Almato Bardioc

sRAG – Beyond hallucination

Mission statement

Bardioc sRAG revolutionises access to knowledge by transforming complex data structures into meaningful contexts. Our goal is to provide accurate, precise, personalised, and transparent information using semantic intelligence for every user – anytime, anywhere.

Summary

Almato Bardioc sRAG is a powerful semantic retrieval augmented generation (sRAG) system that combines cutting-edge artificial intelligence, semantic graph technology, and advanced knowledge management in a flexible, scalable platform. It significantly outperforms traditional RAG systems by leveraging the capabilities of semantic graph databases in combination with modern artificial intelligence and natural language processing (NLP):

+ No hallucinations, i.e. accurate rather than just good-sounding responses
+ Enhanced contextualisation and precision
+ Dynamic linking of knowledge
+ Continuous optimisation through feedback mechanisms
+ Explainability and traceability of results
+ Advanced personalisation
+ Scalability and flexibility

Bardioc sRAG enables companies to make complex data more accessible and understandable, allowing them to generate accurate, precise, and contextual responses to queries. By combining semantic search with natural language processing, Bardioc sRAG facilitates superior information processing that surpasses traditional systems.

The architecture of Bardioc sRAG is characterised by its flexibility: it can be deployed as Software as a Service (SaaS), Platform as a Service (PaaS), or an on-premises solution, offering companies maximum adaptability. Support for hyperscalers such as AWS, Azure, and OpenStack makes the system universally applicable, while the highest security standards and data protection requirements, including ISO certifications and GDPR compliance, enable its operation in regulated environments.

The system is capable of integrating, linking, and continuously optimizing information from a wide range of sources, making it a central solution for knowledge management, data processing, and responding to customer and employee inquiries. Through the integrated feedback mechanism and continuous model adjustments, Bardioc sRAG reliably delivers high-quality, context-based results.

In summary, Bardioc sRAG provides companies with a future-proof solution to optimise their knowledge and data processing, drive innovation, and make strategic decisions based on solid foundations. Businesses benefit from a powerful tool that can be tailored to their individual needs while meeting the highest standards in scalability, security, and personalisation.

Contents

1.	**Introduction**	5
2.	**Overview**	6
2.1	Introduction to retrieval augmented generation (RAG)	7
2.2	Semantic retrieval augmented generation (sRAG)	7
2.3	Limitations of large language models	7
3.	**Almato Bardioc sRAG platform**	9
3.1	Functionality and features	10
3.1.1	Accuracy	10
3.1.2	Data contextualization and linking of information	10
3.1.3	Efficient search and knowledge access	10
3.1.4	Flexible and dynamically scalable knowledge	11
3.1.5	Explainability of results	11
3.1.6	Advanced personalisation	11
3.1.7	Augmented data integration	11
3.2	Technology / architecture	12
3.2.1	sRAG frontend	13
3.2.2	sRAG and LLM module	13
3.2.3	Knowledge core and graph database	13
3.2.4	Knowledge APIs and message bus	13
3.2.5	Ontology manager, data manager, access manager	13
3.2.6	Data protection and security	14
3.2.7	Integration and linking of knowledge sources	14
4.	**System requirements and operation**	15
4.1	Software as a service / platform as a service	16
4.2	Hyperscaler	16
4.3	On-premises	16
4.4	System and integration requirements	17
5.	**Application examples**	18

1. Introduction

Almato Bardioc sRAG is a sophisticated semantic retrieval augmented generation (sRAG) system that combines cutting-edge artificial intelligence with the capabilities of the Bardioc semantic data platform to provide superior access to knowledge, significantly improving information processing and utilisation. By integrating a powerful, semantically structured knowledge network, Bardioc sRAG delivers contextually deep, accurate, and precise responses to queries, surpassing traditional language models (LLMs), search, and generation systems that often suffer from hallucinations. Designed to efficiently navigate complex data landscapes, it identifies relevant content and transforms it into personalised, easily understandable, and interpretable responses. Bardioc sRAG unlocks new possibilities for businesses to access internal knowledge, optimising the specific context of data for corporate requirements in data analysis, information retrieval, and knowledge management.

The power of enterprise knowledge unfolds with Bardioc sRAG - an advanced AI system that recognizes, links and refines data. It provides precise answers, deep insights and tailored solutions to take information processing to a new level. Bardioc sRAG enables intelligent knowledge management.

2. Overview

2. Overview

2.1 Introduction to retrieval augmented generation (RAG)

In practical use within companies, the benefits of generative artificial intelligence can only unfold if sensitive internal company data, in addition to the vast repository of publicly available information, is utilised for generating new content. While large language models (LLMs) are capable of performing a wide range of general tasks, they reach their limits when faced with more complex, knowledge-intensive tasks that require specialised expertise. In particular, LLM systems cannot provide »correct« answers, as knowledge representation is not integrated as a feature in an LLM. To address this problem, Lewis et al. (2020) introduced a hybrid approach called retrieval augmented generation (RAG), which enhances LLMs by retrieving relevant information in a targeted manner. The core idea of RAG is to combine a large pre-trained language model with a dedicated information retrieval component. This allows the model to access external information sources during text generation, retrieving specific knowledge and incorporating it into the output. Technically, this is achieved by linking the original prompt with the retrieved relevant documents or text passages to create an extended context-based prompt. This extended prompt then serves as the input for the generation model.

2.2 Semantic retrieval augmented generation (sRAG)

The Almato Bardioc sRAG system is based on the Almato Bardioc semantic data platform and offers several significant advantages compared to conventional RAG systems. These advantages stem from the semantic data platform's ability to depict relationships and meanings between information in a structured way and to place these in the context of the knowledge query.

As a result, Almato Bardioc sRAG offers deeper contextualisation, better personalisation, a more flexible knowledge structure, greater explainability, and the assurance that an answer is correct compared to traditional RAG systems. These advantages contribute to the system delivering more relevant, accurate, precise, and easily understandable responses.

2.3 Limitations of large language models

When users receive incorrect information from LLMs, they quickly lose confidence in the model. In professional applications of LLMs within regulated fields such as medicine, finance, public administration, or law, erroneous information due to misleading conclusions and the associated consequences is unacceptable.

The problem of so-called hallucinations in LLMs refers to the fact that these models sometimes generate incorrect or fabricated information, even though it sounds certain and plausible. The root cause of hallucinations lies in the following conceptual limitations of LLMs:

No representation of knowledge: The transformer algorithms underlying an LLM depict a system that optimises the sequencing of words and sentence fragments to produce language that is highly comprehensible to humans. Specific context is statistically represented through very large parameter vectors. However, the system as a whole does not store knowledge in the sense of causal relationships. As a result, LLMs lack the inherent ability to verify facts directly.

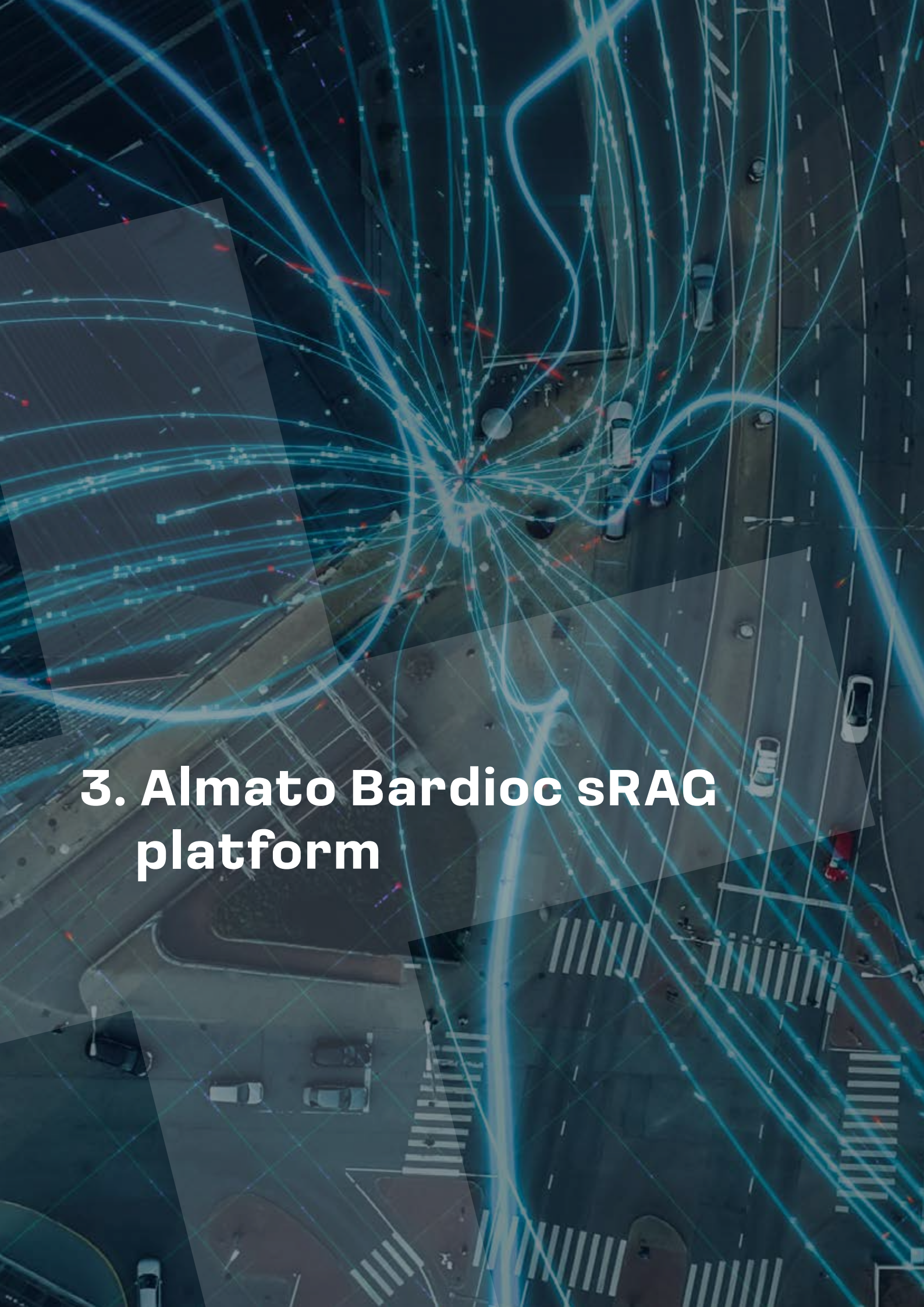
Purely statistical evaluation: An LLM does not contain information about the correctness of the data or an understanding of the world. Data is sequenced purely based on statistical considerations. Texts are generated by combining predicted words and sentences based on patterns from the training data. This explains why a proof of concept with LLMs and limited data often yields good results, as only the desired outcomes are supplemented, making them the only relevant answers in the system. In practice, however, with extensive and heterogeneous data, a different picture emerges, as numerous correct answers may exist, but they lack statistical relevance.

New information is statistically irrelevant at first: When a system is supplied with new information or entirely new findings are introduced, they are statistically irrelevant due to their small quantity and poor integration with existing information, even if they replace outdated or incorrect data. This necessitates retraining LLMs in large steps rather than being able to continuously add new information.

Reliability and validation: An LLM generates answers based on patterns in the training data, which means it can provide incorrect or inaccurate information (hallucinations) if the training data is incomplete or biased. For particularly complex or rare questions, an LLM encounters knowledge gaps and attempts to fill them with plausible but fabricated answers. However, in applications related to knowledge management, it is essential that knowledge is validated and consistently reliable. An LLM neither has an integrated method nor the necessary data to verify or validate the information.

Explainability and transparency: An LLM is a »black box« *by design*, meaning the reasons behind a specific answer or decision cannot be traced retrospectively. LLMs are designed with generative behaviour to creatively and fluently produce new texts. However, this can lead to incorrect conclusions due to overgeneralisation. For effective knowledge management, it is crucial that both the sources and the logic behind the information are transparent and traceable. This is particularly critical in regulated fields such as healthcare, the public sector, or finance.

Knowledge graphs and semantic data structures help avoiding hallucinations in LLMs by providing structured and verifiable knowledge. Through semantic retrieval augmented generation (sRAG), the accuracy and reliability of generated content can be significantly improved.



3. Almato Bardioc sRAG platform

3. Almato Bardioc sRAG platform

3.1 Functionality and features

3.1.1 Accuracy

Through the semantic data structure of the Bardioc graph, data is no longer linked statistically but follows causality chains used by humans. This enables the Bardioc sRAG model, unlike purely LLM-based implementations of RAG systems, to provide a correct answer rather than just a statistically relevant one. This avoids incorrect or fabricated answers, commonly referred to as hallucinations.

3.1.2 Data contextualization and linking of information

Almato Bardioc not only stores isolated data but also their causal relationships. This means the Almato Bardioc sRAG system can leverage deeper connections between content and concepts. It can semantically link information to better understand questions and generate more precise and accurate answers by forming a concrete query within the information base by means of causal relationships. From the reliably correct answers retrieved, it then generates a context-based response that is easily understandable for humans.

Example: A conventional RAG system might struggle to establish connections between interest rate decisions and logistical processes when asked, »How does monetary policy affect supply chains?« due to the many statistical correlations. However, the semantic data platform Bardioc can build a better understanding through causal graph links between economic variables and supply chain processes, ignoring correlations that lack causal justification and providing more relevant information.

The extended context is generated using the ontology as part of the semantic data structure. The ontology represents a detailed hierarchy and classification of concepts and their relationships.

3.1.3 Efficient search and knowledge access

The semantic structure of Bardioc makes it easier to search for relevant information based on its meaning rather than just the statistical connection of keywords. Traditional systems based on text similarity often return less relevant information, whereas the semantic Bardioc system specifically searches for concepts and their causal relationships, delivering more accurate and precise results.

With this, companies can leverage the capabilities of an LLM to provide human-readable and understandable answers while simultaneously addressing what an LLM cannot do, i.e. ensuring the correctness of these answers. Additionally, users can better understand complex relationships and make more informed decisions by analysing their data in a broader context and making it available for answer generation.

3.1.4 Flexible and dynamically scalable knowledge

Almato Bardioc allows for modelling knowledge in the same way humans do. Bardioc can dynamically expand knowledge without requiring adjustments to existing structures. New information and its relationships can be easily added, allowing the system to continuously learn and amplify its knowledge. Moreover, it can identify causally correct information as accurate answers, even without relying on large data volumes.

Conventional RAG systems often operate based on fixed data models and purely technical representations of relationships (e.g., vectors), making it more challenging to expand them and integrate new knowledge domains. In the semantic Bardioc platform, new nodes and edges can be added to the graph, enabling the knowledge base and its underlying relationships to grow continuously and remain up to date.

3.1.5 Explainability of results

The semantic Bardioc sRAG system offers explainability by being able to extract the paths between connected data points in the Bardioc knowledge graph, thereby answering the question of »Why«. This makes it easier to understand why certain information was considered relevant and how the system arrived at its conclusion.

Example: In a financial reporting, the system could explain how a specific metric is linked to other financial or operational indicators by highlighting the underlying connections in the semantic graph.

3.1.6 Advanced personalisation

The semantic data platform Bardioc can deliver more precise personalised results by modelling user history, preferences, and behaviours within the data and linking them to relevant content. The Bardioc sRAG system can semantically capture these connections between user interests and the knowledge base, enabling the generation of personalised responses. This approach understands the user's implied expertise or available context rather than merely accounting for it statistically.

For instance, a user who regularly makes inquiries about a specific topic can receive more suitable responses aligned with their preferences and interests through data connections, while skipping background information they have already seen multiple times.

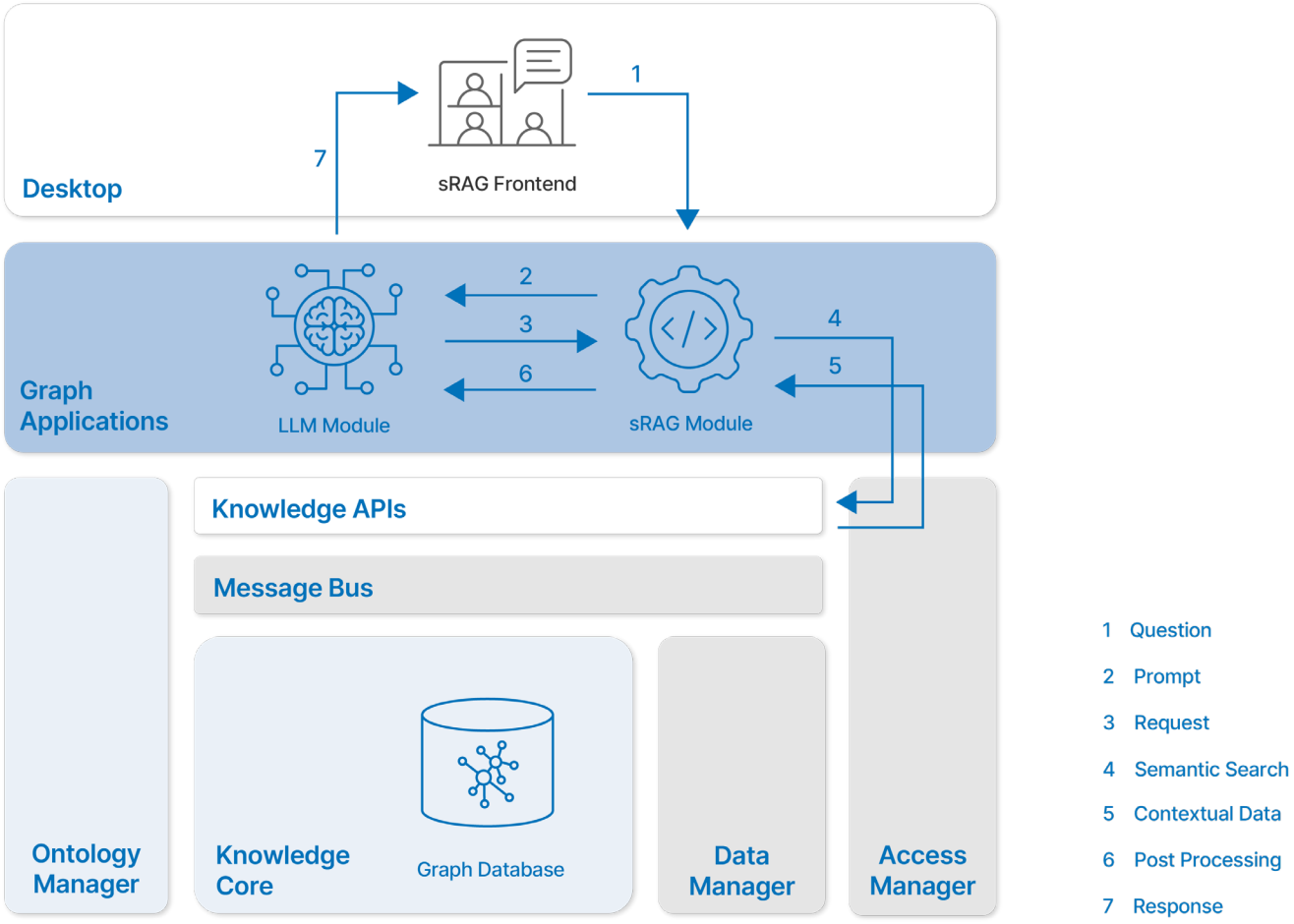
3.1.7 Augmented data integration

Since the semantic data platform Bardioc can easily link various data sources (e.g., structured and unstructured data), the Bardioc sRAG system is able to integrate and utilise heterogeneous data sources more efficiently. This enhances the quality and scope of the generated responses, enabling the Almato Bardioc sRAG system to provide comprehensive answers based on the latest available data and documents.

Conventional systems without such infrastructure often rely on a limited number of data sources and formats, and the integration of structured and unstructured data must be laboriously created for each individual case. This is not the case with the Bardioc sRAG system.

3.2 Technology / architecture

The Bardioc sRAG system operates as an application within the Almato Bardioc Platform, leveraging the platform's full range of security, scalability, and integration capabilities at all times. The architecture of Bardioc sRAG is based on a modular and scalable design that integrates the power of Artificial Intelligence (AI), semantic graph technology, Natural Language Processing (NLP), and Large Language Models (LLM). This architecture is specifically designed to process highly complex data contexts and provide users with accurate, precise, and context-based answers in a language that is comprehensible for humans. The following diagram illustrates the data flow and interactions between the various components of the Bardioc sRAG architecture.



Architecture for semantic retrieval augmented generation based on Bardioc

3.2.1 sRAG frontend

The frontend of the Bardioc sRAG system serves as the interface between users and the system. Here, the user inputs a question (Step 1) in natural language—possibly including domain-specific technical language—which is then forwarded to the sRAG module. After processing and generating a response through the backend components, the answer (Step 7) is presented to the user in the frontend.

3.2.2 sRAG and LLM module

The sRAG module is the central component that processes the user query and integrates the LLM module to extract and prepare relevant information from the available knowledge data. The process begins with the formulation of a prompt (Step 2) based on the user query. Using a specialized, optimized LLM, the user query is translated from natural language into a graph database query (Step 3) and further optimized.

3.2.3 Knowledge core and graph database

The Almato Bardioc graph database forms the core of the architecture, storing all data and their semantic relationships. This semantic graph ensures that the system searches for answers not just based on keywords but through connections based on meaning.

When a user query is submitted, the sRAG module performs a **semantic search (Step 4)** to identify relevant data points within the graph. These are then returned to the sRAG module as **context-related data (Step 5)**. This semantic search ensures the accuracy of the response and significantly enhances the precision of the answers by recognizing deeper connections between data.

3.2.4 Knowledge APIs and message bus

The Bardioc Platform Message Bus serves as the intermediary between the modules and the knowledge database, efficiently coordinating data communication and information exchange. The Knowledge APIs provide standardized interfaces that allow other systems and applications to access the knowledge base. These APIs are flexible and facilitate seamless integration with third-party systems or external data sources.

3.2.5 Ontology manager, data manager, access manager

The Almato Bardioc sRAG system is seamlessly integrated with additional supporting modules:

+ The **ontology manager** handles the semantic models and ontologies that enable to understand the relationships and meanings of the data. This allows for deeper modelling of knowledge and context.
+ The **data manager** is responsible for managing and maintaining the data, including data source integration and data preparation.
+ The **access manager** ensures that only authorised users can access specific data. It plays a key role in security and permission management.

For more information about the supporting modules, please refer to the separate product description for Almato Bardioc.

3.2.6 Data protection and security

Security and data protection are core components of the Bardioc sRAG architecture. All data access, whether to stored or learned information, is secured through strict and granular **access controls**. Data **in transit and at rest** is always encrypted, and all access mechanisms and protocols are digitally signed to prevent tampering.

This ensures that the use of Bardioc sRAG provides a foundation for **compliance with data protection regulations** such as the GDPR. If customer-specific security protocols and encryption methods are required, they can be implemented on a project-specific basis thanks to the modular architecture.

3.2.7 Integration and linking of knowledge sources

A central element of the Bardioc sRAG architecture is the seamless integration of knowledge sources. The system can integrate and link information from various sources using a robust connector framework. This framework enables interfaces to be easily created, monitored, and maintained. The implementation of such interfaces is carried out through standardised APIs, allowing for quick and straightforward connectivity.

The semantic data platform Almato Bardioc enables the integration, semantic linking, indexing, and continuous updating of diverse data sources without the need to retrain the LLMs.

Overall, the architecture of Bardioc sRAG offers exceptional flexibility, security, and adaptability. By combining semantic technologies with powerful AI modules, the system can deliver accurate and precise answers based on a deep understanding of the underlying data. Its high level of integration and adaptability allows seamless embedding into diverse data environments while maintaining the highest security standards.

4. System requirements and operation

4. System requirements and operation

The Bardioc sRAG system is built on the powerful Almato Bardioc data platform. This platform is specifically designed for extreme scalability, high security standards, and speed. To meet these demands, the infrastructure and operational requirements are correspondingly high. The platform offers flexible deployment options and can be used in various operating models.

The features of the Almato Bardioc sRAG system can be utilised as Software as a Service (SaaS) on the multi-tenant central platform. Alternatively, a dedicated platform can be deployed in an enterprise cloud or private cloud as a Platform as a Service (PaaS) model.

4.1 Software as a Service / Platform as a Service

Almato operates the Bardioc Platform in geo-redundant data centres of DATAGROUP located in Frankfurt am Main. The data centre operations are certified according to the ISO 27001 security standards and the ISO 20000 quality standards. Continuous monitoring of all IT components, services, and capacities, along with a high degree of automation, ensures high service levels and cost-efficient operations.

4.2 Hyperscaler

The *Almato Bardioc Platform* is compatible with all major cloud providers, particularly the platforms of leading hyperscalers such as Amazon Web Services, Microsoft Azure, and all cloud platforms based on OpenStack. Reference architectures are available for the deployment of the platform.

4.3 On-premises

For customers with high protection requirements or an explicit make strategy, the technology of the Almato Bardioc Platform and the Bardioc sRAG system is also available for on-premises operation in their own data centre. An on-premises strategy should only be considered by customers with very large rollout requirements due to the necessary infrastructure and system resources.

Production ready:

+ Platform as a Service by Almato in German data centres
+ All hyperscalers (AWS, Azure, OpenStack)
+ On-premises operation in your own data centre

4.4 System and integration requirements

The system and integration requirements for on-premises or cloud installations are determined in customer-specific requirements workshops. These requirements are tailored individually to meet the specific needs of the customer, regardless of whether a SaaS, PaaS, or on-premises solution is implemented.

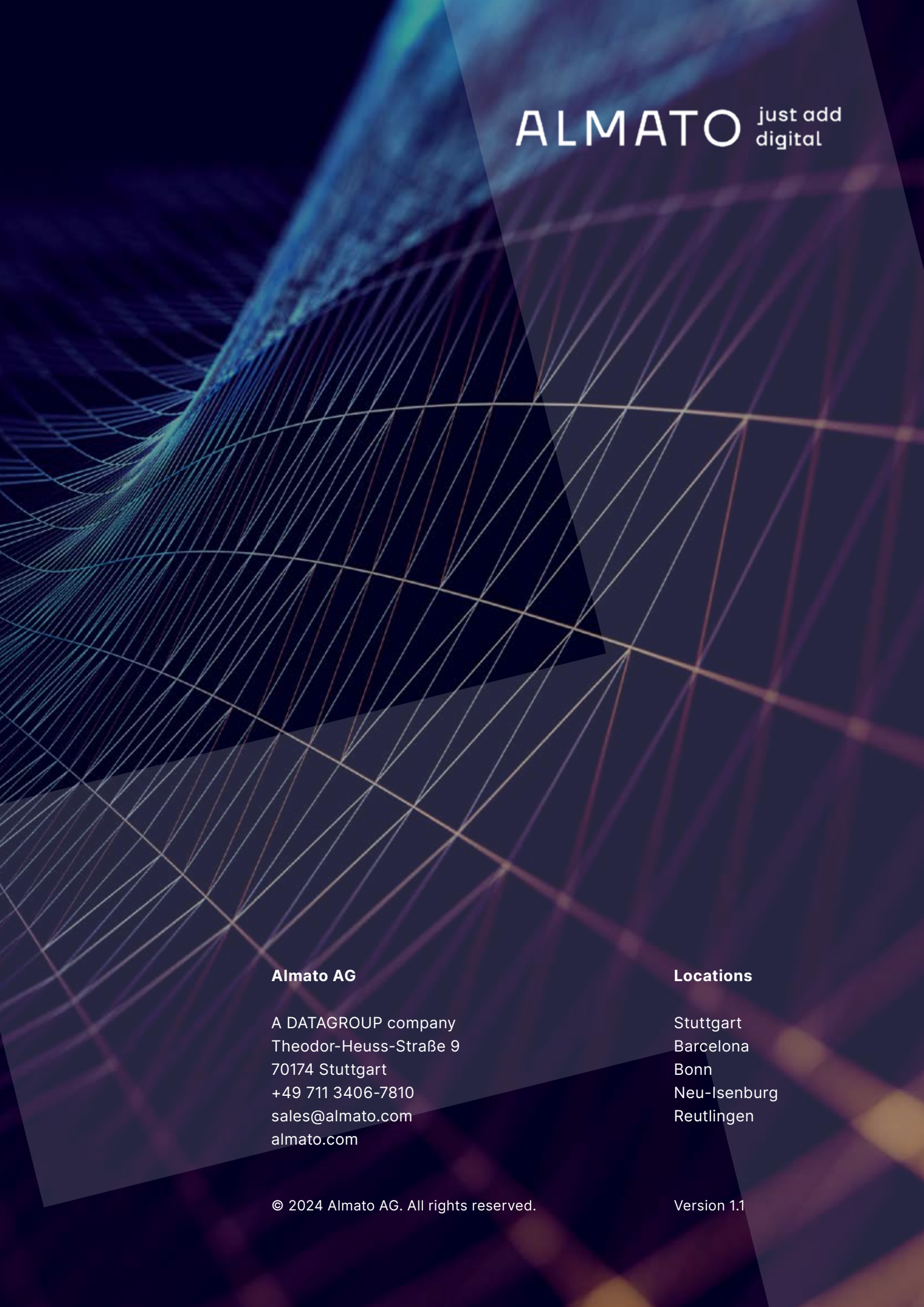
This flexibility allows companies to choose the optimal operating environment for their needs and fully leverage the capabilities of the Bardioc sRAG platform.

5. Application examples

The combination of LLMs with knowledge graphs and semantic data structures into an sRAG opens up numerous use cases for businesses and organizations where the reliability and accuracy of generated information are of great importance:

- **Technical support and troubleshooting:** Complex technical problems require accurate and specific solutions, as incorrect instructions could lead to costly errors. Semantically structured technical documentation and troubleshooting protocols can support LLMs in generating precise instructions for solving technical issues.
- **Knowledge management in healthcare:** Information on rates or treatment procedures can be provided reliably and up-to-date based on semantically structured data, for example, from knowledge databases or treatment protocols.
- **Legal assistance system:** Laws, precedents, commentaries, and expert opinions can provide legal information in the context of the actual inquiry through semantic processing. This can also take into account temporal changes in the legal situation (historicisation).
- **Scientific analyses:** Scientific analyses: Research data, publications, and studies can be evaluated in a context-specific manner to identify suitable literature or to focus on the most relevant and verifiable information.
- **Education and eLearning:** Education systems, in particular, must rely on accurate learning materials based on verified facts. With the aid of sRAG applications, personalised learning content can be generated.

sRAG applications based on Almato Bardioc offer a promising solution for avoiding hallucinations by utilising structured, verified, and verifiable data sources to answer queries. This increases the reliability and accuracy of the generated information, particularly in critical application areas where errors could have severe consequences.



Almato AG

A DATAGROUP company
Theodor-Heuss-Straße 9
70174 Stuttgart
+49 711 3406-7810
sales@almato.com
almato.com

Locations

Stuttgart
Barcelona
Bonn
Neu-Isenburg
Reutlingen

Version 1.1