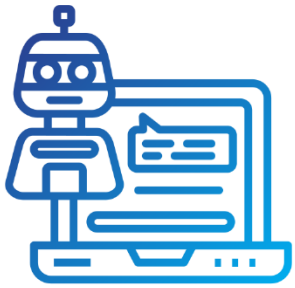
Ref. Ares(2026)1257203 - 04/02/2026



EuroLegalBot

CRIMINAL MATTERS CHATBOT

Enhancing Judicial Cooperation through a European Arrest Warrant (EAW) Chatbot

Chatbot Development and Training

Co-funded by the European Union

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Project Number: **101192405**
JUST-2024-JCOO

D3.3 – Chatbot Development and Training

Table of Contents

1. Change Control

1.1 Document Properties

Deliverable No.		D3.3	
Work Package No.	WP3	**Work Package Title**	Chatbot Development
Author/s		Rafael Tavares (AGF)	
Contributor/s		Camillo Rospigliosi, Claudio Di Lauro (Expleo)	
Reviewer		Irene Mariotti (AGF); Sergio Bianchi (AGF).	
Name		Chatbot Development and Training	
Date		February 2026	
Dissemination Level		PUBLIC	

1.2 Revision History

Version	Date	Comments
1.0	29/11/2025	First draft
1.1	09/01/2026	Expleo Contribution
1.2	03/02/2026	Final version

1.3 Table of Figures

1.4 List of Tables

2. Executive Summary

This deliverable presents the design, development and training of EuroLegalBot, an institutional-grade conversational Artificial Intelligence system developed to support judicial cooperation in criminal matters, with a specific focus on the European Arrest Warrant (EAW).

EuroLegalBot has been conceived in response to the growing need for trustworthy, transparent and legally compliant AI solutions in sensitive legal environments. In such contexts, generic AI systems often prove inadequate due to risks related to hallucinations, lack of traceability, limited control over data provenance and insufficient alignment with European regulatory and ethical requirements. To address these challenges, EuroLegalBot adopts a controlled and auditable architecture based on Retrieval-Augmented Generation (RAG), combined with deterministic multi-agent orchestration and a curated legal knowledge base.

From a technical perspective, the system is built on a modular and extensible architecture that separates document ingestion, processing, storage, retrieval and response generation. This separation of concerns ensures scalability, maintainability and long-term evolvability, while supporting both online and offline deployment scenarios. Particular emphasis has been placed on data sovereignty, auditability and compliance with European standards, including the General Data Protection Regulation (GDPR) and the emerging European regulatory framework on Artificial Intelligence.

A core contribution of the project is the development of a structured and continuously updated knowledge bank, populated through an automated web scraping and ingestion pipeline targeting authoritative European and national legal sources. The resulting dataset comprises more than 6,300 legislative and judicial documents organised by institutional origin and jurisdiction, enabling verifiable, traceable and context-aware responses grounded exclusively in validated legal material.

The chatbot training methodology combines structured data preprocessing with expert-driven question design. Legal experts, including lawyers and magistrates, actively contributed to defining realistic use cases and validation scenarios, supporting iterative testing and fine-tuning of the system. This approach ensures accurate, multilingual and legally sound responses while preserving the role of human judgement.

Overall, this deliverable demonstrates the feasibility of designing and implementing an AI-based legal assistant that meets the stringent technical, legal and ethical requirements of the European institutional context, while providing tangible operational value for judicial cooperation and legal practice.

3. Introduction

The EuroLegalBot project is developed within the framework of the European Union's digital transformation strategy and its commitment to trustworthy, ethical and transparent Artificial Intelligence. The legal and judicial domain represents a particularly sensitive area, characterised by complex normative frameworks, multilingual sources, evolving jurisprudence and strict requirements in terms of accuracy, explainability, traceability and data protection.

In this context, generic AI solutions are often inadequate. Legal practitioners, judicial authorities and institutional actors require specialised systems that are capable of operating on authoritative sources, providing verifiable answers and supporting—rather than replacing—human legal reasoning.

3.1 Need for AI systems in sensitive legal environments

In legal and institutional environments, errors, hallucinations or lack of traceability may have significant legal and operational consequences. For this reason, AI systems must be designed with explicit control mechanisms, deterministic workflows and clear links between outputs and source materials.

EuroLegalBot responds to these needs by adopting a controlled Retrieval-Augmented Generation (RAG) architecture, combined with curated datasets, auditability and domain-specific logic, ensuring that all generated responses are grounded in pre-ingested and validated legal content.

4. Overview of the EuroLegalBot system

4.1 Functional definition

From a functional perspective, EuroLegalBot is an AI-based legal response system that manages the entire legal knowledge lifecycle: document ingestion, semantic processing and enrichment, multimodal indexing, retrieval of relevant contexts, and generation of structured responses. User interaction takes place via a conversational interface, similar to a chat, which allows questions to be asked in natural language, either via text or voice.

The system uses an agency-based Retrieval-Augmented Generation (RAG) model, in which queries are deterministically routed to agents specialised in specific legal domains (e.g., civil, criminal, administrative or tax law). Each agent applies hybrid search rules and domain-specific criteria, ensuring greater accuracy and consistency in responses.

4.2 Problems that the system aims to solve

EuroLegalBot is designed to address recurring problems in the legal context:

- A lot of time spent searching for and reading long and complex documents.
- Difficulty in locating relevant precedents, articles or extracts in large document collections.
- Risk of inconsistency and subjectivity in the interpretation of legal documents.
- Limitations of generic AI solutions for GDPR compliance and data sovereignty.
- Lack of auditability and traceability in responses generated by traditional language models.
- Necessity to use multilanguage tools for EU judicial cooperation mechanisms.

By structuring legal knowledge and applying hybrid search engines and specialised agents, EuroLegalBot significantly reduces manual effort, increases the reliability of responses and mitigates regulatory risks.

4.3 Areas of application for EuroLegalBot

EuroLegalBot can be applied in various contexts and organisations, such as:

- Law firms and corporate legal departments.
- Public agencies and regulatory institutions.
- Compliance, audit and internal control offices.
- Contract analysis, comparative law and case law.
- Support for the preparation of reports, opinions and structured legal documents.

Its modular architecture allows it to adapt to different legal fields and jurisdictions, making it suitable for both internal use and large-scale institutional solutions.

4.4 Positioning of EuroLegalBot relative to existing solutions

Compared to generic chatbots and traditional legal research tools, EuroLegalBot positions itself as an institutional-grade solution. While existing solutions focus on textual search or probabilistic responses from language models, EuroLegalBot combines:

- Hybrid search (vector, structural, metadata and full text).
- Specialised legal agents with deterministic routing.
- Architecture oriented towards regulatory compliance and EU data sovereignty.
- Complete observability and audit trail.

This positioning differentiates it as a trusted platform, suited to environments where explainability and control are as important as performance.

5. System development and architecture

5.1 Overall architectural approach

EuroLegalBot is built on a modular and extensible architecture that separates ingestion, processing, storage, retrieval and response generation. This separation of concerns ensures scalability, maintainability and long-term evolvability of the system.

The system follows a Retrieval-Augmented Generation (RAG) paradigm, in which user queries are answered by combining large language models with a structured and curated legal knowledge base. This approach mitigates the limitations of purely generative models and ensures grounding in authoritative sources.

5.1.1 General objectives

The overall objective of the EuroLegalBot project is to develop and validate a specialised conversational Artificial Intelligence system capable of supporting the analysis, interpretation and exploitation of legal and training materials, while maintaining full alignment with the regulatory and institutional principles of the European Union.

EuroLegalBot aims to:

- facilitate structured access to large volumes of legal documentation;
- reduce the time needed to search, analyse and understand complex content;

- support training and capacity-building processes in the field of law;
- ensure that the answers provided by the system are substantiated, traceable and verifiable.

The project therefore aims to demonstrate that it is possible to apply AI responsibly and effectively in sensitive environments without compromising the quality of legal analysis or data sovereignty.

5.1.2 Technical objectives

From a technical point of view, EuroLegalBot aims to implement a robust, modular and auditable architecture based on software engineering principles suitable for critical systems. The main technical objectives include:

- implementing a RAG (Retrieval-Augmented Generation) architecture oriented towards the legal domain;
- developing a deterministic multi-agent system, with agents specialised by legal area;
- integrating hybrid search engines (vector, text and metadata-based);
- ensuring a clear separation of responsibilities between data ingestion, orchestration, response generation and observability;
- enabling the system to operate both online and offline, while preserving essential functionality;
- ensuring the technical traceability of interactions through structured logs and audit trails.

These technical objectives are fundamental to ensuring predictability, long-term maintenance and the suitability of the system for institutional contexts.

5.1.3 Functional objectives

From a functional point of view, EuroLegalBot aims to provide concrete features that can be used directly by its target audiences. Key functional objectives include:

- enabling conversational interaction with already assimilated legal materials;

- allowing the extraction of relevant information, structured summaries and comparisons between documents;
- supporting the identification of regulatory references, key concepts and relationships between legal texts;
- offering answers based exclusively on the content available in the system;
- adapting the behaviour of the system according to the user profile and context of use (formal, analytical, exploratory);
- providing structured outputs that can be reused in study, training or professional analysis contexts·

The functional objective of EuroLegalBot is not to automate legal decisions, but to provide qualified support for human understanding and work.

5.1.4 Strategic objectives

At a strategic level, EuroLegalBot aims to contribute to the broader objectives of the European Union in the field of digital transformation and trustworthy Artificial Intelligence. The main strategic objectives include:

- strengthening European digital sovereignty by reducing external technological dependencies;
- promoting the use of AI in line with European values, such as transparency, accountability and data protection;
- demonstrating the feasibility of specialised AI solutions for high-risk sectors;
- creating a reusable and scalable technological base for future European projects;
- supporting public institutions and training bodies in modernising their processes without compromising legal certainty.

EuroLegalBot thus positions itself as an example of the strategic application of AI in the public and institutional sector.

5.2 Technological choices

EuroLegalBot's technological choices were guided by requirements of security, compliance, performance and evolutionary flexibility. The goal was not only to select modern technologies, but to adopt components that would support the growth of the system without compromising governance and control.

5.2.1 Criteria for technology selection

The main criteria adopted were:

- Regulatory compliance (GDPR and data sovereignty requirements).
- Possibility of self-hosting and full control of the data cycle.
- Horizontal scalability and modularity.
- Technological maturity and community support.
- Observability, auditability and traceability capabilities.

5.2.2 Reasons for excluding alternative solutions

Fully managed and proprietary solutions (e.g., vector banks or SaaS-only LLMs) were rejected due to:

- Excessive dependence on third parties for the processing of sensitive data.
- Risks associated with extraterritorial legislation and data retention.
- Limitations on customisation and review of the decision-making flow.

Similarly, agent frameworks with non-deterministic behaviour were avoided, as they introduce unpredictability that is incompatible with the legal domain.

5.2.3 Advantages of selected technologies

Adopting components such as Qdrant (self-hosted vector bank), PostgreSQL, LangChain with LangGraph, and container-based infrastructures allows for:

- Full control over data and processing.

- Improved performance in combined semantic and structural searches.
- Explicit and auditable routing of AI decisions.
- Ease of evolution and replacement of components over time.

5.2.4 Impact of technological choices on system performance

Technological choices directly influence EuroLegalBot's performance by enabling:

- Low-latency responses through caching and efficient indexing.
- Independent scalability of critical services.
- Optimised costs by separating event-driven, continuous, and computationally intensive workloads.

These factors ensure that the system maintains adequate response times even as the volume of data and users grows.

5.3 Choice of European Artificial Intelligence

The choice of European language models, such as the French Mistral and the Swiss Apertus, is in line with the European Union's strategic values and objectives in terms of technology. Firstly, there is strong political and institutional momentum to reduce dependence on external suppliers and strengthen European digital sovereignty. The European Parliament emphasises the need to promote an autonomous and secure digital economy by reducing dependence on non-European actors and investing in its own digital infrastructure. In this context, adopting European AI is a strategic decision that responds to the demand for sovereign and locally controlled technologies.

Furthermore, European models have been designed from the outset to meet the legal and ethical requirements of the European context. The AI Initiative of Switzerland, for example, developed Apertus with the explicit aim of creating a reliable, sovereign and inclusive model that serves as an open core technology on which others can build ethz.ch. This reflects a motivation for open innovation: by releasing advanced models under permissive licences (such as Apache 2.0), research, local industry participation and the dissemination of knowledge are encouraged without the constraints imposed by proprietary systems. In short, the choice of European AIs is motivated by a combination of alignment with European values (privacy,

transparency, linguistic plurality) and strategic considerations of technological independence and regional innovation.

5.3.1 Comparison with North American and Chinese AIs

The adoption of European models is also justified when we compare their technical, ethical and legal characteristics with those of North American and Chinese AIs:

- Technical aspects: recent European models demonstrate remarkable technical competitiveness. Mistral 7B (France), for example, outperformed larger US open-source models: it surpassed LLaMA 2 with 13 billion parameters in several benchmarks, despite being about half the sizeventurebeat.com. This efficiency suggests that European models can achieve high performance with optimised architectures, incorporating cutting-edge innovations in research (such as Grouped-Query Attention and Sliding Window for long sequences). In contrast, leading US AIs, such as OpenAI's GPT-4, tend to be large-scale models with superior performance in many tasks, but are closed and accessible only through proprietary APIs. This limits auditability and the possibility of direct customisation by the user.

  In China, companies such as Baidu and Alibaba have also developed large-scale models (e.g., ERNIE Bot, Wen), with a focus on the Chinese language and integration into their ecosystems. In China, companies such as Baidu and Alibaba have also developed large-scale models (e.g., ERNIE Bot, Qwen), with a focus on the Chinese language and integration into their ecosystems; however, these models are generally not available with open weights, making independent testing and local adaptations difficult.

- Ethical and content aspects: European models are designed from the outset with ethical and transparency principles. Apertus, for example, was based on use ‘for the public good’, incorporating multilingualism, transparency and regulatory compliance as fundamental design principles. In addition, its developers applied strict data filters to avoid bias and inappropriate content, removing personal information and complying with deletion requests from original sources even before training. In comparison, North American models such as OpenAI's adopt content moderation policies and internal ethical guidelines; however, they are criticised for a lack of clarity on training data and for possible hidden biases arising from non-transparent datasets. In China, AIs follow ethical guidelines aligned with state regulation, which in practice translates into strong censorship of politically sensitive content. For example, Baidu's ERNIE chatbot will refuse to respond or change the subject when asked about topics considered sensitive

by the government (such as political leaders or historical events), in strict compliance with government censorship requirements.

This difference is significant: while the EU gives priority to cultural diversity and freedom of information compatible with fundamental rights, Chinese models incorporate restrictions that conflict with European values of freedom of expression.

- Legal and regulatory aspects: European AIs offer clear advantages in terms of compliance with local legal regulations. They are developed with European laws on data protection and future regulations on AI in mind, while many foreign solutions face compliance challenges. A well-known example was the sanction imposed by the Italian regulator (Garante) on OpenAI: a €15 million fine for GDPR violations resulting from the improper processing of personal data to train ChatGPT. This shows that even highly sophisticated American models can clash with strict European privacy and transparency laws. In contrast, initiatives such as Apertus have incorporated legal considerations from the outset – it is the first major model designed to comply with the transparency requirements of the future European AI Regulation (AI Act). In practice, this means comprehensive documentation of the dataset and training process, compliance with European copyright laws, and mechanisms for deleting personal data ethz.ch. As for Chinese models, in addition to being subject to state surveillance laws, they have not been designed to comply with Western standards such as the GDPR; their direct use could entail legal risks if European citizens' data were processed outside the standards required by the EU. Therefore, in terms of legal balance, European models are the safest choice and in line with full compliance with existing and emerging legal and regulatory obligations in the European space.

5.3.2 Benefits in terms of digital sovereignty

The use of European AI in the EuroLegalBot project brings unequivocal benefits of digital sovereignty to the European Union and its Member States. In practical terms, this means that the AI infrastructure can be self-managed in IT environments under European jurisdiction, ensuring full control over the data and algorithms involved. As envisaged in the technical planning of the project, the migration to self-hosted AI models in European clouds is part of the strategy to achieve greater control and operational independence. Unlike external AI services, where sensitive data may transit through servers in other jurisdictions, a locally integrated European model ensures that data remains in the EU, subject only to European laws and audits.

From a strategic point of view, this choice strengthens European technological autonomy. The European Union, through specific initiatives and funds, encourages the development of its own AI capabilities so as not to be at the mercy of foreign monopolies. The President of the European Commission has already emphasised that Europe intends to 'set its own standards' and has decided to step up investment in technological sovereignty, including the creation of scale funds and European laws on AI development and cloud computing. Choosing European AI achieves these objectives, signalling that EuroLegalBot is based on locally sourced technologies, promoting regional industry and keeping skills and intellectual property in Europe. In short, digital sovereignty is strengthened by the absence of dependence on external suppliers, by direct control over the evolution of AI (the ability to customise the model according to European needs) and by greater trust from users and authorities when they know that the solution fully complies with the European regulatory and security ecosystem.

5.3.3 Compliance with European principles

European AI models are distinguished by their intrinsic compliance with EU principles and regulations, including privacy, ethics and transparency. First and foremost, there is strong adherence to the GDPR in the design of these AIs. Apertus, for example, was developed in compliance with Swiss data protection laws and the strict European transparency standards set out in the AI Act, applying the principle of data minimisation. Their training set was limited to public and licensed data, filtering out personally identifiable information and even respecting websites' opt-out signals to exclude content from those who explicitly refused. This proactive stance demonstrates alignment with the GDPR, which requires a legal basis for the processing and protection of personal data, and anticipates future obligations to allow the exclusion of individuals from training bases, something that foreign models are still in the early stages of implementing.

In terms of ethics and transparency, European AIs comply with the high standards promoted by EU directives (such as the European Commission's Trusted AI Guidelines). Transparency means that the functioning of the model and its knowledge sources can be examined by third parties. Models such as Apertus and Mistral stand out for their open disclosure of their architecture, weights and comprehensive training documentation, including source code and descriptions of the data used.

This contrasts with proprietary models (e.g. OpenAI GPT-4 or Baidu Ernie), whose datasets and parameters are closed, making independent audits impossible. European openness not only allows for the inspection of biases or errors, but also facilitates legal and ethical

compliance audits by authorities or partners. In principle, this reflects Europe's commitment to transparency and accountability in AI, a value enshrined in both the GDPR (principle of accountability) and the upcoming AI Regulation.

Finally, European AIs are designed to operate within ethical boundaries acceptable in the EU, avoiding discrimination and respecting human dignity. The intentional linguistic diversity in the Apertus training (covering more than 1,000 languages, with 40% of the data in non-English languages ethz.ch) demonstrates a commitment to inclusion and reducing cultural or linguistic bias, in line with the European appreciation of multiculturalism. In short, by adopting European models, EuroBot benefits from solutions that are already GDPR-compliant, with requirements for ethics, transparency and respect for fundamental rights, minimising the risks of non-compliance and strengthening the institutional legitimacy of the project before the bodies of the European Union.

5.4 Overall system architecture

5.4.1 Architectural overview

The EuroBot architecture is based on decoupled microservices, designed to provide security, scalability, and auditability at all stages of the information lifecycle. It follows the Retrieval-Augmented Generation (RAG) standard with agentic orchestration, integrating processing, storage, search, and response generation into a fluid and traceable flow.

5.4.2 Core system components

The architecture consists of functional layers that interact in an orchestrated manner:

- Ingestion Layer: responsible for acquiring and validating legal documents.
- Processing Level: performs OCR, cleaning, semantic enrichment, and embedding generation.
- Indexing and Archiving Layer: Organizes data into vector, relational, and full-text databases.
- Orchestration and Agent Layer: directs requests to specialized legal agents.
- Generation and Delivery Level: produces structured and auditable responses.
- Observability Level: collects logs, metrics, and audit trails for continuous analysis.

5.4.3 Relationship between different modules

The modules are interconnected via an event architecture and RESTful APIs:

1. The user sends a query via the interface.
2. The query is processed and forwarded to the orchestration module.
3. The selected legal agent performs the hybrid search (vector, structural, metadata, and full text).
4. The result is reordered, formatted, and sent to the generative module.
5. The response is sent to the user and recorded for auditing purposes.

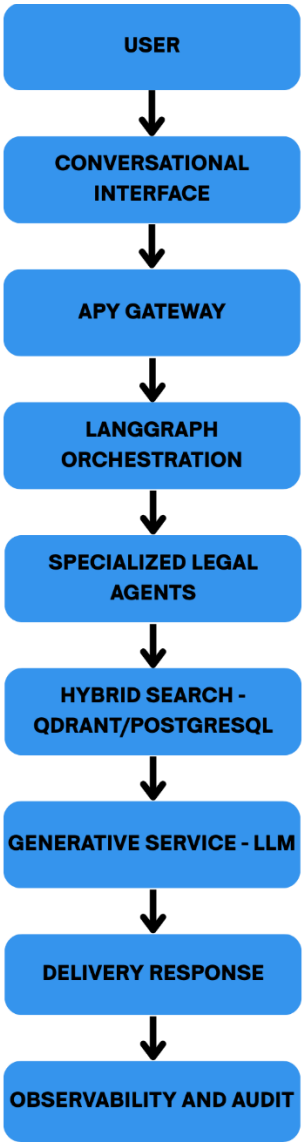


Figure 1: Relationship between different modules

5.4.4 Macro-level data flows

The macro data flow can be divided into two main axes: Knowledge Ingestion and Interactive Query.

a. Knowledge acquisition

1. Document loading.
2. Processing (OCR + semantic enrichment).
3. Embedding generation.
4. Indexing in vector and relational databases.

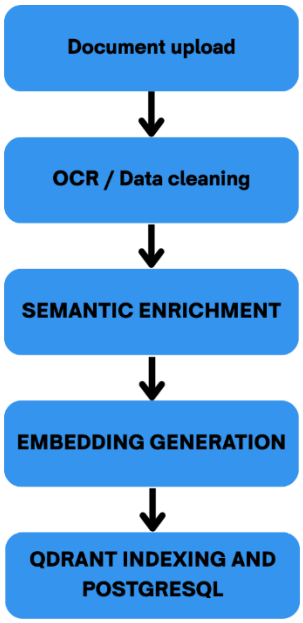


Figure 2: Knowledge acquisition

b. Interactive

1. The user submits a question.
2. The orchestrator identifies the legal domain.
3. The agent performs a hybrid search and reclassification.
4. The LLM generates the response.
5. The response is delivered and recorded.

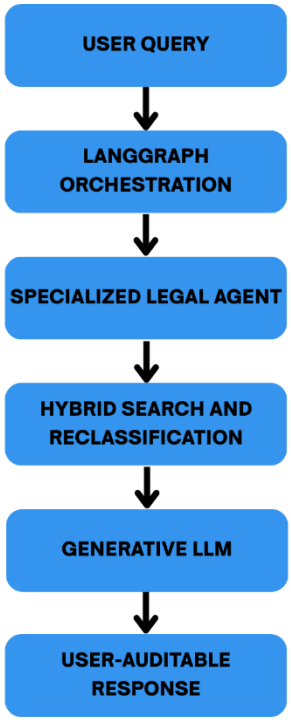


Figure 3: Interactive

5.5 Online vs. Offline

5.5.1 The online system

EuroLegalBot Online is the version of the platform that operates with cloud services, connecting to language models and databases hosted remotely. This configuration allows access to cutting-edge LLM models (such as OpenAI GPT-4o, Claude 3, Mistral, Gemini) via API, as well as the use of managed scalable infrastructures, such as Qdrant Cloud, AWS Lambda, Fargate, and EC2, according to the architecture described in the technical documents.

In online mode:

- Document ingestion is triggered by events, which are processed in serverless services such as AWS Lambda.
- Computing is distributed: microservices are orchestrated with Fargate containers, and vector and relational databases run on EC2 instances for compliance and performance reasons.
- The user interface is via API Gateway and web frontend, connecting to orchestration services and specialized agents for each area of law.
- The generative module communicates with LLM providers via the REST API, supporting secure calls and inbound masking for sensitive data protection.

Online architecture is ideal for environments with high network availability, demand for high-performance models, and a need for horizontal scalability. It allows you to incorporate proprietary models, obtain near real-time responses, and rely on enterprise support and automatic model and infrastructure updates.

5.5.2 The offline system

EuroBot Offline is designed to operate completely locally, without the need for an internet connection. Offline mode prioritizes local execution and data.

The client application (desktop) maintains all navigation, dialog, and document ingestion functionality even in disconnected environments, using local engines via extensions and an optional local server (web-app/src/providers/DataProvider.tsx:121).

The DataProvider initializes on-premises providers and can activate an internal API server to expose the user-loaded model, preserving conversation flow and context retrieval (RAG) in isolated environments. The RAG extension operates on a locally configured VectorDB, which ensures that all ingested documents and retrieved data remain on the client host, without any external transmission.

5.5.3 Functional differences between the two systems

Aspect	EuroLegalBot Online	EuroLegalBot Offline
Model source	Third-party APIs (OpenAI, Claude, Mistral, etc.)	Local engines loaded via extension
Data storage	Managed infrastructure (S3, RDS, Qdrant Cloud, etc.)	Qdrant/PostgreSQL on-premises on the client host
Connectivity	Requires a continuous internet connection	Operates completely without a network**
Scalability	High, via cloud infrastructure	Limited to the local capacity of machines
Privacy and sovereignty	Data can circulate via external providers	Full local control of data and logs
Latency and performance	High, depending on bandwidth and model	Variable, depending on local CPU/GPU
Hybrid scenarios	Allows local fallback for vector retrieval	Allows local models to be used with optional remote provider

Table 1: Functional differences between the two systems

**With the exception of voice data entry, which depends on interaction with a remote server, however, the latter has no interaction with external providers.

5.5.4 Curren limitations of each system

Online:

- Subject to provider policies, costs, and availability (e.g., API limits).
- Potential risk of data leakage in sensitive environments, requiring masking and data processing agreements (DPAs) with external providers.

Offline:

- Quality and latency depend on the model and local hardware.
- Without sufficient GPU or CPU, you must use smaller models or accept longer response times.
- Requires manual installation and updating of local extensions and engines.

5.5 Server infrastructure and technical settings

EuroLegalBot's server infrastructure has been designed to support an AI system applied to a sensitive (legal) domain, where requirements such as data sovereignty, auditability, security by design, and controlled scalability are as critical as performance. The configuration adopted configuration prioritizes execution and storage in a European jurisdiction environment, reducing the regulatory and operational risks associated with extraterritorial dependencies.

5.5.1 Server configuration

For the implementation of EuroBot (especially in institutional/online mode), an infrastructure based on Insiel was adopted, using Data Center functionality and Cloud/IaaS services aimed at regional public administration. Insiel is the "internal" ICT company of the Autonomous Region of Friuli Venezia Giulia (Italy), responsible for the design and management of digital infrastructures and critical services, with a particular focus on continuity and security.

The configuration adopted includes the following blocks at the logical level:

- Scalable computing level (instances/VMs and/or containerized execution for EuroBot services: API, orchestration, ingestion, and processing services);
- Data layer with separate persistence for the purpose:
    - PostgreSQL for structured data (metadata, operational audit, session/user control, logical versioning);
    - Qdrant for vector indexing (embedding), semantic retrieval support (RAG);

- Network and perimeter security, with segmentation between public services (e.g., gateways/APIs) and internal services (e.g., databases and indexes);
- Continuity and redundancy mechanisms (typical strategy for data centers with multiple sites and dedicated connectivity, where applicable).

In its qualification and governance plan, Insiel presents specific certifications and extensions for IT security and cloud services (e.g., ISO/IEC 27001 and extensions such as 27017/27018, according to institutional documentation), strengthening the reference framework for use in regulated scenarios.

5.5.2 reasons for choosing this configuration

The choice of Insiel as the main infrastructure is in line with EuroLegalBot's objectives of

operate with institutional control, European compliance, and traceability. The main motivators are:

a) European sovereignty and jurisdiction (data within the EU): Keeping workloads and data (including legal documents and vector indexes) in an environment located in Italy, within the European Space, reduces exposure to extraterritorial legal regimes and facilitates the consistent application of data protection requirements. Insiel operates as regional infrastructure, with a data center in Trieste and services aimed at public administration, strengthening governance and institutional predictability.

b) Compliance with security and cloud standards for the public sector: For sensitive environments, it is essential to demonstrate appropriate technical and organizational measures, supported by recognized standards. Insiel demonstrates security certifications (including ISO/IEC 27001 and cloud-oriented extensions and data protection), as well as qualification/compliance with specifications for cloud services in public administration (e.g., reference to qualification as a Cloud Service Provider in the AgID context).

c) Technical scalability geared towards data and user growth: EuroLegalBot has a progressive growth nature: more documents ⇒ a higher rate ⇒ a higher recovery volume ⇒ a higher inference/coordination load. The IaaS/Cloud configuration allows you to adjust computing and storage capacity in a controlled manner:
   - Horizontal scaling: adding instances/containers for stateless services (API, orchestration, ingestion), with load balancing and on-demand replication;
   - Vertical scaling: increasing resources (CPU/RAM/IOPS) on critical nodes (e.g., relational database, indexing nodes) when necessary;

- Workload separation: ingestion/processing can scale independently of conversational traffic, avoiding degradation when there are load spikes;
- Scalability by concurrent users: Increase the number of replicas of orchestration and query (retrieval) services to maintain SLAs in terms of latency and throughput.

These principles allow that, as the number of users and the volume of documents increase, the system's processing and response capacity also grows, maintaining operational predictability.

documents, the system's processing and response capacity also increases, maintaining operational predictability.

5.5.3 Performance and security implications

Performance

The configuration adopted has a positive impact on performance by:

- reducing internal latency between services (orchestration banks ↔ APIs ↔) by keeping components close together in the same network domain;
- enabling on-demand scalability at specific levels (for example, scale resources during ingestion during upload spikes without affecting query traffic)
- supporting resilient architectures with geographic/operational redundancy (e.g., reference to disparate data centers and connectivity), promoting service continuity.

Data security and protection

For a legal system, security must be treated as a set of controls:

- Access control: network segregation, RBAC by function (technical vs. institutional), and credential/secret management;
- Encryption: protection of data in transit (TLS) and at rest (when supported by the storage layer), with keys managed under institutional governance;
- Audit and traceability: structured logs of accesses, ingestions, queries, and relevant events, consistent with accountability requirements;
- Organizational controls and certifications: adherence to security standards (ISO/IEC 27001 and related extensions to cloud and data protection), increased trust, and ability to demonstrate compliance to stakeholders and audits.

In summary, the adoption of Insiel as the central infrastructure strengthens EuroBot's position as an institutional solution: data under European jurisdiction, appropriate controls and certifications, and the ability to scale with the growth of documents and users while maintaining governance and predictability.

6. Knowledge bank and dataset construction

6.1 Conceptual architecture of the knowledge bank

The knowledge bank represents the core component of both system development and chatbot training. It is designed as a modular and extensible repository capable of integrating heterogeneous legal materials while preserving legal coherence, provenance and versioning.

Content is organised thematically and by legal instrument, jurisdiction and institutional source, enabling efficient retrieval and contextualisation of information.

6.2 Dataset structure and composition

The dataset currently (02/02/2026) consists of approximately 6,339 legal documents organised into directory-based collections corresponding to courts, tribunals, legal portals and judicial institutions at national and European level. Each directory reflects the institutional origin of the documents and enables a preliminary classification by legal system, level of jurisdiction and territorial scope.

From a quantitative standpoint, the dataset shows a strong concentration of judicial documents originating from higher courts and appellate jurisdictions. The most represented sources include investigating chambers, supreme courts, constitutional courts and high regional courts, with particularly significant coverage of German, Polish, Italian, Spanish and European legal systems.

The heterogeneous distribution of files across directories reflects both institutional relevance and publication practices of the underlying sources. This structure makes the dataset especially suitable for comparative legal analysis, longitudinal studies of jurisprudence, and large-scale text mining and NLP applications on judicial texts.

6.2.1 File distribution by directory

Below is the list of directories included in the dataset, along with the corresponding number of files.

- Almaiura/ → 127 files

  Collection associated with academic or legal content, presumably doctrinal in nature or supporting case law.

- COAGD_of_Luxembourg_Investigating_Chamber/ → 777 files

  Documentation originating from the Investigating Chamber of Luxembourg, with a significant volume of judicial acts.

- Corte_di__Cassazione/ → 487 files

  Judgments and decisions of the Italian Court of Cassation.

- Curia/ → 37 files

  Limited number of documents, likely referring to institutional decisions or communications.

- European__Justice/ → 713 files

  Documents of a European scope, potentially including case law and acts of European Union institutions.

- HRC_of_Appeals_Celle_/ → 63 files
- HRC_of_Appeals_Frankfurt/ → 78 files
- HRC_of_Appeals_Hamm/ → 184 files
- HRC_of_Appeals_Munich_/ → 81 files
- HRC_of_Appeals_Saarbrucken_/ → 412 files
- HRC_of_Appeals_Schleswig_Holstein_/ → 173 files
- HRC_of_Appeals_Stuttgart_/ → 559 files

Group of directories referring to the German High Regional Courts (Oberlandesgerichte), with a heterogeneous distribution of decisions across locations.

- Legal_acts_of__Republic_of__Moldova_National__Court's_Web__Portal_of__Ministry_of__Justice_of__the_Republic__of__Moldova/ → 40 files

  Legal acts originating from the official Moldovan portal of the Ministry of Justice.

- Lex_(Acceso_a_la_legislación_de_la_Unión_Europea)/ → 528 files

  Legislative and legal documents extracted from the EUR-Lex portal.

- Netherland_/ → 353 files

Collection of decisions or legal acts from the Dutch legal system.

- Niedersachse_/ → 2 files

  Marginal directory with an extremely limited number of documents.

- Portal__Orzeczen__Poland/ → 28 files

  Judgments extracted from the Polish judicial decisions portal.

- Portale_del_Massimario/ → 279 files

  Legal maxims and summaries of case law, presumably from the Italian context.

- Public_Spanish_/ → 611 files

  Documentation of Spanish origin, with a substantial volume of public or judicial acts.

- Supreme_Court_of_Poland/ → 674 files

  Judgments of the Polish Supreme Court.

- The__federal__Consitutional__Court/ → 133 files

  Decisions of the German Federal Constitutional Court.

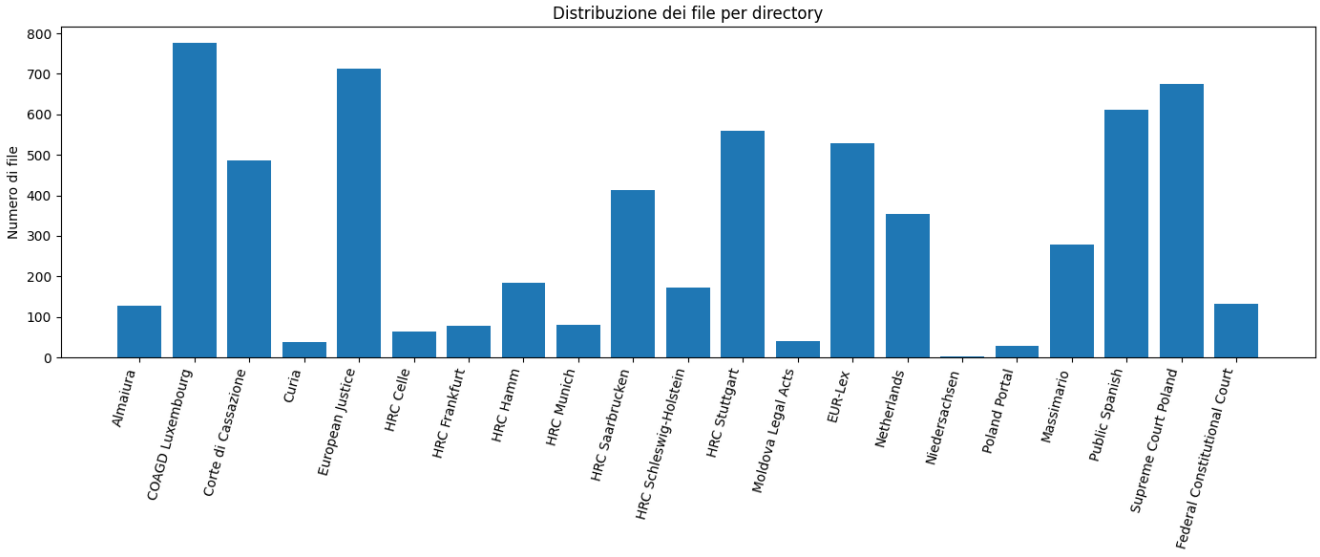


Figure 4: File distribution by directory

6.3 Web-based legal sources and scraping pipeline

A significant portion of the knowledge bank is derived from authoritative and publicly accessible web-based sources related to the European Arrest Warrant and judicial cooperation in criminal

matters. These sources include institutional and official websites at EU and national level, such as European justice portals, national court repositories and legislative databases.

From a technical perspective, data acquisition is implemented through an automated web scraping pipeline specifically developed to query institutional sources in a controlled and repeatable manner. The pipeline is based on scripted crawlers that target predefined URLs corresponding to courts, tribunals and legal portals, as identified during the dataset definition phase.

The scraping process is executed on a weekly schedule. This periodic execution enables:

- progressive acquisition of newly published judicial decisions and legal acts;
- detection and ingestion of updates or amendments to previously collected documents;
- continuous expansion of the dataset while preserving temporal alignment.

Scraped documents are stored in a directory-based structure mirroring their institutional origin (e.g. supreme courts, courts of appeal, constitutional courts, EU-level institutions). Each directory corresponds to a specific source or judicial body and contains a variable number of files, reflecting the publication frequency and output volume of the source itself.

This organisational choice supports several technical objectives: provenance tracking, jurisdiction-based filtering, preliminary legal classification and efficient batch processing during ingestion, indexing and re-indexing phases.

6.4 Case law repositories

The knowledge bank incorporates an extensive corpus of judicial decisions issued by European and national courts. At European level, the dataset includes materials originating from EU judicial institutions and European justice portals. At national level, it integrates decisions from Supreme Courts, Constitutional Courts and Courts of Appeal of selected EU Member States.

From a technical and dataset-design perspective, the case law corpus is characterised by a directory-based segmentation reflecting individual courts or judicial bodies. Each segment contains a varying number of documents, ranging from small collections to directories with several hundred files, depending on the output of the respective institution.

This segmentation enables efficient incremental updates, selective re-indexing and jurisdiction-specific retrieval strategies. Combined with the weekly scraping schedule, it supports a dynamic and continuously expanding repository that remains aligned with evolving jurisprudence while preserving historical versions for audit and traceability purposes.

6.5 Legislative and regulatory frameworks

In addition to case law, the knowledge bank includes EU legislation, national implementing laws and relevant regulatory frameworks governing the European Arrest Warrant and judicial cooperation in criminal matters. This ensures that responses are grounded in applicable legal norms across jurisdictions.

6.6 Data selection, quality control and updating strategy

All collected data undergoes a structured selection and quality control process. Priority is given to official, authoritative and publicly accessible sources. Automated checks are applied during ingestion to ensure format consistency, source traceability and completeness of extracted content.

The periodic scraping schedule enables systematic updating of the corpus, reducing the risk of outdated information while preserving historical versions of documents for audit and traceability purposes.

6.7 Dataset lifecycle and ingestion workflow

From a system development perspective, the dataset is managed as a living asset following a clearly defined lifecycle. The workflow connects data acquisition, preprocessing, storage and retrieval within a unified pipeline that supports both system robustness and legal reliability.

The dataset lifecycle is articulated through a sequence of well-defined technical phases, depicted in Figure 5, which provides a consolidated view of the end-to-end process supporting the EuroLegalBot knowledge bank.

The lifecycle can be summarised as follows:

- Authoritative legal sources: institutional and official websites at EU and national level (courts, tribunals, judicial portals and legislative repositories) constitute the primary data sources.

- Source configuration: authoritative sources are identified and configured as predefined scraping targets, with URL-level control and institutional attribution.
- Automated acquisition (weekly): scripted crawlers periodically retrieve new or updated documents on a weekly basis, ensuring temporal alignment of the corpus.
- Raw document persistence: collected files are stored in directory-based collections reflecting institutional origin, jurisdiction and court level.
- Preprocessing and normalisation: raw texts are cleaned, normalised and enriched with basic metadata to ensure consistency and downstream usability.
- Versioning and incremental updates: new versions of documents are added without overwriting historical data, preserving traceability and auditability.
- Knowledge bank and indexing: processed documents are integrated into the knowledge bank and prepared for retrieval and RAG-based interaction.

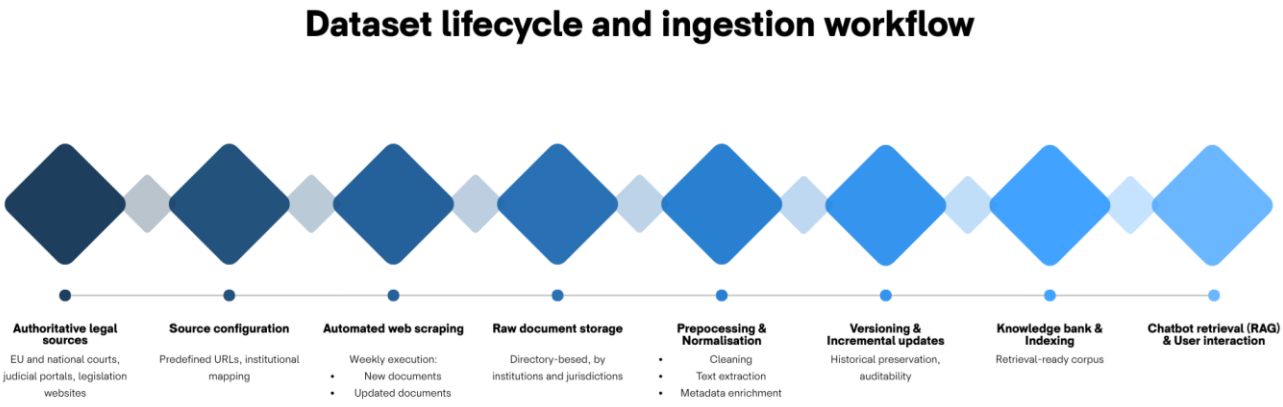


Figure 5: Dataset lifecycle and ingestion workflow

This end-to-end workflow ensures that EuroLegalBot operates on a continuously updated, structured and verifiable legal corpus, bridging system development and chatbot training in a coherent and controlled manner.

7. Chatbot training methodology

7.1 Training objectives

The primary objective of the chatbot training is to enable accurate, context-aware and multilingual responses to legal questions posed by practitioners involved in cross-border judicial cooperation. Typical use cases include clarification of procedural requirements, interpretation of legal provisions and consultation of relevant case law.

7.2 Data pre-processing and structuring

Before being used for training, raw legal data is cleaned, normalised and structured. This includes categorisation by legal theme, legal instrument and jurisdiction, improving retrieval accuracy and contextual relevance.

7.3 Expert-driven question design

Legal experts, including lawyers and magistrates, were actively involved in the chatbot training process through the design of a structured set of legal questions. These questions were conceived to reflect real-life professional scenarios and typical information needs arising in the context of cross-border judicial cooperation.

The questions will be directly submitted to the chatbot during the training, testing and fine-tuning phases (task T3.4 – *Chatbot internal testing and fine-tuning* and deliverable D3.4 – *Chatbot internal testing and fine-tuning*) and were formulated in the languages of the consortium, thereby contributing both to functional validation and to multilingual performance assessment.

The expert-driven questions serve multiple purposes. First of all, they test the accuracy and completeness of the chatbot's responses. Secondly, they assist in identifying gaps or ambiguities in the knowledge base. Finally, they intend to support iterative refinement of the system.

The list below provides an overview of the questions used in the training process:

1. For the purposes of the passive surrender procedure, what information must the European Arrest Warrant contain?
2. For the purposes of the passive surrender procedure, what additional information must the European Arrest Warrant contain when it is issued for the execution of a custodial sentence or a custodial security measure imposed following proceedings in which the person concerned did not appear in person?
3. For which categories of offences is the European Arrest Warrant executed regardless of double criminality?
4. In what form and before which authority may consent to surrender be expressed?
5. What are the grounds for mandatory and optional refusal of surrender?
6. For the purposes of verifying the lawful and effective residence or domicile on Italian territory of the person requested for surrender, which elements must be assessed by the competent Court of Appeal?

7. In the event of competing requests for surrender, which elements must be assessed in order to decide which warrant is to be executed?
8. What procedure must be followed where, in respect of the same person, both a European Arrest Warrant and an extradition request from a third State have been issued?
9. In the active surrender procedure, which authority is competent to issue the European Arrest Warrant?

7.4 Internal testing and fine-tuning

In the final phase of the Work Package, dedicated to internal testing and fine-tuning, the expert-designed questions will be systematically tested against the chatbot in order to assess the accuracy, completeness and legal soundness of its responses. This phase will also support the identification of residual gaps or ambiguities within the knowledge base and response logic.

In addition to the questions already used during the initial training and validation stages, new questions will be progressively added during the internal testing and fine-tuning phase. These additional questions are currently being prepared by the legal experts involved in the project and are intended to further expand thematic coverage, increase linguistic diversity and enhance the overall robustness of the chatbot.

8. Conclusions

This deliverable has documented the design, development and training of EuroLegalBot, highlighting the technical, methodological and strategic choices underpinning the system. The project demonstrates a concrete and viable approach to the application of Artificial Intelligence in sensitive legal and institutional environments, where accuracy, transparency, explainability and data sovereignty are essential requirements.

By adopting a modular Retrieval-Augmented Generation architecture combined with deterministic multi-agent orchestration, EuroLegalBot overcomes many of the limitations associated with generic AI solutions. All responses are grounded in pre-ingested and authoritative legal sources, ensuring full traceability, auditability and legal reliability.

The structured and continuously updated dataset, supported by an automated scraping and ingestion pipeline, represents a key asset of the system. Treating the dataset as a living resource with clear lifecycle management and versioning allows EuroLegalBot to remain aligned with evolving legislation and jurisprudence while preserving institutional robustness and accountability.

The training methodology, based on expert-driven question design and iterative internal testing, further strengthens the legal soundness and practical relevance of the chatbot. This approach ensures that EuroLegalBot effectively supports legal professionals in real-world scenarios without automating or replacing human legal reasoning and decision-making.

In conclusion, EuroLegalBot provides a solid technological foundation for trustworthy AI in judicial cooperation. It illustrates how European values—such as legal certainty, transparency, accountability and digital sovereignty—can be effectively embedded into AI systems. The results achieved within this project offer a scalable and reusable model for future European initiatives aiming to deploy Artificial Intelligence in high-risk and institutionally sensitive domains.

EuroLegalBot
CRIMINAL MATTERS CHATBOT

EuroLegalBot

Project

Enhancing Judicial Cooperation through a European Arrest Warrant (EAW) Chatbot

Project partners

CORTE DI APPELLO DI VENEZIA

EUROPEAN EXPERTISE & EXPERT INSTITUTE
Shaping Expertise across European Justice Systems

RES PVBLICA 1995 ATHINA

AGENFOR INTERNATIONAL

(expleo)

EPLO
European Public Law Organization