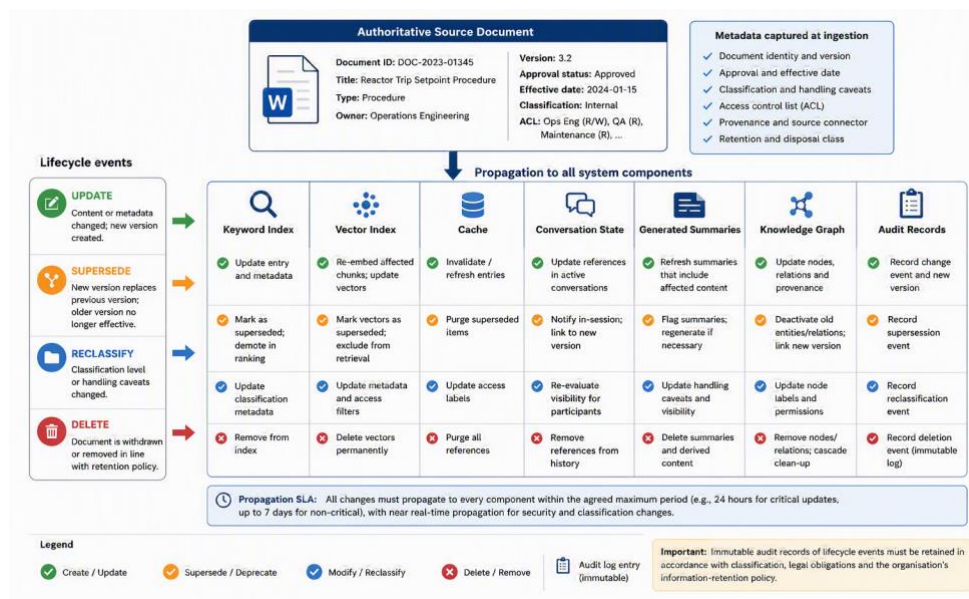


Figure 2 Propagation of document status, permissions and lifecycle changes through the AI-assisted search system.

The system should distinguish between answerable, partially answerable, conflicting and unsupported queries. Where the authorised corpus does not contain sufficient evidence, the system should abstain from generating a substantive answer and state that no adequately supported answer was found. It should not fill evidence gaps from model memory or from content outside the approved retrieval boundary.

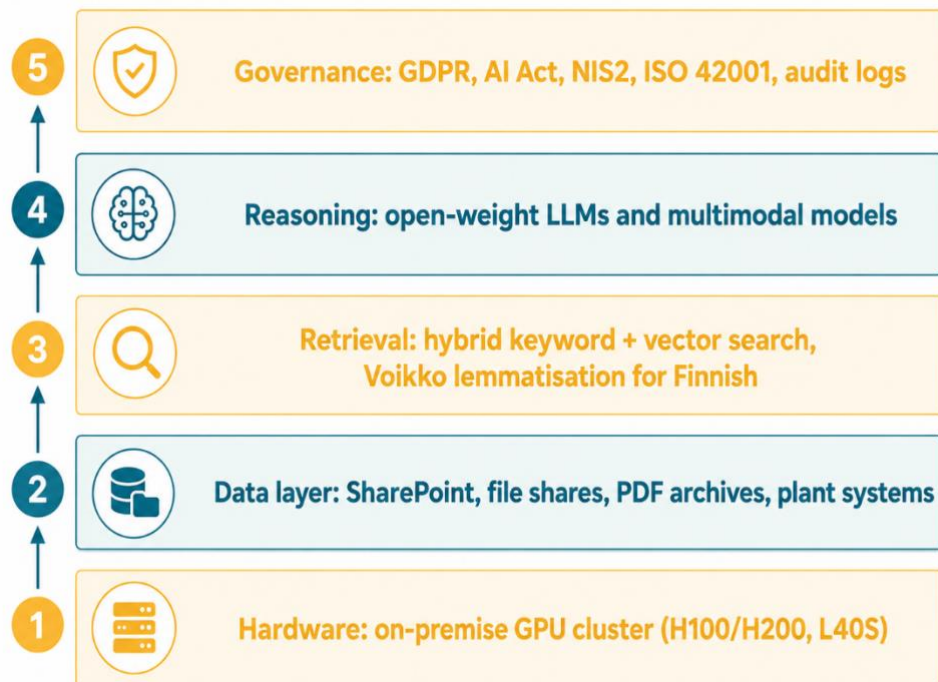


Figure 3 Layered requirements stack for an on-premise nuclear AI search system.

3.3 NON-FUNCTIONAL REQUIREMENTS

As a proposed pilot acceptance criterion, the query latency should remain below 3 seconds at the 95th percentile for a single user, and below 5 seconds at 50 concurrent users. Incremental indexing of 1 000 documents should complete within 15 minutes. Audit-log completeness and document-level access enforcement are non-negotiable at 100%. Comprehensive guidance on AI-specific risk patterns – prompt injection, data poisoning, sensitive-information disclosure – is given by CERT-EU and the OWASP Top 10 for LLM applications (CERT-EU 2024; OWASP Foundation 2025).

3.4 AVAILABILITY, BACKUP AND DISASTER RECOVERY

Production deployment also requires defined availability and recovery objectives. The operator should determine acceptable recovery time and recovery point objectives for source connectors, keyword and vector indexes, the knowledge graph, configuration data and audit logs. The system should provide a degraded mode in which conventional search and source-document access remain available if the generative or vector components are unavailable. Backup restoration, index reconstruction, service failover and preservation of access-control metadata should be verified before production use.

Before a repository is connected, the source owner should complete a corpus-readiness assessment covering document ownership, classification, approval status, version and supersession metadata, access-control quality, duplicate and

uncontrolled copies, scan quality, language distribution and retention obligations. A source should not be admitted to production retrieval where authoritative status or permission mapping cannot be resolved reliably. Corpus remediation and ongoing content stewardship should be budgeted separately from platform implementation, because retrieval quality cannot compensate for missing, contradictory or poorly governed source information.

4 Overview of candidate AI-assisted search solutions

4.1 THREE DEPLOYMENT PATTERNS

The European market offers three patterns for on-premise AI search. The first is a *turnkey enterprise insight engine*: a commercial product such as Mindbreeze InSpire or Sinequa, supplied with pre-built connectors, governance tooling and an administration interface. The second is a *sovereign modular stack*: open-source components – OpenSearch, Qdrant or Weaviate, deepset Haystack, an open-weight large language model – combined into a custom system that the operator owns and maintains end-to-end. The third is a *document-intelligence enrichment layer*: specialist components such as a graph database, an entity extractor or a knowledge-graph indexer that augment either of the above. The first two are the primary focus; the third is treated as an application-stack element in Chapter 9.

4.2 HOW AI-ASSISTED SEARCH WORKS

Figure 4 outlines how an AI-assisted internal search system works.

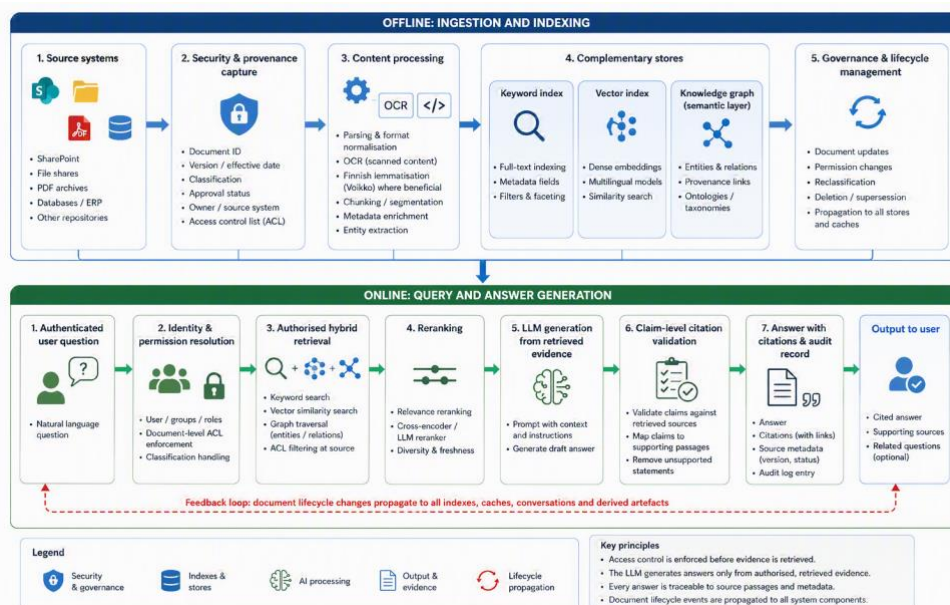


Figure 4 End-to-end workflow for secure AI-assisted search, from document ingestion to a cited answer.

In plain terms, these systems work in two stages. First, they index all connected documents in advance: text is extracted, OCR is applied to scanned pages, chunks of a few hundred words are written to an inverted (keyword) index and converted into vector embeddings for semantic similarity search. Access-control metadata travels with each chunk. Second, when a user submits a query, the system retrieves the most relevant chunks from both indices, reranks them, and passes only those chunks to a language model which drafts an answer. The answer carries citations

to the original passages, so the user can open the source document and verify every statement. Figure 4 outlines the end-to-end workflow for secure AI-assisted search, from document ingestion to a cited answer.

The original retrieval-augmented generation pattern was formalised by Lewis et al. (2020), building on dense passage retrieval work from Karpukhin et al. (2020). Gao et al. (2024) provides a recent survey of the architecture and its variants, including the graph-based extensions discussed in Chapter 9.

In regulated technical environments, provenance is not only a link to a file. Each indexed chunk should carry document identity, version, approval status, effective date, classification, owner and supersession metadata where available. The system should prefer current approved documents and visibly warn when a retrieved passage comes from a draft, obsolete, superseded or uncontrolled copy. This metadata model is a prerequisite for trustworthy citations and for later knowledge-graph curation.

4.3 SHORTLISTED CANDIDATES

Four candidates emerged from the market scan; Table 1 summarises them. The following subsections introduce each candidate and distinguish vendor claims from project-team assessment. Prices are estimates and subjects to change, as they are not based on actual offers.

Table 1 Shortlisted candidate solutions.

Solution	Country	Deployment	Role in evaluation
Sinequa (ChapsVision)	France	On-premise software	Maximum capability, page-level citations
Mindbreeze InSpire	Austria	On-premise appliance	Expected shortest deployment path to production, comparatively low operational burden low TCO
IntraFind iFinder + iAssistant	Germany	On-premise software	Strong sovereignty positioning, cost-optimised commercial option
Sovereign modular stack	Mixed-origin, sovereignty-oriented	Self-managed open-source	Maximum flexibility, no vendor lock-in

4.3.1 Sinequa (ChapsVision, France)

Sinequa is a long-standing enterprise search platform with deep retrieval-augmented generation support and more than 200 connectors. After its 2024 acquisition by ChapsVision it is explicitly positioned as a European sovereign alternative for high-security deployments (ChapsVision 2025). Its early-binding attribute-based access control resolves access rights at index time, so unauthorised documents do not surface even in autocomplete or cached snippets. Citation precision is at page level, which is unmatched in the shortlist. Nordic-language

coverage is provided through multilingual models rather than purpose-built Finnish NLP and needs pilot validation.

4.3.2 Mindbreeze InSpire (Austria)

Mindbreeze ships its platform as a rack-mounted appliance: the operator racks the hardware, configures data sources, and goes live (Mindbreeze GmbH 2025). The appliance model removes most of the integration burden and is the principal reason why Mindbreeze led the Forrester Wave Cognitive Search Platforms ranking for Q4 2025 on current offering (Forrester Research 2025). Pricing is by indexed document volume rather than seats, so cost is predictable. Advanced NLP features – entity recognition and compound splitting – are currently available for German and English only, which is the most material limitation for Nordic deployment. Certifications cover BSI C5 (Bundesamt für Sicherheit in der Informationstechnik 2024), SOC 2 Type II and ISO/IEC 27001.

4.3.3 IntraFind iFinder + iAssistant (Germany)

IntraFind has more than two decades of enterprise-search experience and positions itself on data sovereignty. Its iAssistant product, launched in October 2024, delivers RAG with citations over rights-checked content (IntraFind Software AG 2024). The platform uses OpenSearch (OpenSearch Project 2025) as its backend, supports more than 80 connectors and over 600 file formats, and includes a GDPR module that scans indexed content for personal data. Reference customers include the German federal authority for radioactive waste disposal (BGE) and the Bundeswehr. Annual cost is the lowest of the commercial options at roughly 50–200 kEUR per year.

4.3.4 Sovereign modular stack

The sovereignty-oriented stack assembles European-led open-source components: OpenSearch for keyword search, Qdrant (Qdrant Solutions GmbH 2025) or Weaviate (Weaviate B.V. 2025) for vector storage, deepset Haystack for orchestration (deepset GmbH 2025), and an open-weight LLM such as Mistral Large 3 (Mistral AI 2025) or Qwen 3.5 (Qwen Team, Alibaba Cloud 2025) for generation. It should be noted that Qwen 3.5 is not European. Finnish preprocessing on the pilot could rely on Voikko (Pitkänen 2024), with embeddings from multilingual-E5 or BGE-M3 (Wang et al. 2024; J. Chen et al. 2024), both of which lead the Scandinavian Embedding Benchmark (Enevoldsen et al. 2024). The stack has the strongest long-term flexibility but needs an engineering team of five to ten over six to twelve months to reach enterprise grade, with ongoing operations of two to four full-time equivalents.

In the sovereign stack, vector storage must be treated as sensitive infrastructure because embeddings and retrieved chunks can leak information even when original files remain protected. Self-managed vector databases should be isolated by network segmentation, protected by TLS and authentication gateways, and included in continuous configuration compliance checks. The same access-control model used for keyword search must be enforced before vector results are passed to the language model.

4.4 OTHER VENDORS CONSIDERED

Squirro (Switzerland) is notable for the combination of RAG with knowledge graphs and serves the European Central Bank, the Bank of England and the Deutsche Bundesbank (Squirro 2025); it is most suited to financial-services compliance and is treated as a watch item rather than a primary recommendation. Ayfie (Norway) is the only shortlist-adjacent vendor with explicit Nordic-language specialisation but carries a going-concern warning in its 2024 annual report, which removes it from a primary-solution role. US-headquartered offerings (Microsoft Azure AI Search, Google, AWS Kendra, IBM Watson Discovery, OpenText IDOL, Glean) were considered and excluded on earlier mentioned risk-reduction selection criterion to target European options.

5 Comparative analysis

5.1 ARCHITECTURAL COMPARISON

Table 2 summarises the architectural differences. Sinequa and Mindbreeze represent the enterprise-platform end of the spectrum, with mature pre-packaged solutions and large connector libraries. IntraFind occupies a middle position with a German-owned platform built on OpenSearch. The sovereign stack offers the greatest flexibility but at the cost of substantially higher engineering effort.

Table 2 Architectural comparison of candidate solutions.

Dimension	Sinequa	Mindbreeze	IntraFind	Sovereign stack
Vendor headquarters/component origin	France	Austria	Germany	Mixed-origin, sovereignty-oriented
Deployment	On-prem. software	On-prem. appliance	On-prem. software	Self-managed open-source
Document-level security	Early-binding ABAC	ACL-based	RBAC	Custom implementation
Connectors	200+ native	50+ native	80+ native	Custom assembly
Microsoft 365 integration	Native	Native	Native	Custom development
Scalability	400M+ documents	Appliance scaling	Enterprise scale	Internet-scale possible
Deployment effort	Weeks to months	Days to weeks	Weeks	materially longer than a configured commercial-platform pilot, can be up to 6–12 months
GPU requirement	Optional	GPU-ready appliance	Optional	Required for LLM
Lock-in risk	Medium	Medium-high	Low-medium	Very low

5.2 FUNCTIONAL COMPARISON

Table 3 compares functional capability. The most consequential differences are citation precision and Nordic-language depth. Sinequa returns citations at page level; the other commercial solutions are at source-document level; the sovereign stack returns chunk-level citations whose precision depends on chunk size. None of the commercial vendors provided sufficient public evidence for the production-

grade Finnish NLP today, which is the single most important pilot acceptance criterion.

Table 3 Functional comparison of candidate solutions.

Dimension	Sinequa	Mindbreeze	IntraFind	Sovereign stack
RAG quality	Best-in-class	Strong	Strong	Customisable
Citation level	Page	Source	Document	Chunk (configurable)
Finnish support	Multilingual model	Listed; depth unclear	Not documented in detail	Depends on model. E.g. multilingual-E5 / BGE-M3
Customisability	High (APIs)	Medium (appliance)	Medium	Highest
Knowledge-graph extension	Via ChapsVision modules	Limited	Limited	Via Neo4j / GraphDB
Visual/multimodal retrieval	Built-in multimodal search and image-to-text processing; nuclear drawing retrieval requires pilot validation	Built-in multimodal processing and semantic search over images and multimedia; nuclear drawing retrieval requires pilot validation	Built-in OCR and image/document processing supported; public evidence for visual semantic retrieval is limited	Configurable using multimodal embeddings or a vision-language model with a vector database; custom implementation and pilot validation required

5.3 SECURITY AND TRACEABILITY

All four shortlisted options support fully on-premise deployment, and European ownership reduces specified jurisdictional risk with the CLOUD Act jurisdiction by being European-headquartered. The principal differentiator is access-control enforcement. Early-binding ABAC (Sinequa) resolves access rights at index time so unauthorised documents do not surface in any retrieval path. ACL-based and RBAC models (Mindbreeze and IntraFind) enforce rights at query time. The sovereign stack inherits the properties of its components; Qdrant in particular ships with explicit warnings that self-deployed instances are not secure by default and require hardening (Qdrant Solutions GmbH 2025).

5.4 USABILITY AND MAINTAINABILITY

The trade-off between deployment simplicity and long-term flexibility is the defining choice. Mindbreeze's appliance model needs roughly half a full-time

equivalent person to operate and reaches production in days or weeks. Sinequa requires about one full-time equivalent (FTE) and weeks to months. IntraFind sits in between. The sovereign stack needs 2–4 FTE on an ongoing basis and materially longer build-up time than a configured commercial-platform pilot. Build-up time is dependent on connector scope, security integration, evaluation tooling and available engineering capacity. It can be even up to six to twelve months. None of these figures include the content curation effort, which is a separate budget line and is discussed in Chapter 9. Convenience and product maturity versus control, customisability and engineering responsibility of the shortlisted candidates are outlined in Figure 5.

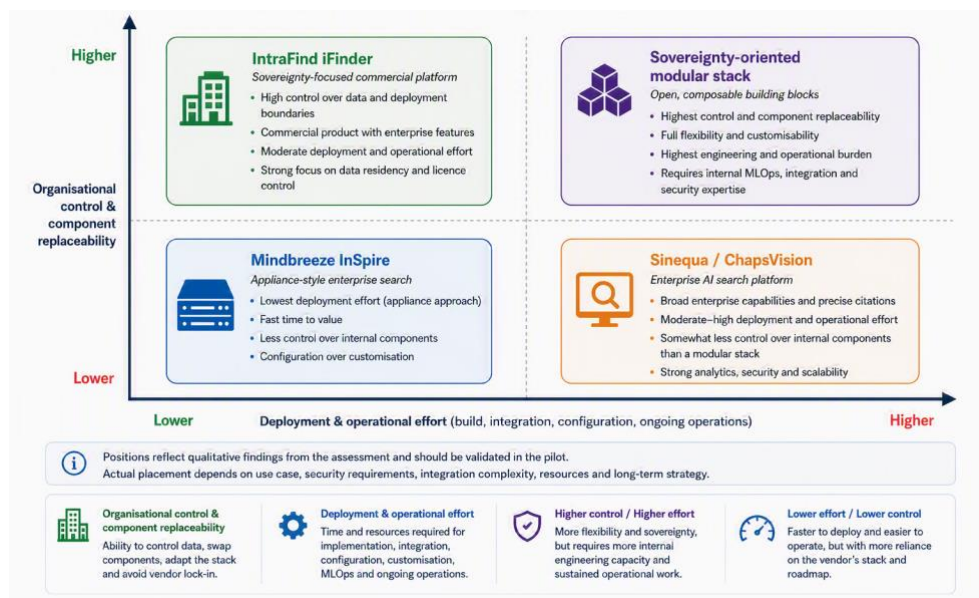


Figure 5 Strategic positioning of the shortlisted solution routes by deployment effort and degree of organisational control.

5.5 OVERALL ADVANTAGES, DISADVANTAGES AND DECISION IMPLICATIONS

Table 4 consolidates the principal trade-offs identified in the architectural, functional, security, maintainability and cost analyses. The assessment is based on publicly available information, vendor documentation and indicative project assumptions. Finnish and Swedish retrieval quality, citation faithfulness, access-control behaviour and actual implementation effort remain pilot-validation questions for all candidates.

Table 4 Summary of principal advantages, disadvantages and uncertainties and decision implications.

Option	Principal advantages	Principal disadvantages and uncertainties	Most appropriate when	Pilot implication
Mindbreeze InSpire	Appliance-based deployment reduces infrastructure assembly and integration work. Estimated operational staffing is low relative to the other options. Commercial cost is predictable in the report's indicative model. The platform has relevant security certifications and analyst recognition.	Finnish-language depth is unclear and advanced NLP is described mainly for German and English. Citation precision appears to be at source level rather than page level. The appliance approach provides less component-level flexibility and creates medium-high lock-in risk. Licensing, hardware configuration, telemetry, air-gap operation and exportability require vendor confirmation.	The priority is a controlled and comparatively rapid Phase 1 pilot with limited internal platform-engineering capacity.	Preferred initial pilot candidate, subject to satisfactory evidence on air-gap operation, access control, Finnish and Swedish retrieval, citation faithfulness, lifecycle handling and commercial terms.
Sinequa / ChapsVision	Broad enterprise-search capability, extensive connector coverage, early-binding attribute-based access control and page-level citations. Strong fit for complex information environments and potentially multi-classification content.	Highest estimated commercial cost in the full-deployment comparison. Deployment and operation are expected to require more specialist effort than the appliance route. Nordic-language retrieval still requires pilot validation.	Citation precision, complex security structures and broad enterprise integration are more important than minimising cost or deployment effort.	Strong benchmark candidate for comparative testing, especially where page-level citation precision and access-control complexity are decisive.

Option	Principal advantages	Principal disadvantages and uncertainties	Most appropriate when	Pilot implication
IntraFind iFinder + iAssistant	Lowest estimated commercial TCO in the report. Strong data-sovereignty positioning, OpenSearch-based architecture and relevant public-sector references. Reported support for many connectors and file formats.	Finnish support is not publicly documented in sufficient detail. Citation capability appears less precise than Sinequa's. Access-control and multimodal behaviour require validation in the intended environment.	Budget, European ownership and a relatively open commercial architecture receive high priority.	Include in the request-for-information or quotation stage and consider as a cost-optimised pilot candidate if the vendor can provide convincing Nordic-language and security evidence.
Sovereign modular stack	Highest component-level control and customisability, very low conventional vendor lock-in and the greatest freedom to select Finnish preprocessing, embeddings, models, vector stores and knowledge-graph components.	Highest implementation risk and operational burden. Access control, lifecycle consistency, auditability, security hardening and enterprise connectors become the operator's responsibility. The report's estimate indicates materially greater staffing needs and a longer build-up period.	The operator has sustained platform-engineering capacity and regards component sovereignty and replaceability as strategic requirements.	Treat as a separate technical prototype or longer-term development track, rather than requiring it to meet the same implementation schedule as configured commercial products.

No candidate is superior on every criterion. Mindbreeze is favoured only under the Phase 1 assumption that rapid controlled deployment and low operational burden have high decision weight. Sinequa becomes more attractive when citation granularity and complex access-control requirements dominate. IntraFind becomes more attractive when commercial cost and European data-sovereignty positioning dominate. The sovereign modular stack becomes more attractive when long-term component control and replaceability outweigh implementation speed and staffing requirements.

Consequently, the preferred pilot route should be conditional on the operator's priorities and on the evidence obtained through procurement dialogue and pilot testing.

5.5.1 Procurement questions before full deployment

Before procurement, require each vendor to document: on-premise and air-gap operation mode; telemetry and update mechanisms; source-code or escrow position where relevant; exportability of indexes, metadata, audit logs and evaluation datasets; supported identity providers and ACL edge cases; page- or chunk-level citation capability; Finnish and Swedish language evidence; security certifications; support access model; and lifecycle terms for GPU, model and software upgrades.

6 Modality load: text, scanned documents, diagrams

This chapter examines how different content types affect the computational burden placed on underlying models. In short, text is inexpensive to process, whereas scanned PDFs and engineering drawings are not (Figure 6). The difference in cost is substantial enough to influence several key architectural decisions. These are illustrative configurations and internal assumptions and not yet detailed plans.

6.1 TEXT

For ordinary text – policies, procedures, reports, e-mail threads, structured data exported as text – a moderately sized open-weight language model is sufficient. A model in the 7–14 billion parameter range with 4-bit quantisation runs on a single L40S 48 GB or A100 40 GB GPU and returns answers within typical interactive latency. A 32-billion-parameter model such as Qwen 3.5-32B raises answer quality, runs on a pair of A100 80 GB cards, and still leaves headroom for embedding and reranking. For demanding Finnish or Swedish queries, the dominant constraint is not parameter count but the embedding model and the lemmatisation step. The Scandinavian Embedding Benchmark consistently shows that English-benchmark performance does not transfer cleanly to Nordic-language (Danish, Swedish and Norwegian) retrieval (Enevoldsen et al. 2024); the practical implication is that embedding-model choice and Voikko preprocessing matter more than scaling up to the largest LLM available.

6.2 SCANNED DOCUMENTS AND IMAGES

A substantial share of nuclear archives is held as image-based scanned PDFs with no embedded text layer. Two routes coexist. The traditional route applies OCR – Tesseract, ABBYY FineReader, PaddleOCR – as a preprocessing step and then treats the recovered text as plain text. The newer route uses a multimodal language model that ingests page images directly. Either route works; the choice depends on accuracy on Finnish typography and on infrastructure budget. Multimodal models such as Qwen3-VL (Qwen Team, Alibaba Cloud 2026) and InternVL 2.5 (Z. Chen et al. 2024) have widened OCR coverage to more than 30 languages and are robust to low-light, blurred and tilted scans. DeepSeek-OCR demonstrates a different angle: representing dense document text as compressed visual tokens reduces context-window consumption substantially, which lowers the cost of indexing large archives. The downside of multimodal routes is that they are typically 5–10 times more expensive in GPU time per page than text-only OCR followed by a text LLM.

6.3 ENGINEERING DRAWINGS AND DIAGRAMS

Engineering drawings – piping and instrumentation diagrams (P&IDs), isometric drawings, electrical schematics, equipment layouts – need two distinct capabilities. The first is text-in-image extraction: reading the labels, dimensions, component identifiers and callouts on a drawing. This is a layout-aware OCR task and is

achievable with current tooling, though accuracy on dense technical notation is lower than on standard documents. The second is visual semantic search: locating a diagram by describing it in words, such as “coolant loop schematic” or “valve arrangement for reactor inlet”. Multimodal embedding models – CLIP, SigLIP and the embedding side of Qwen3-VL or InternVL 2.5 – can place both images and text in a shared semantic space. Accuracy on natural photographs is now high; accuracy on engineering drawings remains materially lower and is the principal reason why we recommend treating visual semantic search as a Phase 2 capability rather than a Phase 1 commitment.

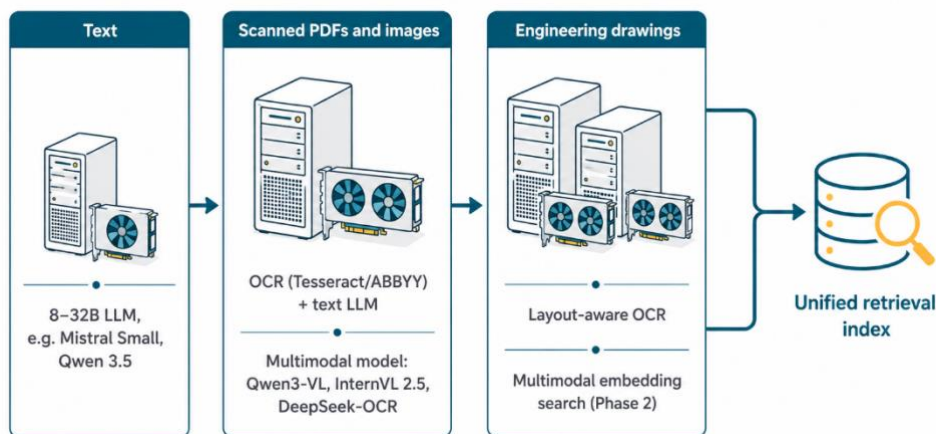


Figure 6 Modality-driven model selection for the on-premise pipeline.

6.4 FEASIBILITY SUMMARY

Implementing text-only RAG with citations is well within current open-source and commercial tooling, with mature reference deployments in adjacent regulated industries. Idaho National Laboratory (2024) demonstrate this for nuclear documentation at Idaho National Laboratory, and American Nuclear Society Newswire (2025) report a joint INL benchmark suite for evaluating RAG and LLMs specifically for nuclear use cases.

Multimodal extraction from scanned documents is feasible and improving rapidly, with production-grade open-weight models available under permissive licences. The infrastructure implication is estimated to be 5–10× more GPU time per page; this should be sized in the cost model.

Publicly available evidence is currently insufficient to demonstrate dependable visual semantic retrieval from nuclear engineering drawings with their specialised typography and high information density. Phase 1 should remain on text-in-image extraction; visual semantic search should be revisited in Phase 2 with empirical validation against the operator’s own drawing corpus.

6.4.1 Recommended pilot scope

Phase 1 should focus on text documents, searchable PDFs, scanned PDFs after OCR, Finnish/Swedish retrieval, citations and access control. Phase 2 should add multimodal interpretation of engineering drawings, visual similarity search, diagram-level entity extraction and deeper knowledge-graph curation. This avoids overloading the pilot with the most expensive and least mature modality before the basic trust requirements are proven.

Before visual semantic retrieval is promoted beyond Phase 2 research, it should be evaluated on a labelled drawing corpus containing P&IDs, electrical schematics, equipment layouts and revised drawing versions. The evaluation should distinguish text-label recognition, drawing-type classification, visual-similarity retrieval and component-level extraction. Results from photographs or general document-image benchmarks should not be treated as evidence of engineering-drawing performance.

7 Latest model landscape (May 2026)

The open-weight model landscape moved decisively during 2025 and early 2026 (Table 5). This chapter gives the operator a short, current map of what is available and where each option sits in terms of licence and Nordic-language coverage. The AI model landscape is developing rapidly, and the selections presented in this chapter reflect the information available in May 2026. New model releases, licence changes and performance improvements may alter their relative suitability for on-premise deployment.

7.1 TEXT-ONLY LARGE LANGUAGE MODELS

Mistral Large 3 is a sparse mixture-of-experts model with 675 billion total parameters and roughly 41 billion active parameters, trained on 3 000 NVIDIA H200 GPUs and released under the Apache 2.0 licence (Mistral AI 2025). It is among the strongest European open-weight options for general use, with on-premise serving through vLLM, TGI or Ollama. Nordic languages are present in training data but are not the optimisation target.

Meta released the Llama 4 family with two-tier variants – Scout and Maverick – supporting context windows of 128k tokens and beyond and demonstrating strong general performance (Meta Platforms 2025). The licence retains the custom Meta terms with the 700M-MAU clause, which restricts redistribution but not on-premise enterprise use; the licence should be reviewed by legal before deployment. Llama is natively multimodal, but in this category the comparison focuses on its text generation use.

Qwen 3.5 ships under Apache 2.0 with dense and mixture-of-experts variants up to roughly 397 billion total parameters with 17 billion active (Qwen Team, Alibaba Cloud 2025). It currently leads the open-weight space on graduate-level reasoning benchmarks and supports the broadest multilingual coverage of any open-weight family, which is directly relevant for Nordic deployments.

DeepSeek-V3 and the V3.x line are released under the MIT licence and provide one of the strongest cost-per-token ratios on self-hosted inference (DeepSeek-AI 2025). The newer V4 generation extends context to 1 million tokens and is the strongest open model on coding and reasoning benchmarks as of released preview in April 2026.

Google's Gemma 4 sits in the 2–31 billion parameter range under a permissive licence, optimised for single-GPU inference (Google DeepMind 2025). It is the most efficient option for constrained-hardware deployments where a 70 B-class model is not feasible. Gemma is natively multimodal, but in this category the comparison focuses on its text generation use.

For Finnish specifically, the EuroLLM-22B research line is the most promising European sovereign option (Martins, Fernandes, et al. 2024); the consortium-funded OpenEuroLLM programme is expected to deliver production-ready models covering all EU languages during 2026–2027. The Viking 7B/13B/33B line

from Silo AI / AMD (Silo AI / AMD 2024) was trained on the LUMI supercomputer (CSC – IT Center for Science 2024) on roughly 2 trillion tokens including Finnish, Swedish, Norwegian, Danish and Icelandic. As of May 2026, Viking remains a set of research checkpoints with a 4 096-token context window, which is too short for RAG; we recommend monitoring it for a future Phase 2 upgrade rather than relying on it in Phase 1.

Table 5 Selected open-weight text LLMs available for on-premise deployment, May 2026.

Model family	Open licence	Active params	Context	Notes for Nordic deployments
Mistral Large 3	Apache 2.0	41 B / 675 B total	128k	European; multilingual coverage; strong reasoning
Llama 4 (Maverick/Scout)	Meta custom	17 B / 400 B total	128k–10M	Long-context leader; licence review needed
Qwen 3.5	Apache 2.0	17 B / 397 B total	256k	Strongest multilingual; instruction tuned
DeepSeek-V3.x / V4	MIT	37 B / 671 B total	up to 1M	Best cost per token; coding strong
Gemma 4	Apache 2.0	2–31 B dense	256k	Single-GPU friendly; permissive; Multimodal; efficient local deployment
EuroLLM-22B	Open weight	22 B dense	32k	All 24 EU languages; research-grade
Viking 7/13/33B	Apache 2.0	7–33 B dense	4k (research)	Best Finnish coverage; not production-ready

7.2 MULTIMODAL MODELS FOR SCANNED DOCUMENTS AND DIAGRAMS

Qwen3-VL extends the Qwen series into vision, with OCR over 32 languages and robustness to low-light, blurred and tilted scans (Qwen Team, Alibaba Cloud 2026). Variants range from 8 to 72 billion parameters with the same Apache 2.0 licence as the text models. InternVL 2.5 spans 1 to 78 billion parameters under Apache 2.0 and is competitive with proprietary counterparts on document and chart understanding (Z. Chen et al. 2024). Meta’s Llama 3.2 Vision provides general image understanding and integrates cleanly with the rest of the Llama ecosystem (Meta Platforms 2025). DeepSeek-OCR is the latest open vision model from DeepSeek-AI and is purpose-designed for dense-document OCR with substantial token compression; its open weights make it a natural Phase 1 component for scanned archives.

7.3 EMBEDDING MODELS

For retrieval, multilingual-E5 and BGE-M3 are relevant multilingual candidates: both have been evaluated on Finnish as part of the MIRACL benchmark, while multilingual-E5 also performs strongly on the Scandinavian Embedding Benchmark. However, SEB covers Danish, Swedish and Norwegian rather than Finnish. Finnish retrieval should therefore be evaluated separately using

representative Finnish-language documents and queries from the intended deployment domain (Wang et al. 2024; J. Chen et al. 2024; Enevoldsen et al. 2024). LaBSE, despite a strong multilingual reputation, performs poorly on the fine-grained discrimination needed for retrieval and should be avoided.

8 On-premise infrastructure and cost model

8.1 REFERENCE HARDWARE

Three illustrative GPU tiers are used for comparative planning for the on-premise nuclear-search scenarios (Table 6). Specifications and list pricing in May 2026 are summarised below; bulk procurement and integrator margins shift these by 10–20%. See also 8.5 for uncertainties related to the pricing.

Table 6 Reference GPU tiers for on-premise nuclear AI search (May 2026 indicative pricing).

GPU	Memory	Approx. unit price	Power	Best fit
NVIDIA L40S	48 GB	8–11 kEUR	350 W	Pilot tier, single-GPU inference of 7–14B models, embedding nodes (NVIDIA Corporation 2024b)
NVIDIA H100 80 GB	80 GB	25–35 kEUR	700 W	Department-tier inference of 32–70B models, OCR pipelines (NVIDIA Corporation 2024a)
NVIDIA H200 141 GB	141 GB	30–40 kEUR	700 W	Enterprise-tier inference of 70–400B-class MoE models, multimodal pipelines (NVIDIA Corporation 2025)

The H200 can provide higher inference throughput than the H100 for memory-intensive large-model workloads because of its larger and faster memory subsystem. The actual gain depends on the model, numerical precision, context length, batch size, serving engine and workload mix. Hardware selection should therefore be based on a reproducible benchmark using the intended pilot workload. For RAG applications, concurrent-request handling, KV-cache capacity, embedding placement and retrieval latency may be more important than peak generation throughput alone.

8.2 THROUGHPUT ENVELOPE FOR NUCLEAR-SCALE CORPORA

Table 7 gives realistic envelopes derived from the GPU specifications above and from typical batched-serving benchmarks. Numbers should be read as order-of-magnitude planning input; vendor-specific benchmarks deliver tighter figures.

Table 7 Indicative retrieval and generation envelopes per inference node, May 2026.

Tier	GPU config	Model class	Tokens/s (peak)	Concurrent users at 95 th -percentile latency ≤ 5 s
Pilot	1×L40S 48 GB	8–14B text LLM	1 500–2 500	5–10

Tier	GPU config	Model class	Tokens/s (peak)	Concurrent users at 95 th -percentile latency ≤ 5 s
Department	2×H100 80 GB	32B text + 8B vision	6 000–10 000	30–60
Enterprise	4×H200 141 GB	70–400B MoE + 32B VL	18 000–30 000	100–250

The concurrency values in Table 7 are initial capacity assumptions rather than acceptance commitments. Hardware sizing should therefore follow a reproducible load test using the selected model, representative prompt and output lengths, and the intended retrieval pipeline.

8.3 FIVE-YEAR TOTAL COST OF OWNERSHIP

Table 8 and Table 9 present a five-year cost of ownership comparison across the four shortlisted solutions, at two scales: a pilot enclave covering one business unit and a full deployment of roughly 1 000 users. Numbers carry the usual uncertainty of indicative pricing and are intended to support an initial budget conversation rather than to substitute for vendor quotations.

Table 8 Five-year cost of ownership: pilot enclave (about 100 users, single business unit, kEUR).

Cost component	Mindbreeze	Sinequa	IntraFind	Sovereign stack
Platform licensing	175	250	100	100
GPU hardware (1×L40S)	11	11	11	11
Server infrastructure	included	30	included	25
Implementation and integration	40	60	45	90
Operational staffing (5 years, 0.25–0.5 FTE)	80	160	125	320
Power and cooling (5 years)	15	20	15	18
Indicative five-year total	321	531	296	564

Table 9 Five-year cost of ownership: full deployment (about 1 000 users, multi-TB corpus, kEUR).

Cost component	Mindbreeze	Sinequa	IntraFind	Sovereign stack
Platform licensing	950	1 400	380	380
GPU hardware (purchase)	80	80	80	80
Server infrastructure	included	180	included	100
NVIDIA AI Enterprise (5 years)	80	80	80	80
Implementation and integration	135	180	150	270
Staffing (5 years)	320 (0.5 FTE)	640 (1.0 FTE)	500 (0.75 FTE)	1 280 (2.0 FTE)

Cost component	Mindbreeze	Sinequa	IntraFind	Sovereign stack
Power, cooling, DC (5 years)	68	90	75	80
Hardware refresh (year 4)	55	70	65	75
Indicative five-year total	1 688	2 720	1 330	2 345

At pilot scale, IntraFind and Mindbreeze sit close to each other and well below the other two options. At full-scale deployment, the sovereign stack's apparent licensing advantage is absorbed by a roughly fourfold increase in operational staffing relative to Mindbreeze. IntraFind remains the lowest-cost commercial option at full scale.

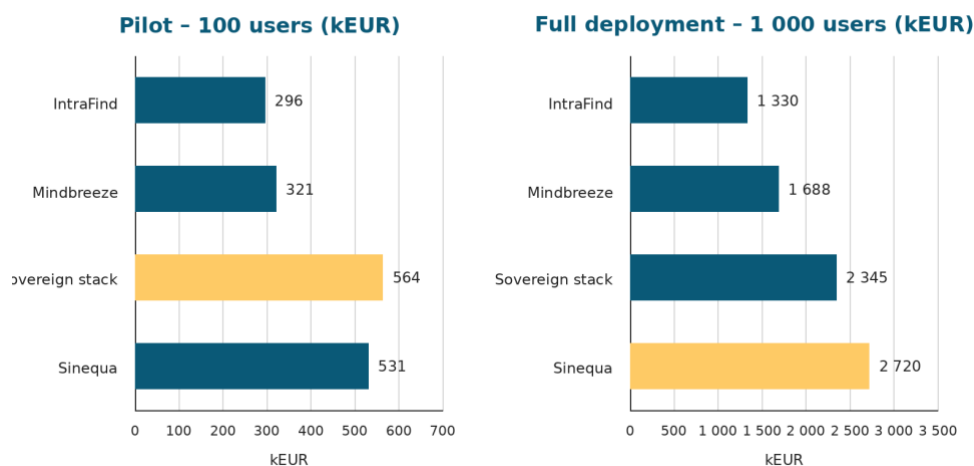


Figure 7 Five-year cost-of-ownership comparison at pilot and full-deployment scales.

8.4 OPERATIONAL CONSIDERATIONS

Power and cooling for a four-GPU H100 or H200 server reach 12–20 kEUR per year in electricity alone, depending on local tariffs. NVIDIA AI Enterprise adds roughly 4.5 kEUR per GPU per year for production NIM deployment. Server-room footprint is modest – a single 4U chassis covers most pilot scenarios – but airflow and rack-PDU capacity should be verified before procurement.

8.4.1 Monitoring and operational response

Production operation requires end-to-end observability across ingestion, permission resolution, keyword and vector indexes, model serving and answer generation. Monitoring should cover failed or quarantined documents, stale ACLs, index divergence, deletion propagation time, query latency, GPU utilisation, unsupported-claim rate and citation-validation failures. Every material change to

the model, embedding model, OCR engine, chunking rules, prompts or retrieval parameters should be versioned, regression-tested and capable of rollback. The operator should define incident-response procedures for suspected access leakage, poisoned content, incorrect citations, failed updates and compromise of imported model or container artefacts.

8.5 COST-MODEL ASSUMPTIONS AND UNCERTAINTY

The cost estimates are intended for comparative planning and are not substitutes for formal quotations. Hardware prices, licence terms, staffing rates and energy costs may change materially between pilot planning and procurement, even during the production of this report. The estimates should therefore be interpreted as indicative range rather than point forecasts. Before procurement, the model should be updated using vendor quotations and organisation-specific assumptions for labour cost, electricity, cooling, infrastructure redundancy, storage, backup, integration, security assurance, model evaluation and content curation. Knowledge-graph construction and ongoing semantic curation should be reported separately because their cost depends strongly on corpus quality, ontology scope and the level of human review required.

Source status in Table 6 to Table 9: hardware memory and power values are manufacturer specifications; prices, throughput, concurrency, staffing and licence costs are indicative project assumptions at the time of the report writing.

9 Application stack: knowledge graph and RAG

This chapter describes a target architecture centred on a curated knowledge graph that functions as a semantic and relationship layer complementing the keyword and vector indexes. Retrieval-augmented generation pipelines operate across these components to support chat, insight extraction and dashboarding. The architecture supports chat, insight extraction, and dashboarding while leveraging, rather than replicating, the capabilities of the underlying search engine.

9.1 WHY A KNOWLEDGE GRAPH

Plain RAG works well for question answering over a flat document corpus, but it struggles with multi-hop questions (“Show me all maintenance records for valves linked to coolant loop B, in the last five years, where the supplier is still active”) and with sensemaking across entities (“How does the issue raised in this safety report relate to incidents at peer plants?”). A knowledge graph stores entities – components, locations, suppliers, incidents, regulations – and the relationships between them. Hogan et al. (2021) give a comprehensive overview of the underlying technology. Edge et al. (2024) show that combining knowledge-graph reasoning with RAG outperforms baseline RAG on complex sensemaking tasks by a substantial margin; Peng et al. (2024) survey the broader literature.

The recommended approach is to treat the knowledge graph as the operator’s curated *semantic layer*, populated from authoritative internal sources – asset hierarchies, document metadata, organisational charts, incident records – and from entities extracted from indexed documents. The graph database does not replace the search engine; it augments it. Production-grade integrations between knowledge graphs and RAG are now mature, with Neo4j providing reference implementations using Microsoft GraphRAG and LangChain (Neo4j 2025).

9.2 REFERENCE ARCHITECTURE

Figure 8 sketches the proposed stack. From bottom to top, the data-source layer draws content from the operator’s SharePoint, file shares, PDF archives and selected plant or business systems. The ingestion layer then normalises the content: documents are parsed, scanned material is OCR-processed, Finnish text is lemmatised with Voikko, and relevant entities and metadata are extracted. The storage layer keeps the processed material in three complementary forms: a keyword index for exact and lexical search, a vector index for semantic similarity search, and a knowledge graph for curated entities, relationships and provenance. The retrieval layer queries these stores together by combining keyword search, vector search and graph traversal; a reranker then reorders the candidate passages before they are passed onwards. The reasoning layer hosts the open-weight text and multimodal models discussed in Chapter 7 and uses the retrieved evidence to generate grounded outputs. Three application surfaces sit on top: a conversational assistant for natural-language questioning with citations, an insight-extraction

service that produces structured summaries and trend analyses, and a dashboarding service that publishes graph-derived indicators such as incident counts, supplier exposure and regulatory deadlines.

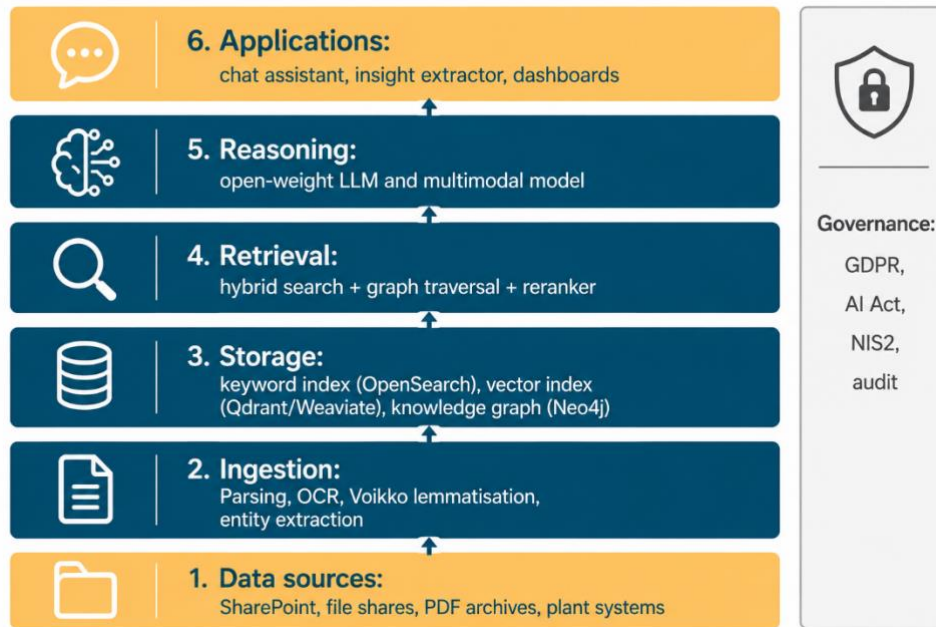


Figure 8 Reference application architecture: a knowledge graph as the curated semantic layer, with hybrid retrieval and RAG as the access pattern.

9.2.1 Document Lifecycle and index consistency

A production system should distinguish the logical source document from its individual revisions. Every indexed revision should carry a content hash, source version, approval status, effective period, classification, ACL version and the versions of the extraction, chunking and embedding components used. A revision should become searchable only after its keyword entries, vector entries, metadata and provenance records have been written and verified. Partial or failed indexing must not expose an incomplete revision. Deletion, replacement, reclassification and permission changes should be propagated to all retrieval stores, caches, conversation state and derived knowledge-graph entities within an agreed maximum period. Periodic reconciliation against the source system should complement event- or delta-based updates.

9.3 APPLICATION PATTERNS

Three patterns cover most expected use cases.

9.3.1 Chat with citations

A user submits a natural-language question in Finnish, Swedish or English. The retrieval layer runs a hybrid query and, if the question is multi-hop, expands it through the knowledge graph (for example, from the named component to its

parent system, suppliers and related incidents). The reasoning layer returns passage-, chunk-, page- or source-document-level citations according to the capability validated for the selected platform. The target is the most precise citation level that remains stable, traceable and directly verifiable against the original source.

9.3.2 Insight extraction

Periodic batches over the corpus generate structured outputs: trend summaries of maintenance findings by system, regulatory-deadline calendars, supplier-risk maps. The knowledge graph provides the joins that would otherwise need to be written by hand against each source system. Results are stored back into the graph and exposed to dashboards. These results require review/status labels before derived outputs are entered in the curated graph.

9.3.3 Dashboarding

A small set of dashboards reads from the graph – not from the unstructured corpus directly – and presents indicators that the operator already tracks: incident counts by category, document review status, training-record completeness, regulatory deadlines. The dashboards become reproducible because the joins are persistent in the graph rather than re-derived from text each time a chart is refreshed.

9.4 CURATION AND GOVERNANCE

The graph is only as good as the curation that populates it. The recommended approach is a phased build: start with the asset hierarchy and organisational chart, both of which exist in structured form already; add entities extracted from documents in a second phase, with human review of high-impact entities such as suppliers and incidents; add cross-document relationships in a third phase. Microsoft's GraphRAG reference implementation – adopted by Neo4j and Squirrelro among others – automates much of this (Edge et al. 2024; Neo4j 2025). Graph construction costs are reported to be 10–100× higher than baseline RAG indexing at corpus level, which should be factored into the project plan.

The governance model should cover not only source documents and knowledge-graph content but also the complete AI configuration. Model versions, embedding models, tokenisation and lemmatisation settings, chunking rules, rerankers, prompts and retrieval parameters should be version-controlled and linked to evaluation results. Any material change should trigger regression testing against the approved Finnish, Swedish and English test sets before production release. The system should also retain sufficient configuration and audit information to reproduce how a given answer was generated.

The operating model should assign named ownership for source onboarding, permission mapping, content authority, model and retrieval configuration, security monitoring, evaluation-set maintenance and user support. Users should be trained to treat generated answers as navigation and decision-support aids rather than authoritative records, to verify safety-relevant claims against the cited source and to report incorrect, unsupported or access-inappropriate outputs. High-impact

extracted entities and AI-generated graph relationships should require explicit review before they enter the curated semantic layer.

10 Reference deployments and benchmarks

Several adjacent deployments, benchmark initiatives and international nuclear-sector activities demonstrate the feasibility of the architecture above and provide useful reference points for the scoping of a Nordic nuclear pilot.

The Nuclear Research Data System team at Idaho National Laboratory has demonstrated retrieval-augmented generation on nuclear research data and reports material improvements in accessibility and interpretability of dataset content (Idaho National Laboratory 2024). Atomic Canyon and Idaho National Laboratory have published the first benchmark suite for evaluating RAG and LLMs in nuclear applications, which gives a potential public reference for Nordic nuclear pilot acceptance testing (American Nuclear Society Newswire 2025). The International Atomic Energy Agency's guidance on deploying AI applications in nuclear power and its 2025 Symposium on AI and Nuclear Energy provide the international policy context (International Atomic Energy Agency 2024, 2025). The Nuclear Energy Agency's AIxpertise initiative is the natural place to publish the results of the Nordic nuclear pilot to the wider nuclear community (Nuclear Energy Agency, OECD 2025).

In adjacent regulated sectors, IntraFind's deployments at the German federal authority for radioactive waste disposal (BGE) and the Bundeswehr are the closest commercial reference points (IntraFind Software AG 2024). Squirro's deployments at the European Central Bank, the Bank of England and the Deutsche Bundesbank are the closest financial-services references and confirm that knowledge-graph-augmented RAG is suitable for environments with formal access-control regimes (Squirro 2025).

11 Risks and considerations

11.1 FINNISH LANGUAGE DEPTH

This is the single highest-priority technical risk. Publicly available benchmark evidence is insufficient to establish production-grade Finnish retrieval performance. The pilot should therefore use a dedicated Finnish in-domain evaluation set containing inflected forms, compound words, abbreviations, technical identifiers and cross-language queries. The Scandinavian Embedding Benchmark may be used as supplementary evidence for Swedish, Danish and Norwegian, but not as a substitute for Finnish evaluation. Voikko-based lemmatisation is a technically plausible way to improve Finnish lexical retrieval and should be tested on both the indexing and query paths. Its benefit, and any interaction with embedding retrieval, should be measured against the untreated baseline during the pilot.

11.2 CITATION FAITHFULNESS

Many products advertise citations as a feature; far fewer can guarantee that each citation actually supports the statement it accompanies. The mitigation is to build a test harness that scores each generated sentence against the retrieved sources, and to make citation precision a formal key performance indicator during the pilot.

A citation should be treated as correct only if the cited passage directly supports the specific sentence or claim in the generated answer. The pilot should therefore evaluate citation precision at claim level, not merely whether an answer contains links. This is especially important for summaries that combine several passages, because a plausible answer with weak citations is more dangerous than a conventional search result that simply returns too many documents.

11.3 PROMPT INJECTION AND INDIRECT ATTACKS

When the system retrieves a document, it processes the document's content as input to the language model. An attacker can embed instructions inside a document that the model later follows – for example to ignore access controls or to produce misleading summaries. CERT-EU and the OWASP Top 10 for LLM applications give the relevant guidance (CERT-EU 2024; OWASP Foundation 2025). The pilot should include adversarial testing that places malicious instructions in test documents and verify that the system does not act on them.

11.4 ACCESS-CONTROL LEAKAGE

The leakage paths in a RAG system are subtle: cached snippets that survive a document's deletion, AI summaries that paraphrase restricted facts without quoting them, conversation history shared across users. The mitigation is to favour early-binding security (Sinequa) or, in the sovereign stack, to enforce ACLs on the vector index as well as on the source documents, and to run a clearance step-down

test in which a lower-clearance user must not see higher-classification content in any retrieval path.

11.5 QUERY AND INTERACTION-DATA GOVERNANCE

Search queries, retrieved passages, generated answers, conversation histories, user feedback and evaluation traces may contain sensitive information even when the underlying documents remain correctly protected. These data should therefore be classified, access-controlled, encrypted and retained according to an explicitly approved policy. Conversation state must not be shared across users or security contexts, and a change in a user's permissions should invalidate any cached or persistent context that was created under the previous permissions. Administrative access to query and answer logs should be restricted and audited. The pilot should verify configurable retention, deletion and pseudonymisation of interaction data and should determine which records are required for security monitoring, reproducibility and regulatory accountability.

11.6 REGULATORY EXPOSURE

The regulatory classification of the proposed system depends on its intended purpose and operational use. A knowledge-management tool that retrieves documents and supports human review may be treated differently from an AI system that controls or directly influences safety-related processes. Before deployment, the operator should document the intended purpose, user responsibilities, prohibited uses and level of human oversight, and complete the applicable AI Act classification, GDPR assessment and NIS2 review. Relevant transparency requirements should be addressed before operational use, while later high-risk requirements should be monitored against the revised EU implementation timetable.

11.7 VENDOR VIABILITY AND MARKET CONSOLIDATION

The sector has consolidated rapidly: Sinequa was acquired by ChapsVision in 2024, Silo AI by AMD in 2024. The mitigation is to favour options built on open standards – OpenSearch backend, model-agnostic LLM integration, exportable indices – so that component replacement is feasible without full system migration.

11.8 AI SOFTWARE SUPPLY-CHAIN SECURITY

An on-premise deployment remains dependent on externally developed models, libraries, container images and software updates. Model weights and deployment artefacts should therefore be obtained from approved sources, verified by cryptographic checksums or signatures where available, scanned before introduction into the protected environment and recorded in a software and model inventory. The operator should define controlled procedures for offline updates, vulnerability remediation, dependency review, administrator access and vendor support. Components that require external telemetry, cloud-based licence

validation or remote administrative access should be identified before procurement and evaluated against the security boundary.

12 Conclusions and recommendations

12.1 SUMMARY OF FINDINGS

The European market now offers viable on-premise alternatives to US cloud-dependent search platforms. Four shortlisted options cover the design space: Mindbreeze InSpire for the shortest path to production, Sinequa/ChapsVision for the deepest enterprise capability and page-level citations, IntraFind for the strongest sovereignty and lowest commercial cost, and a sovereign modular stack for maximum flexibility at the cost of substantially higher in-house engineering. Finnish-language depth is the dominant technical risk across all options and must be validated empirically. The open-weight model landscape – Mistral Large 3, Llama 4, Qwen 3.5, DeepSeek V3.x/V4, Gemma 4 – is mature enough to support a fully sovereign deployment in May 2026; multimodal models such as Qwen3-VL, InternVL 2.5 and DeepSeek-OCR cover scanned-document workloads adequately for Phase 1, with visual semantic search over engineering drawings deferred to Phase 2. A knowledge-graph layer on top of the search index is the most defensible target architecture for chat, insight extraction and dashboarding.

12.2 RECOMMENDATIONS

The four candidates address different strategic priorities and are not directly interchangeable. For the specific purpose of a Phase 1 pilot, Mindbreeze is identified as the preferred starting candidate because the pilot is intended primarily to validate trusted retrieval rather than to develop an enterprise search platform from individual components.

The recommendation is based on the following considerations:

- **Deployment model:** The appliance approach reduces the amount of infrastructure assembly and component integration required before retrieval testing can begin.
- **Operational burden:** The report's indicative assessment assigns Mindbreeze a lower ongoing staffing requirement than Sinequa and the sovereign modular stack.
- **Cost:** Mindbreeze is not the lowest-cost candidate, but its indicative five-year pilot cost is close to IntraFind and materially lower than the estimates for Sinequa and the sovereign modular stack.
- **Security assurance:** The documented certifications provide relevant supporting evidence for security governance, although they do not by themselves demonstrate compliance with the operator's nuclear-security requirements.
- **Commercial maturity:** Analyst recognition and a packaged deployment model reduce product-maturity risk relative to a custom-built stack.
- **Fit with the pilot objective:** The lower expected integration burden allows the pilot to concentrate on Finnish and Swedish retrieval, citation

faithfulness, access-control enforcement, OCR, lifecycle propagation and concurrent-user behaviour.

These advantages do not establish that Mindbreeze is the best final production solution. Important uncertainties remain. Public evidence for Finnish-language depth is insufficient, source-level citations may not meet use cases that require page-level traceability, and the appliance model may limit component replaceability and increase lock-in. Air-gap operation, telemetry, update mechanisms, model options, index exportability and commercial terms also require confirmation.

Mindbreeze should therefore proceed only if the pre-pilot documentation review establishes that it can operate within the approved security boundary and if it accepts the same evaluation corpus, metrics and access-control tests as the comparison candidates. Failure to meet any mandatory security or access-control criterion should remove it from further consideration.

Preferred Phase 1 pilot candidate: Mindbreeze InSpire.

Subject to satisfactory due diligence and competitive quotation, Mindbreeze is recommended as the preferred Phase 1 pilot candidate for the reasons set out above. Its appliance-based deployment is expected to allow the pilot to focus on retrieval quality, citation faithfulness, access-control enforcement, OCR and lifecycle behaviour. The recommendation should be reconsidered if pre-pilot documentation or pilot testing identifies deficiencies in security-boundary compliance, Nordic-language retrieval, citation granularity, access control, exportability or commercial terms.

Comparative benchmark candidate– run Sinequa in parallel.

Sinequa should be included in the comparative pilot or in a structured demonstration using the same evaluation corpus. Its page-level citation capability and early-binding access-control model make it particularly relevant where precise source verification and complex security classifications are decisive. Its higher indicative cost and operational demands should be assessed against measurable improvements in citation precision, retrieval quality and security behaviour.

Alternative for maximum control – prototype the sovereign stack.

The sovereign modular stack should be considered as a separate prototype or long-term development path. It offers the greatest control, adaptability and component replaceability, but it should not be evaluated against the same implementation schedule as configured commercial products. Proceeding with this option would require committed platform-engineering capacity, explicit ownership of security integration and lifecycle consistency, and a credible long-term operating model.

Cost-optimised commercial candidate – include IntraFind.

IntraFind should be included in vendor dialogue and quotation comparison. Based on the report's indicative assumptions, it provides the lowest commercial TCO and a strong European data-sovereignty position. It could become the preferred candidate if it demonstrates Nordic-language retrieval, citation faithfulness, access-

control behaviour and operational maturity comparable to the other commercial platforms.

Final selection

No production selection should be made solely from the desk study. Final selection should follow a common pilot protocol, formal procurement criteria, verified vendor responses and an updated TCO model based on quotations. The result may differ from the initial preference if another candidate demonstrates materially better retrieval, citation, security or lifecycle performance.

12.3 PILOT VALIDATION PLAN

Before testing begins, the operator and reference group should approve the evaluation protocol, metric definitions, acceptance thresholds and adjudication process. The golden corpus, queries and relevance judgements should be version-controlled and frozen for the formal comparison. Results should be compared with a conventional keyword-search baseline and, where relevant, with the organisation's current search process. Human relevance and citation judgements should use documented criteria and more than one reviewer for a representative sample. Results should be reported with sample sizes, uncertainty ranges and failure categories rather than only aggregate percentages. Any tuning performed after viewing test results should be evaluated against a separate hold-out set. Figure 9 outlines the stage-gates and explains how those tests lead to a decision. Additionally, Table 10 outlines an example scorecard to identify gates and relative priorities.

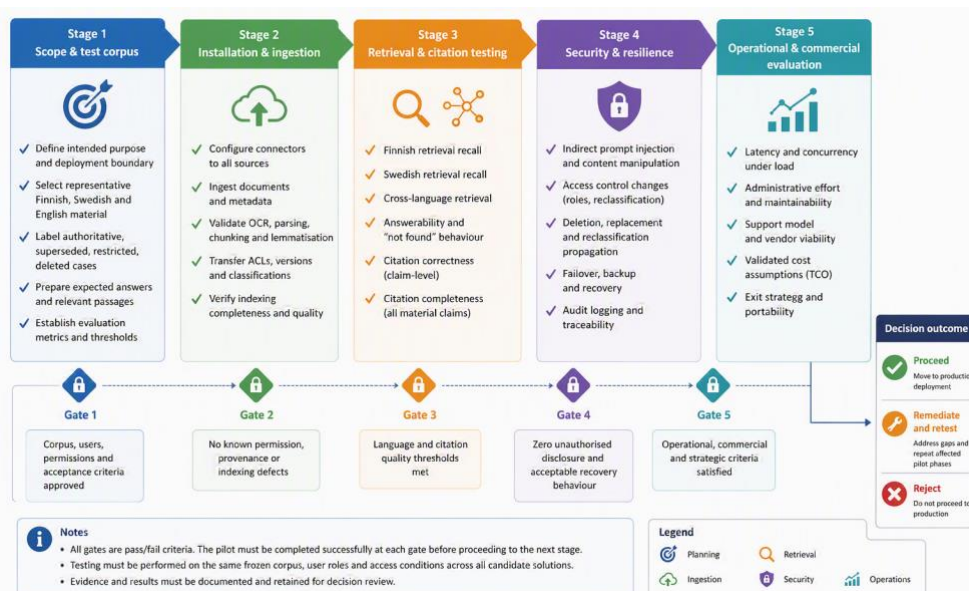


Figure 9 Proposed pilot sequence and evidence gates for progression toward production deployment.

Pilot evaluation should begin with a version-controlled golden corpus and question set. Each query should identify the relevant document revision, relevant passage, authorised test principals, expected answerability and acceptable answer language. Results should be reported separately for Finnish, Swedish and English and for exact identifiers, inflected forms, compound words, abbreviations, cross-language queries and “not found” cases. A separate adversarial corpus should cover stale permissions, deleted documents, conflicting versions, prompt-injection text, misleading citations and content duplicated under different access rights.

An 8–12-week comparative pilot may be sufficient for configured commercial platforms where the necessary connectors and deployment components already exist. A sovereign modular-stack prototype should be planned separately as a platform-development activity, with a materially longer schedule and staged exit gates for retrieval, access control, local generation and enterprise-source integration.

With the commercial platforms, weeks 1–2 cover environment setup and connector configuration. Weeks 3–5 cover retrieval and language-model configuration, embedding evaluation using the dedicated Finnish, Swedish and English in-domain test sets, supplemented by SEB evidence for Swedish where relevant, and controlled A/B testing of Voikko-based lexical preprocessing. Weeks 6–8 cover functional testing, including Finnish recall, citation accuracy, hallucination rate and concurrent-user latency. Weeks 9–10 cover security audit and adversarial testing. Weeks 11–12 cover scoring and recommendation. Acceptance thresholds, in priority order: Finnish retrieval recall@10 $\geq 85\%$; Swedish and English recall@10 $\geq 90\%$; citation accuracy $\geq 95\%$; hallucination rate $\leq 5\%$; query latency P95 ≤ 3 s single-user and ≤ 5 s at 50 concurrent users; access-control enforcement 100%; OCR character-level accuracy $\geq 98\%$ on Finnish scans.

Table 10 Example scorecard to identify gates and relative priorities

Decision area	Suggested role	Example evidence
Security boundary and data flows	Mandatory gate	Approved architecture and vendor documentation
Access-control enforcement	Mandatory gate	No leakage in positive, negative and permission-change tests
Finnish retrieval	High-weight score	Recall and relevance results on the frozen Finnish test set
Swedish and English retrieval	High-weight score	Language-specific recall and relevance results
Citation faithfulness	High-weight score	Claim-level citation assessment
Document lifecycle	Mandatory gate	Deletion, replacement, reclassification and cache-propagation tests

Decision area	Suggested role	Example evidence
Deployment and maintainability	Medium-weight score	Recorded configuration effort, skills and operational tasks
TCO	Medium-weight score	Comparable vendor quotations and organisation-specific staffing estimates
Extensibility and exit	Medium-weight score	Exportability, API coverage, model replaceability and migration provisions
Multimodal capability	Lower Phase 1 weight	OCR and text-label extraction results

Next steps:

1. Complete or document the screening decision for the EU AI Act risk classification and a GDPR DPIA before the pilot begins.
2. Decide to proceed with a commercial or sovereign demonstration.
 - a. Issue formal quotation requests to Mindbreeze and Sinequa; include IntraFind as a cost-optimised third option.
 - b. Either issue formal quotation requests for a Sovereign stack demonstration option or select/hire the team to build it.
3. Assemble a representative pilot corpus of at least 50 000 documents spanning Finnish, Swedish and English.
4. Brief the reference group on the pilot success metrics and agree the governance framework that will cover the production deployment.

13 Glossary

Term	Definition
AI	Artificial Intelligence. Computer systems designed to perform tasks that normally require human intelligence, such as understanding language, identifying patterns, retrieving information and generating content.
ABAC	Attribute-Based Access Control. Resolves access rights using attributes of the user, document and environment. Used by Sinequa for early-binding security.
ACL	Access Control List. List of users and groups permitted to access a resource.
API	Application Programming Interface. A defined interface that allows software components and systems to exchange data and invoke functions.
BPE	Byte-Pair Encoding. A sub-word tokeniser used by most large language models; splits Finnish words at sub-optimal boundaries.
CLOUD Act	U.S. Clarifying Lawful Overseas Use of Data Act (2018). Gives U.S. authorities access to data held by U.S.-headquartered companies regardless of location.
DPIA	Data Protection Impact Assessment, required under GDPR Article 35.
FTE	Full-Time Equivalent, i.e., one full-time person over the relevant period.
GPU	Graphics Processing Unit. A processor designed for highly parallel computation and commonly used to accelerate LLM inference, embeddings, reranking, OCR and multimodal processing.
GraphRAG	Retrieval-augmented generation that uses a knowledge graph as well as a vector index for retrieval; introduced by Edge et al. (2024).
LLM	Large Language Model. Neural network trained on text to generate and interpret natural language.
MAU	Monthly Active Users. A measure of the number of distinct users active during a month.
MoE	Mixture of Experts. Sparse model architecture where only a subset of parameters is active per token; standard at scale in 2026.
NLP	Natural Language Processing. The branch of AI concerned with analysing, understanding, retrieving and generating human language.
OCR	Optical Character Recognition. Conversion of scanned page images into machine-readable text.
RAG	Retrieval-Augmented Generation. Combines information retrieval with a language model to ground answers in retrieved documents.
RBAC	Role-Based Access Control. An access-control model in which permissions are assigned to organisational or system roles and users receive access through their assigned roles.
SEB	Scandinavian Embedding Benchmark. Standard benchmark for evaluating text embeddings on Scandinavian languages (Enevoldsen et al. 2024).
TCO	Total Cost of Ownership over a defined horizon (here, five years).
TLS	Transport Layer Security. A protocol used to encrypt and authenticate network communications.
Voikko	Open-source Finnish morphological analyser and lemmatiser.

14 References

American Nuclear Society Newswire. 2025. "Atomic Canyon Partners with INL on AI Benchmarks for RAG and Large Language Models in Nuclear Applications." <https://www.ans.org/news/2025-09-11/article-7362/atomic-canyon-partners-with-inl-on-ai-benchmarks/>.

Bundesamt für Sicherheit in der Informationstechnik. 2024. "Cloud Computing Compliance Criteria Catalogue (C5)." https://www.bsi.bund.de/EN/Topics/CloudComputing/Compliance_Criteria_Catalogue/Compliance_Criteria_Catalogue_node.html.

CERT-EU. 2024. "Security Guidance: Generative AI in Cybersecurity — Balancing Innovation and Risk." <https://cert.europa.eu/publications/security-guidance/generative-ai-in-cybersecurity-balancing-innovation-and-risk/>.

ChapsVision. 2025. "Sinequa Enterprise Search and RAG Platform." <https://www.chapsvision.com/en-us/platform/sinequa/>.

Chen, Jianlv, Shitao Xiao, Peitian Zhang, et al. 2024. "BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation." *arXiv Preprint arXiv:2402.03216*. <https://arxiv.org/abs/2402.03216>.

Chen, Zhe et al. 2024. "InternVL 2.5: Scaling up Vision-Language Models for General-Purpose Multimodal Understanding." <https://arxiv.org/abs/2412.05271>.

CSC – IT Center for Science. 2024. "LUMI Supercomputer: Pan-European Pre-Exascale System in Kajaani, Finland." <https://www.lumi-supercomputer.eu/>.

DeepSeek-AI. 2025. "DeepSeek-V3 Technical Report." <https://arxiv.org/abs/2412.19437>.

deepset GmbH. 2025. "Haystack Open-Source Framework and Enterprise Platform." <https://haystack.deepset.ai/>.

Edge, Darren, Ha Trinh, Newman Cheng, et al. 2024. "From Local to Global: A Graph RAG Approach to Query-Focused Summarization." <https://arxiv.org/abs/2404.16130>.

Energiforsk AB. 2025. "Digitalisation in Nuclear Power Programme." <https://energiforsk.se/en/programmes/digitalisation-in-nuclear-power/>.

Enevoldsen, Kenneth, Márton Kardos, Niklas Muennighoff, and Kristoffer Nielbo. 2024. "The Scandinavian Embedding Benchmarks: Comprehensive Assessment of Multilingual and Monolingual Text Embedding." In *Advances in Neural Information Processing Systems* 37. https://proceedings.neurips.cc/paper_files/paper/2024/hash/4746bb91bd073ec7eef930d5775122ba-Abstract-Datasets_and_Benchmarks_Track.html.

European Commission. 2025. "Implementation Timeline of the EU AI Act." <https://ai-act-service-desk.ec.europa.eu/en/ai-act/timeline/timeline-implementation-eu-ai-act>.

European Parliament and Council. 2016. "Regulation (EU) 2016/679 on the Protection of Natural Persons with Regard to the Processing of Personal Data (GDPR)." <https://eur-lex.europa.eu/eli/reg/2016/679/oj>.

European Parliament and Council. 2022. "Directive (EU) 2022/2555 on Measures for a High Common Level of Cybersecurity Across the Union (NIS2)." <https://eur-lex.europa.eu/eli/dir/2022/2555/oj>.

European Parliament and Council. 2024a. "Annex III: High-Risk AI Systems Referred to in Article 6(2)." <https://artificialintelligenceact.eu/annex/3/>.

European Parliament and Council. 2024b. "Article 99: Penalties." <https://artificialintelligenceact.eu/article/99/>.

European Parliament and Council. 2024c. "Regulation (EU) 2024/1689 of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)." <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>.

Forrester Research. 2025. "The Forrester Wave: Cognitive Search Platforms, Q4 2025." <https://www.forrester.com/report/the-forrester-wave-tm-cognitive-search-platforms-q4-2025/>.

Gao, Yunfan, Yun Xiong, Xinyu Gao, et al. 2024. "Retrieval-Augmented Generation for Large Language Models: A Survey." *arXiv Preprint arXiv:2312.10997*. <https://arxiv.org/abs/2312.10997>.

Google DeepMind. 2025. "Gemma Family of Open Models." <https://ai.google.dev/gemma>.

Hogan, Aidan, Eva Blomqvist, Michael Cochez, et al. 2021. "Knowledge Graphs." *ACM Computing Surveys* 54 (4): 1–37. <https://dl.acm.org/doi/10.1145/3447772>.

Idaho National Laboratory. 2024. "Nuclear Research Data System (NRDS): Demonstrating Retrieval-Augmented Generation for Nuclear Knowledge Management." <https://nrds.inl.gov/about>.

International Atomic Energy Agency. 2024. "Considerations for Deploying Artificial Intelligence Applications in the Nuclear Power Industry." IAEA Nuclear Energy Series. <https://www.iaea.org/publications/15866/considerations-for-deploying-artificial-intelligence-applications-in-the-nuclear-power-industry>.

International Atomic Energy Agency. 2025. "International Symposium on Artificial Intelligence and Nuclear Energy." <https://www.iaea.org/sites/default/files/ai-symposium-summary-20260224-draft.pdf>.

IntraFind Software AG. 2024. "iFinder and iAssistant: Enterprise Search and RAG with Citations." <https://intrafind.com/en/iassistant>.

- IntuitionLabs. 2025. "NVIDIA AI GPU Pricing Guide: H100, H200 and Successor Accelerators." <https://intuitionlabs.ai/articles/nvidia-ai-gpu-pricing-guide>.
- Karpukhin, Vladimir, Barlas Oğuz, Sewon Min, et al. 2020. "Dense Passage Retrieval for Open-Domain Question Answering." *Proceedings of EMNLP*. <https://aclanthology.org/2020.emnlp-main.550/>.
- Laikari, A. et al. 2018. "Wireless in Nuclear: Feasibility Study." Energiforsk. <https://energiforskmedia.blob.core.windows.net/media/24836/wireless-in-nuclear-energiforskrapport-2018-513.pdf>.
- Laikari, A. et al. 2021. "Industrial Internet of Things in Nuclear." Energiforsk. <https://energiforsk.se/media/29219/industrial-internet-of-things-in-nuclear-energiforskrapport-2021-726.pdf>.
- Lewis, Patrick, Ethan Perez, Aleksandra Piktus, et al. 2020. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." *Advances in Neural Information Processing Systems* 33. <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>.
- Martins, Pedro Henrique, Patrick Fernandes, et al. 2024. "EuroLLM: Multilingual Language Models for Europe." <https://arxiv.org/abs/2409.16235>.
- Meta Platforms. 2025. "Llama 4 Family: Open-Weight Models with Long-Context Windows." <https://ai.meta.com/llama/>.
- Mindbreeze GmbH. 2025. "Mindbreeze InSpire: Cognitive Search Appliance." <https://www.mindbreeze.com/mindbreeze-inspire>.
- Mistral AI. 2025. "Mistral Large 3: Sparse Mixture-of-Experts Open-Weight Model Under Apache 2.0." <https://mistral.ai/models>.
- Neo4j. 2025. "Integrating Microsoft GraphRAG into Neo4j." <https://neo4j.com/blog/developer/microsoft-graphrag-neo4j/>.
- Nuclear Energy Agency, OECD. 2025. "Harnessing AI: First Workshop of the New NEA Joint Project AIxpertise." https://www.oecd-neo.org/jcms/pl_111870/harnessing-ai-first-workshop-of-the-new-nea-joint-project-aiexpertise.
- NVIDIA Corporation. 2024a. "NVIDIA H100 Tensor Core GPU Specifications." <https://www.nvidia.com/en-us/data-center/h100/>.
- NVIDIA Corporation. 2024b. "NVIDIA L40S GPU Specifications." <https://www.nvidia.com/en-us/data-center/l40s/>.
- NVIDIA Corporation. 2025. "NVIDIA H200 Tensor Core GPU Specifications." <https://www.nvidia.com/en-us/data-center/h200/>.
- OpenSearch Project. 2025. "OpenSearch: Open-Source Search and Analytics Engine." <https://opensearch.org/>.

- OWASP Foundation. 2025. "OWASP Top 10 for Large Language Model Applications, 2025 Edition." <https://owasp.org/www-project-top-10-for-large-language-model-applications/>.
- Peng, Boci, Yun Zhu, Yongchao Liu, et al. 2024. "Graph Retrieval-Augmented Generation: A Survey." *arXiv Preprint arXiv:2408.08921*. <https://arxiv.org/abs/2408.08921>.
- Pitkänen, Harri. 2024. "Voikko: Open-Source Finnish Morphological Analyser and Spell-Checker." <https://voikko.puimula.org/>.
- Qdrant Solutions GmbH. 2025. "Qdrant Vector Database: Security Guide." <https://qdrant.tech/documentation/guides/security/>.
- Qwen Team, Alibaba Cloud. 2025. "Qwen 3.5: Open-Weight Apache 2.0 Language Models." <https://github.com/QwenLM/Qwen3>.
- Qwen Team, Alibaba Cloud. 2026. "Qwen3-VL: Multimodal Large Language Model Series." <https://github.com/QwenLM/Qwen3-VL>.
- Silo AI / AMD. 2024. "Viking 7B/13B/33B Open Nordic Language Models." <https://huggingface.co/LumiOpen>.
- Squirro. 2025. "Enhanced RAG with Knowledge Graphs." <https://squirro.com/customers/>.
- United States Congress. 2018. "Clarifying Lawful Overseas Use of Data Act (CLOUD Act), Public Law 115-141." <https://www.congress.gov/bill/115th-congress/house-bill/4943>.
- Wang, Liang, Nan Yang, Xiaolong Huang, et al. 2024. "Multilingual E5 Text Embeddings: A Technical Report." *arXiv Preprint arXiv:2402.05672*. <https://arxiv.org/abs/2402.05672>.
- Weaviate B.V. 2025. "Configuring Role-Based Access Control in Weaviate." <https://docs.weaviate.io/deploy/configuration/configuring-rbac>.

ON-PREMISE AI-ASSISTED SEARCH FOR THE NORDIC NUCLEAR INDUSTRY

This report evaluates practical options for deploying AI-assisted search and document intelligence in Nordic nuclear organisations without compromising data sovereignty, security or traceability. It compares European enterprise-search platforms with a sovereign modular open-source architecture and assesses their suitability for handling technical reports, scanned documents, drawings, multilingual material and regulated knowledge workflows.

The results support a phased pilot approach in which retrieval quality, citation faithfulness, Finnish and Swedish language performance, cybersecurity requirements and total cost of ownership are validated before production deployment. The report proposes a target architecture where a curated knowledge graph and retrieval-augmented generation pipelines together enable safer expert search, insight extraction and future dashboarding for nuclear-sector knowledge management.

A new step in energy research

The research company Energiforsk initiates, coordinates, and conducts energy research and analyses, as well as communicates knowledge in favor of a robust and sustainable energy system. We are a politically neutral limited company that reinvests our profit in more research. Our owners are industry organisations Swedenergy and the Swedish Gas Association, the Swedish TSO Svenska kraftnät, and the gas and energy company Nordion Energi.

