AIGOV

Implementing ethical, trustworthy and fair Artificial Intelligence Systems in Public Sector

D3.1 Evaluation Results and Lessons Learnt

Version-Status:	V1.0 Final
Date:	30/09/2025
Dissemination level:	PU

The research project was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the **"2nd Call for H.F.R.I. Research Projects to support Faculty Members & Researchers"** (Project Number: 2412).

Deliverable factsheet

Project Number:	2412
Project Acronym:	AIGOV
Project Title:	Implementing ethical, trustworthy and fair Artificial Intelligence Systems in Public Sector
Principal Investigator:	Konstantinos Tarabanis
Scientific Area:	SA9. Management & Economics of Innovations
Scientific Field:	9.4 ICT enabled Innovation, Digitisation and Industrial Renewal
Host Institution:	University of Macedonia
Collaborating Organization:	Region of Central Macedonia, Greece

Deliverable title:	D3.1 Evaluation Results and Lessons Learnt
Deliverable number:	D3.1
Official submission date:	30/09/2025
Actual submission date:	30/09/2025
Author(s):	Areti Karamanou, Dimitrios Zeginis, Maria Zotou, Evangelos Kalampokis, Konstantinos Tarabanis

Abstract:	This document is the deliverable, entitled D3.1 “D3.1 Evaluation Results and Lessons Learnt”, of the third work package of the AIGOV project. It reports the results of T3.1, T3.2, and T3.3 of WP3 regarding the pilot operation, the results of the evaluation activities along with the lessons learnt from the pilot operation and evaluation.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

The following table presents the acronyms used in the deliverable in alphabetical order.

Abbreviation	*Description*
AI	Artificial Intelligence
DG	Directorate-General
EC	European Commission
EU	European Union
GDPR	General Data Protection Regulation
GNN	Graph Neural Network
KPI	Key Performance Indicator
LLM	Large Language Model
ML	Machine Learning
NLP	Natural Language Processing
PCS	Pilot Case Study
R&D	Research and Development
SME	Small Medium Enterprise
WP	Work Package
XAI	Explainable Artificial Intelligence

Executive Summary

The AIGOV project conducted four pilot studies to assess the practical application of responsible, transparent and trustworthy Artificial Intelligence (AI) in public-sector contexts. These pilots evaluated technical performance, interpretability, fairness, legal compliance and organisational feasibility, generating valuable insights that informed both the methodological frameworks developed in earlier work packages and future adoption pathways for AI in government.

Across the pilots, predefined KPIs and evaluation strategies were applied to ensure that results were comparable, rigorous and contextually meaningful. The key findings are summarised below:

Pilot 1 - Explainable Graph Neural Networks for Housing Policy Analysis. The use of Explainable Graph Neural Networks demonstrated strong potential for uncovering complex socio-economic patterns and supporting evidence-based policymaking. Although predictive accuracy was satisfactory, the primary added value stemmed from the interpretability of results.

Lesson learnt: Interpretability is essential for adoption in public administration; further effort is required to ensure data quality and stability of explanations.

Pilot 2 - Evaluation of Large Language Models for Legal Text Interpretation. Large Language Models (LLMs) exhibited promising capabilities in extracting legal concepts and relationships from legislative and regulatory texts. Performance varied across models and prompting techniques.

Lesson learnt: Deploying LLMs in legal workflows requires stringent human oversight, risk assessment and alignment with regulatory and ethical requirements.

Pilot 3 - Application of LLMs in Legal Documentation and Evidence-Based Analysis. The pilot showed notable efficiency gains in generating summaries, structuring legal documents and accelerating research tasks. However, challenges emerged regarding thematic bias and the need for domain-specific tuning.

Lesson learnt: LLMs deliver the greatest value when used as assistive tools rather than autonomous sources of legal reasoning.

Pilot 4 - Generative AI for Multilingual and Multicultural Democratic Consultation

Generative AI enabled more inclusive, multilingual and accessible public consultations. Key challenges included ensuring cultural neutrality and managing biases in generated content.

Lesson learnt: AI can strengthen democratic participation, but must be accompanied by robust oversight mechanisms and safeguards for representational fairness.

Across all pilots, several overarching lessons emerged:

- Data governance is foundational: AI performance and fairness depend critically on high-quality, well-structured and ethically governed datasets.
- Interpretability and transparency are non-negotiable: Public-sector agencies require clear, explainable outputs to support trust, accountability and decision-making.
- Organisational readiness matters: Successful adoption necessitates staff training, clarifying roles, and embedding AI workflows into public-sector processes.
- Compliance by design: Legal, ethical and societal considerations must be integrated from the earliest stages of system design.

1 Introduction

The aim of this section is to introduce the background of the work pursued within Task 3.1 "Evaluation Strategy", Task 3.2 "Pilot operation", and Task 3.3 "Evaluation and Lessons learnt" of the AIGOV project. The scope and the objective that the current document has set out to achieve are presented in sub-section 1.1. The intended audience for this document is described in sub-section 1.2, while sub-section 1.3 presents the relationship of this deliverable with previous deliverables of the project. Finally, sub-section 1.4 outlines the structure of the rest of the document.

1.1 Scope

This document constitutes the deliverable "Evaluation Results and Lessons Learnt AIGOV Holistic Framework" (henceforth, referred to as D3.1) of the AIGOV project.

The main objective of D3.1 is to present the results of the three tasks performed under WP3 of the AIGOV project, namely

- Task 3.1 "Evaluation Strategy",
- Task 3.2 "Pilot operation", and
- Task 3.3 "Evaluation and Lessons learnt".

The specific objectives of WP3, therefore, include:

- To design a concrete evaluation framework that can be used in order to assess the quality of AIGOV frameworks, and how this applied to the CO.
- To execute and assess the impact of the pilot organised by the CO
- To identify important lessons learnt that can help public administrations in applying AI technologies and using AIGOV frameworks.

1.2 Intended Audience of this Deliverable

The intended audience for this document is public administration, policy-makers, and anyone interested in deploying Artificial Intelligence in the public sector.

1.3 Relationship with Previous Deliverables

Deliverable D3.1 finalises the development path initiated in WP1 and continued in WP2 of the project AIGOV. Specifically, D1.1 mapped the state of AI governance and public-sector readiness, while D1.2 defined the AIGOV Ecosystem, a conceptual model built on four pillars (Collection, Construction, Evaluation, Translation) and four components (Data, Algorithms, Models, Applications) with various internal and external types of stakeholders. These artefacts directly informed the frameworks developed in D2.1, which formalized the ethical and procedural dimensions of trustworthy AI through the AIGOV Data Value Cycle, the Framework for Trustworthy, Fair, and Accountable AI and the AIGOV Transformation and

Adoption Framework. D3.1 now evaluates the artefacts of D2.1 through four case studies, closing the AIGOV innovation cycle from design to real-world validation.

1.4 Structure

The structure of the document is as follows:

- Section 2 provides an overview of the methodological foundations underpinning the evaluation activities, including the AIGOV Transformation and Adoption Framework, the Government Data Value Cycle, and the AIGOV Framework for Trustworthy, Fair, and Accountable AI. These frameworks establish the conceptual basis for assessing AI systems in public-sector environments.
- Section 3 presents the evaluation strategy applied across all Pilot Case Studies. It outlines the baseline assessment approach, the definition of evaluation scenarios, and the key performance indicators (KPIs) used to ensure coherent, comparable, and responsible evaluation across domains.
- Section 4 details the operation and evaluation of the Pilot Case Studies. For each pilot, the section follows the phases of the AIGOV Transformation and Adoption Framework, covering readiness assessment, responsible use-case design, build-and-test activities, and evaluation. Each pilot demonstrates the application of AIGOV principles in different public-sector contexts, from housing analytics to legal reasoning and multilingual deliberation.
- Section 5 synthesises the lessons learned across the pilot cases. It highlights how the pilots collectively contribute to the development of actionable guidance for trustworthy and value-driven AI adoption in public administrations.
- Section 6 concludes the document, summarising the key insights, policy implications, and recommendations for future work, particularly regarding scaling, governance, and sustainable implementation of AI systems in the public sector.

2 Overview of the WP2 AIGOV Frameworks Underpinning the Evaluation

The evaluation activities carried out in WP3 are grounded in the methodological artefacts developed in WP2 and documented in Deliverable D2.1. Together, these artefacts:

- the AIGOV Government Data Value Cycle (Section 2.1),
- the AIGOV Framework for Trustworthy, Fair, and Accountable AI (Section 2.2), and
- the AIGOV Transformation and Adoption Framework (Section 2.3)

form the AIGOV Holistic Framework, which provides the conceptual, ethical, and operational foundation for the pilot activities conducted in WP3.

This section presents a concise overview of the three frameworks to establish the methodological basis on which the WP3 evaluation strategy was built. Each framework contributes a different but complementary perspective:

- the Data Value Cycle defines the requirements and processes for creating high-quality, interoperable, and ethically governed public-sector data;
- the Trustworthy AI Framework specifies the principles and guidelines necessary to ensure transparency, fairness, accountability, and inclusiveness in AI systems; and
- the Transformation and Adoption Framework outlines the organisational and procedural steps required for responsible and sustainable AI integration into public services.

By briefly revisiting these frameworks, this section clarifies how they shape the evaluation criteria, pilot design considerations, and lessons learnt presented in later sections of this deliverable.

2.1 The AIGOV Government Data Value Cycle

The AIGOV Government Data Value Cycle (Figure 1) provides a comprehensive and structured approach for managing public-sector data in ways that enable ethical, high-quality, and AI-ready data ecosystems. It recognises that governments rely on multiple heterogeneous datasets, structured, unstructured, and dynamic, and that responsible data management is essential for producing trustworthy AI outcomes. Inspired by international best practices (e.g., OECD, EU Data Strategy, SEMIC), the cycle operationalises the full data lifecycle into seven interconnected stages, each accompanied by requirements, challenges, and guidelines tailored to public-sector realities.

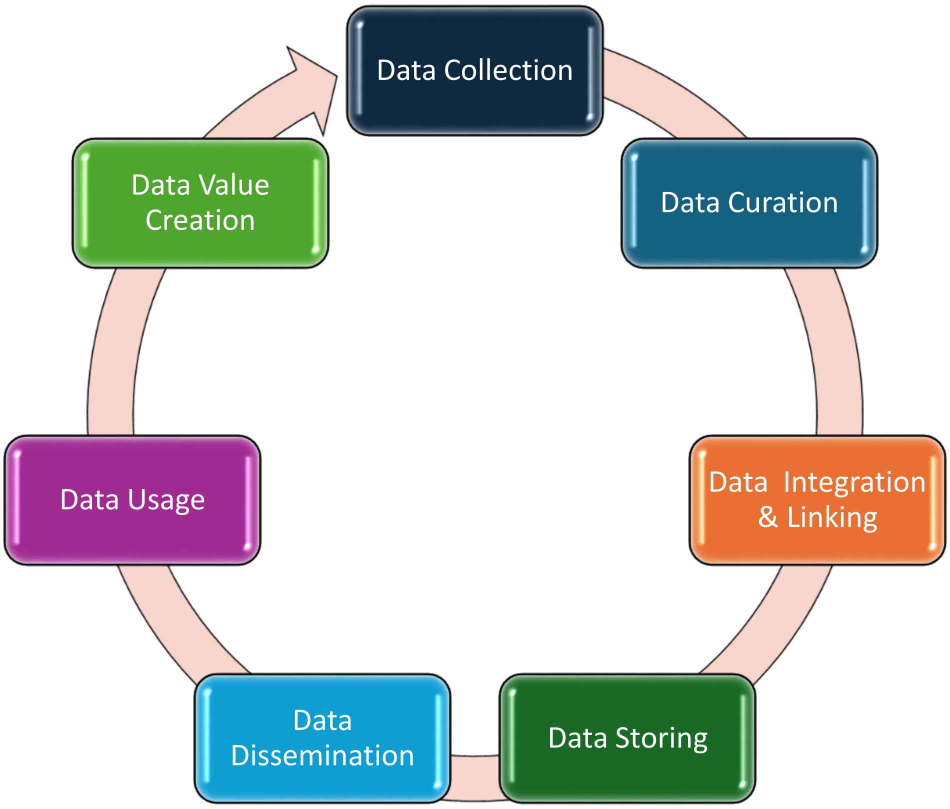


Figure 1 The AIGOV Data Value Cycle

The first stage, **Data Collection**, concerns the acquisition of diverse data sources, including administrative records, sensor data, textual documents, geospatial datasets, and external partner data. In the context of Generative AI, the scope of relevant data expands significantly, as large language models rely heavily on unstructured text (PDFs, reports, clinical notes, legal documents). This stage requires establishing legal and ethical grounds for data access, ensuring compliance with GDPR, and enabling access to high-value datasets such as dynamic sensor and mobility data.

The second stage, **Data Curation**, refers to activities that ensure data accuracy, quality, completeness, representativeness, and readiness for reuse. Public-sector data often suffer from inconsistencies due to legacy systems, human-generated errors, missing entries, or outdated formats. Curation includes cleaning, annotation, enrichment, validation, and documentation of all transformations. It also requires identifying and mitigating biases that may propagate unfairness into AI outputs. In the era of LLMs, textual data may need preprocessing (OCR, entity extraction, segmentation), and dynamic data may require continuous anomaly detection.

The third stage, **Data Integration & Linking**, addresses the longstanding challenge of siloed government information. By using common vocabularies, shared ontologies, and semantic standards (DCAT-AP, RDF, OWL, SEMIC Core Vocabularies), administrations can link datasets across departments and levels of government. Integration is essential for developing a holistic view of public services and for enabling AI systems to reason across domains. Privacy-preserving techniques such as pseudonymisation and federated analysis are critical when combining sensitive datasets.

The fourth stage, **Data Storing**, ensures that governments maintain secure, scalable, and sustainable data infrastructures. Beyond traditional relational databases, Generative AI requires support for vector databases that store embeddings and enable semantic search. Metadata catalogues, provenance documentation, backup policies, encryption, and access control mechanisms must be in place. Data storage solutions should comply with GDPR, national data policies, and principles of privacy- and security-by-design.

The fifth stage, **Data Dissemination**, turns internal datasets into public assets by ensuring discoverability, accessibility, interoperability, and reuse. Dissemination requires clear licensing, machine-readable formats, multilingual support, and dissemination pipelines for real-time data via APIs. By adhering to open data directives and FAIR principles, administrations promote transparency, innovation, and public trust. Rigorous quality assurance and risk-based anonymisation are critical before releasing data publicly.

The sixth stage, **Data Usage**, focuses on the ethical and purposeful use of data by public authorities. It emphasises the importance of securing a social licence for data-driven decision making, ensuring meaningful human oversight, and preventing algorithmic harm. This stage evaluates whether data are being used proportionately, transparently, and in ways that serve public value objectives.

Finally, **Data Value Creation** constitutes the transformation of managed data into concrete benefits such as improved public services, faster administrative procedures, better-informed policies, and enhanced citizen understanding. Generative AI expands this potential dramatically by enabling advanced applications such as semantic search, summarisation, automated document drafting, decision support, and virtual public-service assistants.

Together, these seven stages form a coherent model that guides public administrations in creating the foundations required for responsible AI development and deployment.

2.2 The AIGOV Framework for trustworthy, fair, and accountable AI

The AIGOV Framework for Trustworthy, Fair, and Accountable AI defines the ethical, governance, and operational requirements for deploying AI, especially Generative AI, in public administration contexts. Recognising that AI systems can influence decisions affecting citizens' rights, access to services, and legal entitlements, the framework was designed to ensure that AI implementations support democratic values, comply with EU regulations (e.g., AI Act, GDPR), and preserve public trust. It is structured around four pillars, each containing concrete guidelines and associated assessment tools:

1. Transparency and accountability,

2. Data management and access,
3. Comprehensibility and multilingual support, and
4. Data Interoperability and reusability.

The first pillar, **Transparency and Accountability**, establishes that AI systems must be explainable, auditable, and overseen by humans. Public administrations cannot rely on black-box models that produce opaque outcomes. Instead, models should provide understandable rationales for their outputs, enabling scrutiny by civil servants, auditors, and citizens. This includes the documentation of model architecture, datasets, assumptions, limitations, and known risks. Accountability mechanisms must clearly allocate responsibilities, ensuring that administrative and political actors retain oversight of AI-supported decisions. Methods such as SHAP, LIME, model cards, and transparency registers support this pillar.

The second pillar, **Responsible Data Management and Access**, emphasises that trustworthy AI depends on high-quality data. Data must be accurate, timely, granular, ethically collected, and accessible through interoperable mechanisms such as APIs. Public administrations must manage sensitive and personal data responsibly, ensuring compliance with data protection laws and preventing misuse. Dynamic data streams, such as sensors, mobility data, or environmental monitoring, require robust pipelines that maintain integrity, timeliness, and metadata documentation. This pillar connects directly to the AIGOV Government Data Value Cycle, ensuring coherent governance across the entire AI lifecycle.

The third pillar, **Comprehensibility and Multilingual Support**, recognises the linguistic and cultural diversity of European societies. AI outputs must be understandable to people with different languages, literacy levels, and educational backgrounds. This requirement is essential for inclusive public services. Generative AI systems must be able to generate accurate and culturally appropriate responses in multiple languages, including under-represented ones such as Greek. Evaluating multilingual performance, cultural sensitivity, and clarity of outputs is therefore central to this pillar. Techniques such as instruction tuning, multilingual fine-tuning, and human-in-the-loop validation support implementation.

The fourth pillar, **Interoperability and Reusability**, ensures that AI systems are sustainable and capable of evolving. Public administrations cannot afford fragmented or proprietary solutions that cannot interoperate with existing systems or datasets. This pillar requires the use of open standards, semantic models, modular architectures, and reusable components. By aligning with the European Interoperability Framework (EIF) and FAIR principles, administrations can avoid lock-in, reduce costs, and enable cross-sector collaboration. In the context of LLMs, this includes the use of interoperable datasets, linked-data formats, and integration frameworks (e.g., LangChain, LlamaIndex).

2.3 The AIGOV Transformation and Adoption Framework

The AIGOV Transformation and Adoption Framework aims to facilitate public authorities to explore how public services will need to be redesigned to leverage the impact of AI. It supports public administrations to determine current strengths and weaknesses, set achievable goals, and construct transformation plans by taking into account all the AIGOV methods, guidelines, and tools in order to achieve ethical, trustworthy, and fair adoption of AI technologies, with particular emphasis on Generative AI.

The AIGOV Transformation and Adoption Framework integrates the AIGOV Government Data Value Cycle (T2.1) and the AIGOV Framework for Trustworthy, Fair, and Accountable AI (T2.2), translating these into actionable steps to guide ethical, fair, and sustainable AI adoption in the public sector.

The AIGOV Transformation and Adoption Framework supports public administrations in redesigning their services and organisational processes to leverage AI responsibly and effectively. While the first two frameworks address data governance and ethical AI design, this third framework focuses on the organisational, strategic, and operational changes required to ensure that AI adoption is feasible, beneficial, and aligned with public interest. The framework is structured as a five-phase lifecycle, providing a clear roadmap for transitioning from initial consideration to sustainable long-term use.

The first phase, **Assess Readiness and Context**, evaluates whether AI is appropriate for the intended public-sector challenge. Many administrative problems do not require AI, so this phase emphasises proportionality and public value. It assesses strategic alignment, stakeholder needs, data readiness (based on the AIGOV Data Value Cycle), legal and ethical compliance, and organisational capability. Readiness assessment tools help identify strengths, weaknesses, and prerequisites before moving forward.

The second phase, **Design Ethical and Value-Aligned Use Cases**, focuses on defining how AI can meaningfully improve public services. This phase requires translating administrative pain points into concrete AI-enabled interventions that respect fairness, transparency, inclusiveness, and multilingual requirements. Tools such as service blueprints, process maps, problem framing, and requirement analysis help ensure that AI interventions are not only technically feasible but also aligned with public values and user needs. Clear impact metrics, success criteria, and risk mitigations are defined here.

The third phase, **Build and Test**, transforms conceptual use cases into functional prototypes. This includes model development, data integration, interface design, and technical implementation using appropriate AI techniques (e.g., LLMs, RAG pipelines, classifiers). Ethical and legal testing plays a central role, including bias assessment, explainability checks, GDPR compliance verification, and validation of human-in-the-loop mechanisms. User testing with civil servants, domain experts, and (where applicable) citizens ensures usability, clarity, and relevance. Iterative improvements refine the system before operational deployment.

The fourth phase, Deploy and Scale Responsibly, concerns operational integration of the AI system into real administrative workflows. Deployment is conducted gradually through controlled pilots, monitoring, and feedback loops that track performance, fairness, accuracy, and user satisfaction. Capacity building is essential: public servants must receive training in interpreting AI outputs, managing risks, and collaborating with AI systems. Governance structures, escalation protocols, and accountability measures must be in place before broad deployment.

The fifth phase, **Govern, Evaluate, and Sustain**, establishes long-term oversight mechanisms and ensures continuous improvement. AI systems require ongoing monitoring because data distributions change, regulations evolve, and societal expectations shift. This phase includes periodic audits, model updates, documentation management, compliance reviews, and stakeholder feedback channels. Lessons learnt from pilot deployments inform broader organisational transformation, future AI initiatives, and the refinement of internal guidelines.

3 Evaluation Strategy

This section defines the unified evaluation strategy for the pilot case studies of the AIGOV project, in accordance with Task 3.1 of the AIGOV project. It establishes a coherent assessment framework that integrates strategic context, baseline analysis, methodological foundations, and Key Performance Indicators (KPIs) across pilots. The strategy ensures that each pilot is evaluated in a systematic, transparent, and comparable manner while respecting its unique objectives, data sources, and technological focus.

The evaluation strategy builds upon three core components:

- The AIGOV Transformation and Adoption Framework, guiding the phased analysis of readiness, design, implementation, and impact assessment.
- The AIGOV Government Data Value Cycle, ensuring that data quality, interoperability, governance, and responsible data use are consistently embedded.
- The Framework for Trustworthy, Fair, and Accountable AI, defining principles for legality, ethics, transparency, fairness, human oversight, explainability, robustness, and socio-technical alignment.

Specifically, the evaluation of the pilot case studies is anchored in Phases 1-3 of the AIGOV Transformation and Adoption Framework, which provides the principal methodological structure for assessing organisational readiness, responsible design, and controlled testing of AI systems in public-sector contexts. Given that WP3 focuses on pre-deployment evaluation rather than large-scale operationalisation, these phases constitute the most relevant subset of the framework for guiding empirical assessment activities.

Crucially, the Transformation and Adoption Framework is designed as an integrating framework: it operationalises and synthesises the two additional conceptual artefacts developed in WP2, namely the AIGOV Government Data Value Cycle and the AIGOV Framework for Trustworthy, Fair, and Accountable AI. Consequently, the evaluation conducted in WP3 is not limited to procedural criteria but systematically incorporates the data-governance and ethical-principles dimensions embedded within the AIGOV methodological family.

More specifically, the AIGOV Government Data Value Cycle contributes the evaluative criteria pertaining to data availability, quality, curation, interoperability, governance, and responsible use. These dimensions are instrumental for assessing whether the technical and organisational preconditions for AI deployment are met within each pilot context.

The AIGOV Framework for Trustworthy, Fair, and Accountable AI provides the normative requirements for lawful, ethical, transparent, equitable, and socially aligned AI operation. These criteria include explainability, transparency, fairness, multilingual accessibility, privacy protection, and appropriate human oversight, principles that are indispensable for assessing AI systems deployed in public administration.

Within WP3, these frameworks function in a complementary manner. The Transformation and Adoption Framework determines the procedural logic of the evaluation (i.e., how readiness, design, and testing are assessed), while the Data Value Cycle and the Trustworthy AI Framework specify the substantive criteria that each pilot must satisfy for responsible AI adoption. In other words, the procedural framework defines the structure of evaluation, and the supporting frameworks define its content.

Accordingly, each pilot is systematically mapped against:

- Phases 1-3 of the AIGOV Transformation and Adoption Framework, including readiness assessment, value-aligned use-case design, and controlled prototyping
- The requirements, challenges, and guidelines of the AIGOV Government Data Value Cycle
- The pillars and guidelines of the AIGOV Framework for Trustworthy, Fair, and Accountable AI

3.1 Purpose and Scope of the Evaluation Strategy

Based on the description of Task3.1, the Evaluation Strategy aims to:

- Define the baseline, including strategic context, stakeholders, operational challenges, organisational capacity, data readiness, and legal constraints for each pilot.
- Specify clear and traceable evaluation criteria, including cross-pilot and pilot-specific KPIs.
- Establish unified evaluation scenarios to assess technological, organisational, ethical, and socio-economic implications.
- Ensure feedback loops into the design and implementation of tools, frameworks, and methods developed in AIGOV.
- Safeguard public value, ensuring that AI applications contribute to transparency, inclusiveness, fairness, and accountability.

The strategy applies uniformly to all pilots while accommodating differences in maturity, domain, technical scope, and available datasets.

3.2 Overview of the Four Pilot Case Studies

AIGOV includes four pilot case studies (PCSs). The pilots were selected to cover a broad spectrum of administrative challenges, data types, legal and ethical risks, and AI capabilities. Together, they enable a multi-dimensional evaluation of AI adoption in the public sector, ranging from predictive analytics to legal reasoning and democratic participation. The pilots also provide the empirical grounding necessary for testing the AIGOV Transformation and Adoption Framework, the Government Data Value Cycle, and the AIGOV Framework for Trustworthy, Fair, and Accountable AI.

The four pilots address various domains and are based on various state-of-the-art AI technologies:

- Predictive policy analytics using Explainable Graph Neural Networks
- Legal interpretation and domain-specific reasoning using Large Language Models
- Multilingual and multicultural democratic deliberation supported by generative AI

Below, each pilot is described in more detail.

Pilot Case Study 1 (PCS1) - Explainable Graph Neural Networks for Housing Market Analysis [17]. This pilot investigates whether Explainable Graph Neural Networks (GNNs) can support evidence-based policymaking by predicting whether housing prices in Scottish data zones lie above or below the national average. Using linked open statistical datasets, the pilot explores how spatial relationships and socio-economic indicators can be modelled more effectively than with traditional machine-learning approaches. A key objective is to ensure transparency, auditability, and explainability, critical requirements for the public sector, through techniques such as surrogate models and GNN-specific explainability tools. The pilot provides insights into how advanced predictive methods can enhance public value while remaining accountable and interpretable.

Pilot Case Study 2 (PCS2) - Evaluating Open and Proprietary LLMs on the EU VAT Directive [20]. This pilot evaluates the capacity of state-of-the-art open and proprietary Large Language Models to interpret and answer questions about the EU VAT Directive. The legal domain presents a high-precision environment in which incorrect reasoning, hallucinations, or ambiguous interpretations may carry significant risks. The pilot assesses the legal accuracy, reliability, and reasoning quality of nine LLMs across multiple families (Claude, GPT, Mistral, Llama), analysing how their outputs align with the directive's requirements. The study provides critical insights into the strengths, limitations, and safe-use guidelines for LLMs in tax administration, compliance, and regulatory contexts.

Pilot Case Study 3 (PCS3) - LLM-Enabled Legal Reasoning and Document Processing in Public Administration. The third pilot extends the analysis of LLMs beyond the VAT Directive to broader legal and administrative tasks, such as classification, summarisation, argument extraction, and cross-lingual analysis of legal texts. This pilot examines how different model families behave when applied to structured and unstructured legal corpora, identifying strengths and weaknesses in reasoning, reproducibility, fairness, and transparency. The findings illuminate the conditions under which LLMs can safely augment legal workflows, always under human oversight and help define governance requirements to ensure accuracy, accountability, and compliance with legal standards.

Pilot Case Study 4 (PCS4) - AI-Supported Multilingual and Multicultural Deliberation [16]. This conceptual pilot develops a framework and technical blueprint for AI-enabled multilingual deliberation, addressing barriers in democratic participation caused by linguistic fragmentation, cultural diversity, and unequal access to deliberative tools. The pilot explores how generative AI, including speech recognition, machine translation, summarisation, cultural-aware argument extraction, and explainability tools, can support more inclusive, accessible, and equitable democratic dialogue. Because AI in democratic contexts carries substantial ethical and legal risks, this pilot focuses heavily on governance safeguards, fairness across languages, cultural sensitivity, and mechanisms that ensure AI remains advisory and under strict human oversight.

3.3 Baseline Assessment Across Pilots

The baseline assessment provides a unified understanding of the initial conditions from which the four AIGOV pilot case studies begin, before any evaluation activities or methodological interventions are introduced. Although each pilot operates within a different policy domain and engages with different AI technologies, from graph neural networks and legal reasoning models to multilingual generative AI, the assessment reveals a set of shared structural challenges as well as pilot-specific starting points. These baseline conditions shape the feasibility of the pilots, the design of evaluation scenarios, and the definition of Key Performance Indicators.

Across all four pilots, the strategic motivations reflect a common ambition: public administrations seek to enhance analytical capacity, interpretive accuracy, and participatory inclusiveness through responsible use of AI. In the housing analysis pilot (PCS1), the baseline challenge lies in improving the quality and interpretability of evidence used in regional housing policy. In the two legal reasoning pilots (PCS2 and PCS3), the baseline is shaped by the complexity of regulatory texts and the need for models whose outputs are both precise and interpretable within contexts carrying legal obligations. In the multilingual deliberation pilot (PCS4), the starting point is a democratic environment characterised by linguistic fragmentation, cultural diversity, and uneven capacity for participation, which traditional tools and processes struggle to accommodate. Together, these strategic contexts demonstrate a growing recognition in European public administrations that AI can offer value only when it enhances human decision-making, reduces cognitive burden, and strengthens democratic or administrative legitimacy.

The stakeholder landscape at baseline also reflects important differences. PCS1 through PCS3 involve stakeholders primarily inside the public administration: data analysts, legal experts, IT specialists, and regulators who will ultimately oversee and interpret AI outputs. PCS4 expands this landscape significantly, involving citizens, civil society groups, interpreters, and democratic facilitators, given its focus on public deliberation. Across all pilots, however, the baseline reveals a need for clearer accountability structures and greater capacity for human oversight, particularly in contexts where AI-generated outputs may influence legal decisions, administrative processes, or democratic dialogue.

Data readiness varies markedly between pilots and forms one of the most consequential baseline factors. PCS1 benefits from structured, interoperable statistical data that is well suited for modelling, although spatial imbalance and missing variables represent identifiable constraints. In PCS2 and PCS3, data availability is high, but the symbolic complexity and semantic ambiguity of legal texts require careful preparation, domain expertise, and benchmarking resources. PCS4 presents the lowest level of data readiness because multilingual deliberation involves live speech, written contributions, video streams, and culturally nuanced expressions, data that is not currently standardised or supported by well-established pipelines within public administrations. This variation underscores that each pilot begins with different levels of maturity in relation to the AIGOV Data Value Cycle, particularly regarding data collection, cleaning, annotation, storage, and governance.

Legal, ethical, and governance constraints also shape the baseline conditions for each pilot. PCS1 operates in a relatively low-risk environment involving open structured data, though transparency and fairness remain essential. PCS2 and PCS3 sit at a significantly higher regulatory threshold: AI used in interpreting legal texts must meet strict accuracy, traceability, and accountability standards, as errors may have compliance or policy implications. PCS4 is situated in the most sensitive domain, as AI systems that support political participation or democratic processes may be classified as high-risk under the EU AI Act, triggering stringent requirements for documentation, human oversight, and bias mitigation. The baseline reveals that the significance of governance safeguards increases progressively from PCS1 to PCS4, mirroring the increasing societal impact of the domains involved.

Organisational and capability capacity at baseline is equally heterogeneous. PCS1 requires mostly technical skills in data engineering and modelling. PCS2 and PCS3 require interdisciplinary expertise combining legal interpretation, AI governance, and linguistic analysis, alongside the capacity to validate large model outputs. PCS4 demands the broadest set of competencies, including advanced linguistic expertise, intercultural communication, democratic facilitation, and real-time AI supervision, most of which are not yet widely available within public administrations. Across all pilots, the baseline shows that while research teams possess the necessary technical skills, the public bodies that would ultimately adopt such systems often lack mature structures for AI governance, risk management, and cross-disciplinary collaboration.

Taken together, the baseline assessment shows that each pilot begins from a different level of readiness, shaped by the nature of the domain, the sensitivity of the data, and the institutional capacity available. Despite this, all pilots share a need for clear human oversight mechanisms, responsible design principles, and consistent governance frameworks. The differences across pilots are not limitations but important inputs for designing tailored evaluation scenarios, appropriate KPIs, and realistic pathways for the adoption of trustworthy and public-value aligned AI solutions. This cross-pilot baseline therefore forms the foundation for the evaluation strategy that follows. This cross-pilot baseline therefore forms the foundation for the evaluation strategy that follows.

3.4 Evaluation Scenarios

The evaluation scenarios define the lens through which each pilot will be assessed. Rather than focusing solely on technical performance, the scenarios incorporate organisational, ethical, legal, and societal dimensions, ensuring that pilot evaluation aligns closely with the AIGOV Transformation and Adoption Framework and the principles of Trustworthy, Fair, and Accountable AI. They also introduce a coherent structure that can be applied consistently across all pilots, enabling cross-domain comparison while still preserving the specificities of each use case.

The first scenario, Technical Performance, evaluates whether the AI components function effectively within their intended operational or conceptual context. In PCS1, this involves assessing the predictive capacity and robustness of Explainable Graph Neural Networks on linked Scottish housing datasets, including comparisons with classical machine-learning baselines. PCS2 focuses on the legal accuracy and reasoning quality of LLM responses to questions derived from the EU VAT Directive, where even minor interpretive errors may have significant regulatory implications. PCS3 extends technical performance assessment to classification, summarisation, argument extraction, and multi-lingual legal analysis tasks performed by LLMs over large legal corpora. In PCS4, the performance scenario remains conceptual but examines the anticipated capabilities of AI components such as speech recognition, machine translation, summarisation, and argument mining within multilingual deliberation environments. Across all pilots, this scenario ensures that each proposed AI system demonstrates the baseline functional adequacy required before questions of governance or public value can be addressed.

The second scenario, Explainability, Transparency, and Traceability, responds to the foundational public-sector requirement that AI systems must produce interpretable and auditable outputs. In PCS1, explainability is achieved through surrogate models and GNN-specific tools that reveal feature importance and spatial relationships, ensuring that predictions about housing prices can be scrutinised and justified by policymakers. PCS2 evaluates whether LLMs provide sufficiently transparent reasoning to support legal interpretation, assessing the extent to which responses can be traced back to the VAT Directive. PCS3 examines the reproducibility and interpretability of legal reasoning outputs, with particular attention to the risk of inconsistent or opaque LLM justifications. PCS4 emphasises culturally sensitive transparency, requiring that multilingual translations, condensations, and argument extraction processes be explainable and traceable to original contributions. This scenario operationalises the Trustworthy AI principle that public-sector AI must enable users to understand, challenge, and contextualise model outputs.

The third scenario, Ethical, Legal, and Risk Compliance, evaluates alignment with GDPR, the EU AI Act, and broader ethical guidelines for public-sector AI use. PCS1 must consider fairness and potential geographic or socio-economic bias in predictive modelling. PCS2 and PCS3, which operate in legal and regulatory domains, require heightened safeguards due to the sensitivity of legal reasoning, the risk of hallucinations, and the potential consequences of erroneous outputs. PCS4 lies closest to the high-risk classification foreseen under the AI Act because it concerns democratic processes; thus, ethical and legal compliance is paramount, including data protection for speech and text contributions, cultural fairness, and the prevention of AI-mediated distortions of public discourse. This scenario ensures that pilot outputs do not inadvertently harm individuals or undermine trust in public institutions.

The fourth scenario, Public-Sector Value, Usability, and Adoption, assesses whether the AI systems deliver meaningful value to public administrations and their stakeholders. PCS1 examines the extent to which explainable GNNs improve analytical capacity and support more evidence-based housing policy. PCS2 and PCS3 evaluate whether LLMs can reliably augment legal workflows without replacing human judgment, increase efficiency, or improve comprehension of complex regulatory texts. PCS4 considers whether AI-supported translation and argumentation tools can enhance inclusiveness, reduce cognitive burdens, and strengthen shared understanding in multilingual deliberations. This scenario also evaluates human oversight mechanisms, accessibility, and trustworthiness—factors critical to the integration of AI tools within real administrative environments.

Together, these evaluation scenarios provide a holistic and balanced basis for assessing the four pilots. They ensure that technical rigour is complemented by considerations of public value, ethical integrity, organisational feasibility, and legal compliance. By applying these scenarios consistently across pilots, the evaluation strategy strengthens comparability while respecting domain-specific differences, ultimately supporting a coherent understanding of how AI systems can be responsibly adopted across the European public sector.

3.5 Integration with AIGOV Frameworks

Each pilot is explicitly mapped against:

(1) The AIGOV Transformation & Adoption Framework, and specifically, Phases 1-3 of the framework.
(2) The guidelines of the AIGOV Government Data Value Cycle (Table 1)
(3) The pillars and guidelines of Framework for Trustworthy, Fair, and Accountable AI (Table 2)

Table 1 Relation of the pilot case's with the WP2's AIGOV Government Data Value Cycle

Step	*Guideline*	PCS1	PCS2	PCS3	PCS4
Data collection	G1.1 Use standards to structure government data used in AI.	**Strong**	Moderate	Moderate	Moderate
	G1.2 Use Application Programming Interfaces (APIs) to provide access to government data including dynamic data.	**Strong**	-	-	Conceptual
	G1.3 Implement dynamic data pipelines in order to enable continuous data ingestion and processing.	Moderate	-	-	Conceptual
Data curation	G2.1 Establish a data-quality management framework. Define measurable quality indicators (accuracy, completeness, timeliness, consistency) and monitor them periodically.	Moderate	Moderate	**Strong**	Conceptual
	G2.2 Combine automation with human-in-the-loop validation. Use automated cleaning and anomaly-detection tools while maintaining expert supervision for critical datasets.	Moderate	Moderate	Strong	Strong (Conceptual)
	G2.3 Apply data-imputation and augmentation techniques. Address incomplete or small datasets using statistical imputation, synthetic data generation, or controlled	Moderate	-	Moderate	Conceptual

	data augmentation for textual corpora.				
	G2.4 Implement verifiable and repeatable curation workflows. Maintain logs and metadata describing each modification, ensuring reproducibility and auditability.	**Strong**	Moderate	**Strong**	Conceptual
	G2.5 Promote ethical and unbiased data practices. Integrate fairness checks, bias detection algorithms, and human review throughout the curation pipeline.	Moderate	**Strong**	**Strong**	**Strong**
	G2.6 Invest in data-literacy and capacity-building. Train data stewards, AI engineers, and public servants to understand data provenance, bias, and quality assurance processes.	Conceptual	Conceptual	Moderate	**Strong**
	G2.7 Leverage scalable, interoperable tools. Employ machine-learning-based curation, crowdsourcing, and collaborative platforms consistent with European interoperability standards.	**Strong**	Moderate	Moderate	Moderate
Data Integration & Linking	G3.1 Adopt semantic-interoperability standards.	**Strong**	-	Moderate	Moderate
	G3.2 Use Linked Data and API-based architectures.	**Strong**	-	-	Conceptual
	G3.3 Establish data-governance agreements.	Moderate	**Strong**	**Strong**	**Strong**

	G3.4 Ensure privacy-preserving linkage.	Moderate	Moderate	**Strong**	**Strong**
	G3.5 Foster trusted partnerships.	Moderate	Moderate	Moderate	**Strong**
	G3.6 Monitor integration performance.	Moderate	Moderate	Moderate	Moderate
Data Storing	G4.1 Prioritize privacy, security, and compliance.	**Strong**	**Strong**	**Strong**	**Strong**
	G4.2 Implement efficient data discovery mechanisms.	Moderate	Moderate	**Strong**	Moderate
	G4.3 Track and document data provenance.	**Strong**	Moderate	**Strong**	Moderate
	G4.4 Strengthen data governance frameworks.	Moderate	**Strong**	**Strong**	**Strong**
	G4.5 Integrate advanced storage solutions for AI.	Moderate	Moderate	**Strong**	Moderate
	G4.6 Ensure interoperability and sustainability.	**Strong**	Moderate	Moderate	Moderate
Data Dissemination	G5.1 Apply open data principles systematically.	**Strong**	-	-	Conceptual
	G5.2 Use standardized metadata and identifiers.	**Strong**	Moderate	**Strong**	Moderate
	G5.3 Provide machine-readable, multilingual, and API-enabled access.	**Strong**	Moderate	Moderate	**Strong**
	G5.4 Establish data release workflows with quality assurance.	Moderate	Moderate	**Strong**	Conceptual
	G5.5 Ensure security and ethical oversight.	Moderate	**Strong**	**Strong**	**Strong**
	G5.6 Promote awareness and capacity building	Conceptual	Conceptual	Moderate	**Strong**

	G5.7 Encourage feedback and collaboration.	Moderate	Moderate	Moderate	**Strong**
Data Usage	G6.1 Obtain and sustain social licence.	Moderate	**Strong**	**Strong**	**Strong**
	G6.2 Pursue purposeful and proportionate analytics.	**Strong**	**Strong**	**Strong**	**Strong**
	G6.3 Embed ethics and accountability frameworks.	**Strong**	**Strong**	**Strong**	**Strong**
	G6.4 Promote human oversight.	**Strong**	**Strong**	**Strong**	**Strong**
	G6.5 Build data literacy and analytical capacity.	Moderate	Moderate	Moderate	**Strong**
	G6.6 Communicate results transparently.	**Strong**	**Strong**	**Strong**	**Strong**
	G6.7 Encourage cross-sector collaboration.	Moderate	**Strong**	**Strong**	**Strong**
Data Value Creation	G7.1 Assess and evolve data architecture.	**Strong**	Moderate	**Strong**	**Strong**
	G7.2 Identify and prioritize use cases strategically	**Strong**	**Strong**	**Strong**	**Strong**
	G7.3 Deploy and fine-tune LLMs for public-sector needs.	-	**Strong**	**Strong**	**Strong**
	G7.4 Support multilingualism and inclusivity.	-	Moderate	Moderate	**Strong**
	G7.5 Facilitate LLM integration through frameworks.	-	**Strong**	**Strong**	**Strong**
	G7.6 Maintain human oversight and accountability.	**Strong**	**Strong**	**Strong**	**Strong**

	G7.7 Establish comprehensive risk management.	Moderate	**Strong**	**Strong**	**Strong**
	G7.8 Build organizational skills and roles.	Moderate	Moderate	Moderate	**Strong**
	G7.9 Develop applications jointly with end users.	Moderate	Moderate	Moderate	**Strong**
	G7.10 Communicate transparently.	**Strong**	**Strong**	**Strong**	**Strong**
	G7.11 Start small and scale responsibly.	**Strong**	**Strong**	**Strong**	**Strong**

The four pilots collectively span the entire Government Data Value Cycle, but each aligns with different steps according to the nature of the domain. PCS1 exemplifies a high-maturity use of open government data, semantic standards, and explainable predictive modelling. It best operationalises the early stages of the cycle. PCS2 deals with high-risk legal interpretations of VAT, requiring strict oversight, transparency, and bias monitoring. It demonstrates how LLMs can be responsibly evaluated for regulatory domains. PCS3 focuses on legal document processing and reasoning, emphasizing reproducibility, data accuracy, and fairness. It contributes significantly to governance and accountability mechanisms for AI in legal workflows. Finally, PCS4 aligns most strongly with principles of fairness, inclusiveness, human oversight, and risk management, given the sensitivity of democratic deliberation contexts.

Table 2 Relation of the pilot case's with the WP2's AIGOV Framework for trustworthy, fair and accountable AI

Pillar	*Guideline*	PCS1	PCS2	PCS3	PCS4
Pillar 1: Transparency and accountability	G1. AI models should be explainable.	**Strong**	**Strong**	**Strong**	**Strong**
	G2. AI models and systems, their architecture, data sources, and methodologies used during their creation, training, or fine-tuning should be open to all stakeholders.	**Strong**	Partial	Partial	**Strong**
	G3. Data used in AI should be accurate and timely.	**Strong**	Partial	**Strong**	Partial

Pillar 2: Responsible Data Management and Access	G4. Data used in AI should be provided in various levels of granularity.	**Strong**	Partial	Partial	Partial
	G5. Access to dynamic data (e.g., sensor data) should be enabled through an Application Programming Interface (API).	**Strong**	Partial	Partial	Partial
Pillar 3: Comprehensibility and Multilingual Support	G6. Responses of (Generative) AI should be comprehensible to all stakeholders speaking multiple languages and with different cultural and educational backgrounds.	Partial	Partial	Partial	**Strong**
	G7. AI tools should support retrieval of information from large document collections to enable evidence-based decision-making.	Partial	**Strong**	**Strong**	**Strong**
Pillar 4: Data Interoperability and Reusability	G8. AI models, and especially LLMs, should be built on interoperable datasets, enabling seamless integration, sharing, and reuse across various applications and systems.	**Strong**	Partial	**Strong**	**Strong**

Across all four pilots, the AIGOV pillars are reflected to varying degrees. PCS1 aligns strongly with transparency and data interoperability through its use of explainable GNNs and Linked Data pipelines. PCS2 and PCS3, focusing on legal reasoning, demonstrate high alignment with transparency, accountability, and responsible data use, while highlighting the challenges of proprietary LLM architectures. PCS4 shows the strongest alignment with multilingual support and inclusiveness, addressing complex ethical, cultural, and governance requirements inherent in democratic deliberation contexts. Together, the pilots show complementary coverage of the AIGOV framework, demonstrating its applicability across technical, legal, administrative, and democratic domains.

4 Pilot Operation and Evaluation

4.1 Pilot Case study 1: Explainable Graph Neural Networks: An Application to Open Statistics Knowledge Graphs for Estimating House Prices

4.1.1 Phase 1: Assess Readiness and Context

4.1.1.1 Strategic Context and Public Value

Buying a house is probably one of the major decisions and financial commitments in the life of people. This decision is usually affected by prices that fluctuate due to many factors, such as changes in interest rates, economic growth, government policy, and supply and demand dynamics [5]. Changes in property prices not only reflect broader economic trends but also impact the socio-economic fabric of societies, with far-reaching consequences for home ownership and wealth distribution [1][12]. Between 2010 and 2021, the European Union experienced a surge in house prices by 37%, rents by 16%, and inflation by 17%[1]. Concurrently, the cost of construction for new residences soared by 25%, particularly since 2016. The phenomenon has been exacerbated by recent global crisis events, such as the COVID-19 pandemic and war conflict in Ukraine, resulting in unprecedented inflation levels [21] that have dramatically altered the landscape of housing prices and rent [9][32].

In such a complex environment, a realistic estimation of house prices becomes not just a theoretical exercise but an essential tool for governments, policymakers, and the private sector to form strategies and policies that reflect the ever-changing dynamics of the housing market. Furthermore, predicting housing prices is not only crucial for individual decision making but also for broader economic planning and policy formulation. Understanding the dynamic relationships between housing prices and socio-economic variables can inform investment decisions, fiscal policies, and urban planning initiatives [29]. This study highlights the functional relationships between residential property prices and socio-economic factors like the number of loans, unemployment levels, and market rent, showcasing the predictive power of such models over a substantial period.

In the era of accelerating digitization and advanced big data analytics, due to the growing availability of data, machine learning models have become efficient approaches for predicting house prices [35]. However, traditionally, these approaches rely heavily on socio-economic indicators to estimate house prices, which might not always yield accurate results. On the contrary, making predictions based on features with spatial context, like the presence of schools nearby, the availability of parking facilities or gas stations in the same or nearby neighborhood, and the proximity to public transport, can significantly improve the prediction

[1] https://ec.europa.eu/eurostat/en/web/products-interactive-publications/-/ks-09-21-479

accuracy [35] [48] since they reflect the realistic representation of geographical regions and the connectivity between them. This results from the fact that neighboring regions may influence the house prices of each other due to their proximity and shared features. For example, the presence of schools or hospitals in a region that has high house prices might similarly affect house prices of the nearby regions. Recently, advancements in deep learning have demonstrated the superiority of approaches based on neural networks over both traditional statistical methods and conventional machine learning approaches, especially when working with geospatial data [21]. In the past few years, algorithms from graph machine learning have emerged as efficient candidates for predicting node or edge features from graph-structured data [5]. Specifically, Graph Neural Networks (GNNs) constitute a family of algorithms uniquely tailored for scenarios where the spatial representation of data must be explicitly modeled. These models have been successfully applied in various fields, such as social network analysis [44], traffic prediction [13], and recommendation systems [47], to model spatial relationships and dependencies effectively. Inspired by the widespread use and success of GNNs in these applications, we aim to apply them to house price prediction. Spatial dependencies in house prices have not been adequately modeled in traditional approaches, yet they are crucial for accurate predictions. By leveraging GNNs, we can better capture the spatial interactions and dependencies inherent in housing markets, leading to more precise and reliable predictions. Explainable artificial intelligence can be applied on top of the GNN in order to understand the decisions made by the model, improving its explainability and interpretability.

At the same time, linked data technologies have been used to facilitate the integration of open statistical data on the web. Linked statistical data are highly structured data describing statistical indicators based on a set of dimensions, including a geographical dimension and temporal dimension, and attributes [14]. As a result, linked statistical data formulate open statistics Knowledge Graphs (KGs). The structure of open statistics KGs, which already captures spatial relationships between geographical dimensions, makes them suitable for creating GNNs. Currently, open statistics KGs can be accessed as linked data by many official government data portals (e.g., the official portal for European data https://data.europa.eu/, accessed on 30 March 2024) allowing for their integration and the creation of valuable applications for governments, policy makers, enterprises, and citizens. For example, linked statistical data have been previously used to help house owners, buyers, and investors understand which factors affect and determine the prices of houses in the Scottish data zones [18].

This pilot case study therefore explores the use of Explainable GNNs applied to open statistical data to estimate house prices in Scotland. The focus is on assessing whether advanced AI models can be used responsibly in public-sector contexts where transparency, accountability, and explainability are essential. Toward this end, the pilot case study leverages (i) three GNN variants, namely the Chebyshev Neural Network (ChebNet), Graph Convolutional Network

(GCN), and GraphSAGE, and (ii) a KG modeled as linked statistical data from the Scottish data portal, to predict the probability that the average house prices across Scotland's "2011 data zones" are above Scotland's total average. In order to understand the decisions of the best performed model, both global and local explainability are employed using a global surrogate model and the GNNExplainer framework, respectively.

The objective of this use case is not only to demonstrate predictive performance, but to assess whether advanced AI models such as GNNs can be responsibly applied in a public administration context, meeting requirements for transparency, accountability, explainability, and public benefit.

4.1.1.2 Stakeholders

The stakeholders involved in this pilot case study represent the groups that would interact with or be affected by the development and potential deployment of explainable graph-based machine learning models for housing price prediction in a public administration context. As in Case Study 1, these stakeholders are classified according to the internal and external stakeholder categories defined in D1.2 of the AIGOV Ecosystem.

Internal stakeholders include:

- *Public authorities and public organisations* working in policy areas such as housing, regional development, taxation, social welfare, and urban planning. These actors would be responsible for incorporating model outputs into administrative processes or policymaking.
- *Public servants* working as data analysts, policy advisors, or planning officers could use such predictive tools to complement traditional statistical methods when analysing housing trends or designing interventions.
- *Regulators and policymakers* represent a distinct internal group, as they would need to ensure that predictive models comply with legal frameworks (e.g., GDPR and AI Act) and that the use of AI in sensitive domains such as housing does not generate discriminatory or opaque outcomes.
- *IT stakeholders*, including public-sector data engineers, AI specialists, and developers, would be responsible for integrating open data, maintaining data pipelines, training and evaluating models, and ensuring technical robustness. Public service designers may also become relevant stakeholders when such models are embedded into operational workflows, digital dashboards, or decision support tools intended for non-technical users.

External stakeholders include:

- *Citizens* and residents, who are affected by housing market dynamics and may indirectly benefit from more transparent and evidence-based policymaking.

- *Academic and scientific communities,* which play an important role in advancing methods, validating model performance, and identifying risks and ethical implications.
- *Third parties* such as real estate analysts, civic technology organisations, and independent data science practitioners may engage with such a system by auditing results, developing complementary services, or using the datasets for their own analyses.
- *Businesses* and housing-sector organisations, including real estate companies, mortgage providers, construction firms, and property investment services, may use insights produced by such models to support trend forecasting or strategic decision-making.
- *Customer advocacy groups* and civil society organisations concerned with housing affordability and access may also monitor the fairness and transparency of AI-driven insights used in public administration.

Since this case study is experimental and research-based rather than deployed in an operational environment, stakeholder interactions are conceptual rather than participatory. However, the stakeholder mapping reflects realistic usage conditions and supports alignment with the AIGOV principles of transparency, accountability, explainability, and responsible data reuse.

4.1.1.3 Data readiness

The data used in this pilot case study originate entirely from openly accessible statistical datasets published on the official Scottish Open Government Data (OGD) portal. All datasets were retrieved as linked data, following semantic web standards, which enables interoperability, reuse, and machine-readable access, attributes highlighted as essential in Deliverables D1.1 and D1.2 for enabling responsible and scalable AI development in the public sector.

The data were retrieved from sixteen (16) datasets of the OGD portal classified into seven categories, including health and social care, housing, and crime and justice, resulting in sixty (60) statistical indicators. The majority of the indicators are “Crime and Justice” data (22 indicators or 37.2%), followed by “Housing” indicators (13 indicators or 22%). Excluding the Comparative Illness Factor (CIF) and urban rural classification, which are integer and categorical variables, respectively, the rest of the indicators are numeric.

A total of 6976 observations were extracted using the SPARQL queries. Each observation refers to a Scottish “2011 data zone” accompanied by its associated statistical indicators. This quantity of data zones represents an 86.2% coverage of the entire collection of “2011 data zones” within Scotland. The reason for this disparity is that certain data zones that lack values for one or more statistical indicators are not included. The main year of reference is 2015, while, for indicators pertaining to two- or three-year spans, we have chosen 2014–2015 and

2014–2016 as the designated reference periods. A small part of the data (1.4%) are null values. Comprehensive descriptive analysis of these indicators can be found in our previous work in [19].

The problem's dependent variable is the mean house price, which ranges from GBP 20,604 (Cumbernauld Central, Glasgow) to GBP 1,244,910 (Leith (Albert Street)-03 in the city of Edinburgh) across all data zones. The average cost of a house across all data zones is GBP 163,478; the mean price of houses is higher in 39% of the "2011 Data Zones". Therefore, determining whether a Scottish "2011 Data Zone" average house price is (a) over or (b) under GBP 163,478 is the classification problem that this study addresses. Finally, there is a small imbalance in the data.

4.1.1.4 Governance, Ethics, and Legal Compliance

The deployment of artificial intelligence in the public sector requires alignment with legal frameworks, ethical standards, and transparent governance mechanisms. This section assesses the governance and compliance considerations of the pilot in accordance with the AIGOV Framework and relevant European regulatory instruments.

Since this pilot case study focuses on open government statistical data without personal or identifiable information, it operates in a low legal-risk environment under the EU Artificial Intelligence Act and does not trigger obligations associated with personal data processing under the General Data Protection Regulation (GDPR). The model is therefore classified conceptually as a low-risk AI system, although this classification would require reassessment if the model were integrated into decision-making workflows with material effects on individuals (e.g., resource allocation, taxation, eligibility, subsidies).

From an ethical standpoint, the pilot addresses three core requirements of trustworthy and responsible AI: transparency, explainability, and accountability. The use of Graph Neural Networks introduces inherent complexity in interpretability; however, model explainability is strengthened through the application of XAI methods including a global surrogate model and the GNNExplainer framework. These measures support auditability, traceability, and diagnostic insight, aligning with Pillar 1 of the AIGOV Framework (Transparency and Accountability).

Human oversight remains a foundational ethical requirement. In this pilot, the model functions as a decision-support mechanism, not an automated decision-making system. Any future operational deployment would require maintaining human-in-the-loop controls, establishing clear accountability roles (model developers, domain experts, and responsible decision-makers), and defining escalation pathways in case of anomalous outputs or model drift.

Finally, transparency obligations extend beyond technical explainability. Public communication of methodology, data provenance, model behavior, and limitations is essential to sustain public trust, particularly in sensitive socioeconomic domains such as housing. In a real deployment setting, public administrations would need to ensure that explanations are not only technically defensible but also understandable to non-technical users, aligning with the public value principles of the AIGOV Transformation and Adoption Framework.

4.1.1.5 Organisational and Capability Capacity

In the context of this pilot case study, the assessment of organisational and capability capacity focused on determining whether the required expertise, infrastructure, and supporting conditions existed to carry out the research activities and explore the feasibility of explainable AI methods in a public-sector domain.

The necessary competencies for this case study include:

- Technical AI and machine learning expertise, particularly in Graph Neural Networks (ChebNet, GCN, GraphSAGE) and benchmark modelling.
- Semantic technologies and data integration capabilities, required to query, structure, and transform datasets from the Scottish Open Government Data portal using Linked Data standards and SPARQL.
- Explainable AI (XAI) expertise, including use of GNNExplainer and surrogate modelling (logistic regression) to interpret model decisions.
- Infrastructure access, including computational resources suitable for training and evaluating deep learning models.

Leadership support, organisational change management structures, end-user training, and operational skills assessment were not applicable at this stage, as the case study was not embedded in a live administrative context. However, the experiment provides insight into what would be necessary for future adoption, including:

- Increased data literacy and analytical capacity among policy staff;
- A foundational level of AI governance and oversight capability;
- Technical support roles for data engineers, AI maintainers, and domain interpreters;
- Infrastructure readiness, including secure and scalable environments compliant with public-sector standards.

Thus, while the organisational capacity for this case study was sufficient for research prototyping and evaluation, broader capability development would be required for sustained real-world public administration use. The pilot therefore acts as an early-stage input to future capability planning rather than an assessment of existing public-sector readiness.

4.1.2 Phase 2: Design Ethical and Value-Aligned Use Cases

Phase 2 focuses on translating the contextual analysis and readiness assessment from Phase 1 into a well-defined, responsible, and evaluable AI use case. For this pilot, the goal is not only to test predictive modelling feasibility, but also to ensure alignment with public value, governance safeguards, and responsible deployment principles. The activities in this phase follow the AIGOV Transformation and Adoption framework and include:

(1) problem and service redesign definition (Section 4.1.2.1),

(2) responsible AI use case design (Section 4.1.2.2), and

(3) definition of success metrics and evaluation criteria (Section 4.1.2.3).

The results of this phase serve as the reference framework for the implementation and evaluation activities carried out in Phase 3 (Build and Test), ensuring traceability between design intent and technical execution.

4.1.2.1 Problem and service redesign definition

Housing affordability and real estate market volatility present an ongoing challenge for governments, policy makers, and citizens. Traditional statistical tools used in public administrations often provide delayed, aggregated, or limited insights, making it difficult to anticipate housing trends or understand the underlying socio-economic factors influencing price fluctuations. As identified in Phase 1, existing analytical practices rely largely on tabular indicators and standard econometric models, which may overlook important spatial dependencies that shape market behaviour. As a result, public authorities lack timely and granular evidence to support forward-looking housing policy, budget planning, regional development strategies, or targeted interventions.

In this context, the service challenge is reframed as the need to enhance the analytical capacity of public administrations by incorporating advanced AI-based models capable of capturing geographical relationships, complex indicator interactions, and evolving market dynamics. Rather than replacing existing analytical workflows, the redesigned service concept positions AI as a complementary decision-support layer that can generate higher-resolution insights from openly available linked statistical datasets.

User needs identified during the assessment phase indicate that any analytical system intended for policy use must remain interpretable, transparent, and aligned with legal and ethical expectations. Public servants require not only predictive outputs, but also meaningful explanations that help them understand which variables influence pricing trends and how regional characteristics interact. Policy makers and regulatory bodies further require safeguards to ensure that advanced analytics do not introduce opaque, discriminatory, or biased patterns, particularly in a domain that affects social equity, access to housing, and economic mobility.

Accordingly, the redesigned service concept introduces an Explainable AI solution based on GNNs to generate predictions on whether house prices in specific geographic zones are above or below the national average. These predictions are accompanied by interpretable justifications through explainability methods such as surrogate modelling and GNNExplainer. This allows the analytical process to remain traceable, enabling human experts to validate, challenge, or contextualise model output before it informs decision processes.

The redesign therefore shifts the service from a static descriptive analytics model to an advanced, explainable, and spatially aware predictive system that supports proactive evidence-based decision-making. This approach maintains the role of human judgment while introducing new analytical capabilities aligned with responsible AI adoption in the public sector.

Based on the problem definition and the insights identified during Phase 1, the need for an advanced, transparent, and spatially aware analytical capability was translated into a concrete responsible AI use case: the development and evaluation of GNNs to support housing-market analysis in the public sector. The purpose of the use case is not solely predictive performance but the demonstration of how complex machine learning models can be applied in a manner aligned with public values, legal frameworks, and ethical requirements.

4.1.2.2 Responsible AI Use Case Design

In accordance with the AIGOV Framework for Trustworthy, Fair, and Accountable AI, several design principles guided the development of this use case:

1. Explainability Requirements. Since GNNs inherently operate as complex deep learning models, explainability was established as a core requirement from the outset. The design therefore incorporates both global and local interpretability techniques, including surrogate models (logistic regression) and GNNExplainer, to enable public-sector analysts to understand which factors drive predicted outcomes and how spatial relationships influence results. This ensures that analytical insight, rather than black-box behavior, remains central to the use case.
2. Ethical and Legal Safeguards. Even though the dataset used does not involve personal or sensitive information, ethical safeguards remain essential due to the socio-economic relevance of housing prediction. Safeguards were embedded to prevent misuse of the model for automated decision-making affecting citizens, avoid algorithmic bias in regional classification, and ensure transparency about data sources, assumptions, and model limitations. If deployed operationally, the model would require reassessment under the EU AI Act, and governance structures such as auditability, role allocation, and escalation procedures would be introduced.
3. Human Oversight and Accountability. The system is designed as a human-in-the-loop model, where predictions serve as decision-support inputs rather than automated determinations. Policy analysts, planners, and domain experts retain responsibility for interpreting, validating, and contextualising outputs. This aligns with requirements in

the AIGOV Framework ensuring responsibility remains with accountable public actors, not algorithmic systems.

4. Inclusiveness and Accessibility. While this use case does not engage end users directly, accessibility principles were incorporated by ensuring that results and explanations are interpretable by non-technical audiences, that outputs could be visualised on dashboards or integrated with standard reporting processes, and that model documentation is written in clear language to support transparency and reproducibility.
5. Multilingual considerations may become relevant in future scaling scenarios, particularly in multinational government contexts.

Together, these design elements ensure that the AI use case adheres to responsible innovation principles, allowing the system to be assessed not only based on technical accuracy but also on its alignment with public values, legal compliance, governance requirements, and institutional readiness. The result is a responsible and realistically adoptable AI use case capable of supporting data-driven public service improvement while maintaining trustworthiness and accountability.

4.1.2.3 Success Metrics and Impact Definition

To operationalise the evaluation objectives defined in Phase 2, a structured set of Key Performance Indicators (KPIs) was established. These KPIs serve as measurable criteria for assessing both the technical performance of the model and its suitability for responsible public-sector AI deployment. Rather than focusing solely on machine-learning accuracy, the indicators reflect the multidimensional evaluation approach required under the AIGOV Frameworks capturing transparency, explainability, reproducibility, fairness, and alignment with public value.

The KPIs are grouped into three complementary evaluation dimensions:

- Technical Performance (Table 3), assessing predictive accuracy, robustness, and comparative performance against established modelling approaches;
- Explainability and Transparency (Table 4), examining whether the model's reasoning can be interpreted, audited, and communicated to non-technical stakeholders; and
- Responsible AI and Public-Sector Applicability (Table 5), evaluating ethical safeguards, alignment with governance requirements, fairness considerations, and readiness for operational use in a public-sector ecosystem.

Not all KPIs can be fully quantified in the context of this research-oriented prototype. Accordingly, the tables distinguish between KPIs assessed in this pilot and those defined as forward-looking metrics for future implementation, monitoring, and refinement. Together, these KPIs provide a foundation for iterative evaluation across subsequent phases of the AIGOV Transformation and Adoption Framework, supporting evidence-based decision-making on whether, and how, such a model should be integrated into real-world public administration environments.

KPI	Definition	Status
Accuracy	Percentage of correctly classified data zones	Measured
Precision	Share of true positive predictions among all positive outputs	Measured
Recall (Sensitivity)	Share of true positives among actual positive cases	Measured
F1-Score	Harmonic mean of precision and recall (supports imbalanced datasets)	Measured
Comparative Performance Score	Improvement relative to benchmark models (XGBoost, MLP)	Measured
Model Robustness / Stability	Performance variation across repeated runs or perturbed input data	Partially assessed

Table 3 Technical Performance KPIs for PCS1

KPI	Definition	Status
Explainability Score	Ability to interpret decisions using GNNExplainer and surrogate models	Partially assessed
Traceability Score	Degree to which outputs can be linked to specific features or graph structure	Partially assessed
User Interpretability Index	Ability for non-technical stakeholders to understand model behaviour	Defined (future validation)
Documentation Completeness Indicator	Availability and clarity of data provenance, modelling steps, assumptions, and limitations	Partially assessed

Table 4 Explainability and Transparency KPIs for PCS1

KPI	Definition	Status
Decision-Support Suitability Indicator	Suitability for advisory—not automated—decision-making	Defined (future validation)
Fairness and Bias Assessment Score	Degree of bias across socio-spatial categories (e.g., rural vs. urban)	Defined (future validation)
Reproducibility and Openness Index	Ability of third parties to reproduce the model using available resources	Partially assessed
Public Value Impact Potential	Expected benefit to planning, transparency, and policy development	Defined
Risk Assessment Rating	Identification and severity of potential harms or misinterpretations	Defined

Governance and Human Oversight Readiness	Presence of oversight, accountability, and escalation mechanisms	Defined (future phase)

Table 5 Responsible AI and Public-Sector Applicability KPIs for PCS1

4.1.3 Phase 3: Build and Test

Phase 3 focuses on developing, implementing, and validating the AI solution in a controlled environment. In the context of this pilot case study, this phase consisted of constructing the Explainable Graph Neural Network (GNN) models, preparing the linked statistical datasets, integrating explainability mechanisms, and evaluating the system against the KPIs established during Phase 2.

4.1.3.1 Prototype Development and Technical Implementation

The Build and Test phase operationalised the use case through four technical steps (Figure 2) (data collection, pre-processing, model development, and explainability analysis), which align with the AIGOV Phase 3 activities. Steps 1-3 correspond to prototype construction, while Step 4 aligns with ethical and explainability validation. User testing and feedback integration remain a future activity scheduled for subsequent deployment phases.

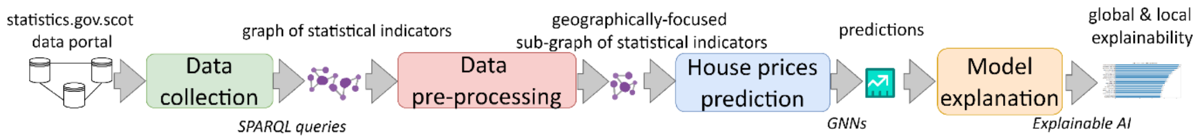
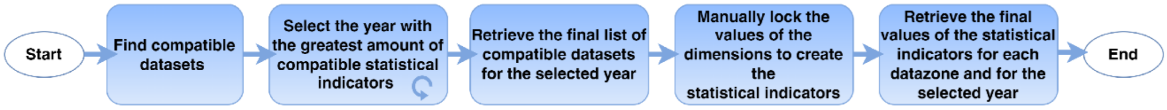


Figure 2 Methodology for the build and test phase of PCS1 [17].

This activity includes the following steps:

(1) Collect data.

Method (Figure 3)*:* This case study utilizes linked statistical data from the Scottish Open Government Data portal. Toward this end, multiple SPARQL queries were submitted to the SPARQL endpoint provided by the data portal. Specifically, the first query was applied to find all compatible datasets in the Scottish data portal (i.e., with the same year of reference and granularity level of the geography dimension) that measure ratio, percent, or score, resulting in 30 datasets. Then, using various years, we repeatedly submitted a second SPARQL query to determine which year had the greatest amount of compatible variables. Year 2015 was the outcome. The third SPARQL query was then submitted to retrieve the final list of datasets that will be used to obtain the statistical indicators. This query searches for datasets that measure ratio, percent, mean, or score (rank) values regarding 2011 data zones and for years 2015, 2014-2015, or 2014-2016. It resulted in 16 datasets. The statistical indicators were then retrieved by manually locking the values of the dimensions of the datasets, resulting in 60 indicators. Finally, an SPARQL query was submitted to retrieve the final values of the 60 indicators for each data zone and for the selected year. A detailed presentation of the method



used to retrieve data and the SPARQL queries, as well as descriptive statistics of the dataset, can be found in [19].

Figure 3 The flowchart for the data collection step [17].

Results: The data were retrieved from sixteen (16) datasets of the OGD portal classified into seven categories, including health and social care, housing, and crime and justice, resulting in sixty (60) statistical indicators. The majority of the indicators are “Crime and Justice” data (22 indicators or 37.2%), followed by “Housing” indicators (13 indicators or 22%). Excluding the Comparative Illness Factor (CIF) and urban rural classification, which are integer and categorical variables, respectively, the rest of the indicators are numeric.

A total of 6976 observations were extracted using the SPARQL queries. Each observation refers to a Scottish “2011 data zone” accompanied by its associated statistical indicators. This quantity of data zones represents an 86.2% coverage of the entire collection of “2011 data zones” within Scotland. The reason for this disparity is that certain data zones that lack values for one or more statistical indicators are not included. The main year of reference is 2015, while, for indicators pertaining to two- or three-year spans, we have chosen 2014–2015 and 2014–2016 as the designated reference periods. A small part of the data (1.4%) are null values. Comprehensive descriptive analysis of these indicators can be found in our previous work in [131].

The problem’s dependent variable is the mean house price, which ranges from GBP 20,604 (Cumbernauld Central, Glasgow) to GBP 1,244,910 (Leith (Albert Street)–03 in the city of Edinburgh) across all data zones. The average cost of a house across all data zones is GBP 163,478; the mean price of houses is higher in 39% of the “2011 data zones”. Therefore, determining whether a Scottish “2011 Data Zone”’s average house price is (a) over or (b) under GBP 163,478 is the classification problem that this study addresses. Finally, there is a small imbalance in the data.

(2) Pre-process data.

Method (Figure 4)*:* This step transforms the integrated statistical indicators into a geo-centric knowledge graph that is suitable for being used by GNN algorithms to predict the house prices in Scottish “2011 data zones”. Toward this end, data zone records with null values in every feature are initially removed. The remaining data are then formatted in a way that is centered on the “2011 data zones”. To achieve this, a geographically focused sub-graph of the original linked dataset was created. The Scottish “2011 data zone” are the central nodes of the sub-graph, each of them connected to the values of the associated features. An illustration of this transformation is shown in Figure 5.

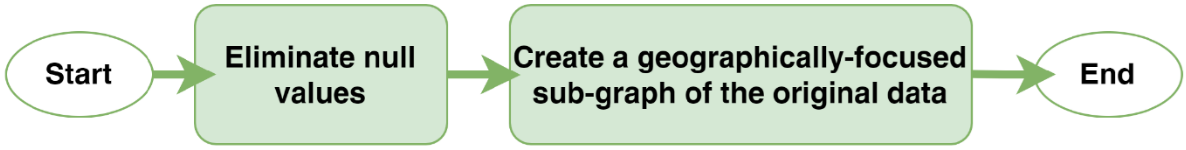


Figure 4 The flowchart for the data pre-possessing step [17].

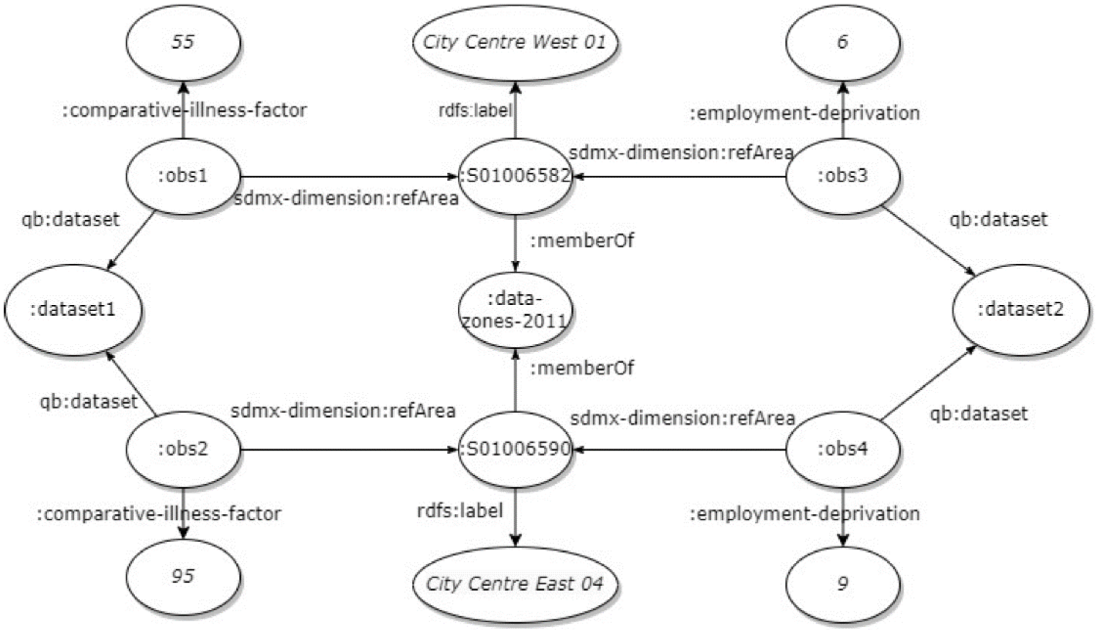


Figure 5 A graph presenting linked statistical indicators from the Scottish data portal. A fragment of the graph will be used to construct the GNNs [17].

Results: The Scottish statistical indicators were retrieved as linked data. The original dataset included 782 data zone records with null values in every feature. These records were removed, resulting in 6014 data zone records.

To facilitate the development of graph neural networks for house price prediction, a sub-graph was extracted from the retrieved dataset that centers on geographical attributes, particularly emphasizing the Scottish "2011 data zones".

In Figure 5, an example of the initially retrieved graph with the Scottish statistical data is presented. The graph comprises four observations derived from two datasets (two observations per dataset). All observations pertain to 2015. The observations of dataset 1 describe "Comparative Illness Factor" (CIF), which is an indicator of health conditions, for two data zones, namely "City Centre West 01" and "City Centre East 04". The two data zones are neighboring areas within Aberdeen, a city in North East Scotland and the third most populous city in the country. The value of CIF is 55 and 95 for the "City Centre West 01" and the "City Centre East 04" of Aberdeen, respectively. Similarly, the observations of dataset 2 describe the percentage of employment deprivation in the same data zones, which is 6% in the "City Centre West 01" and 9% in the "City Centre East 04".

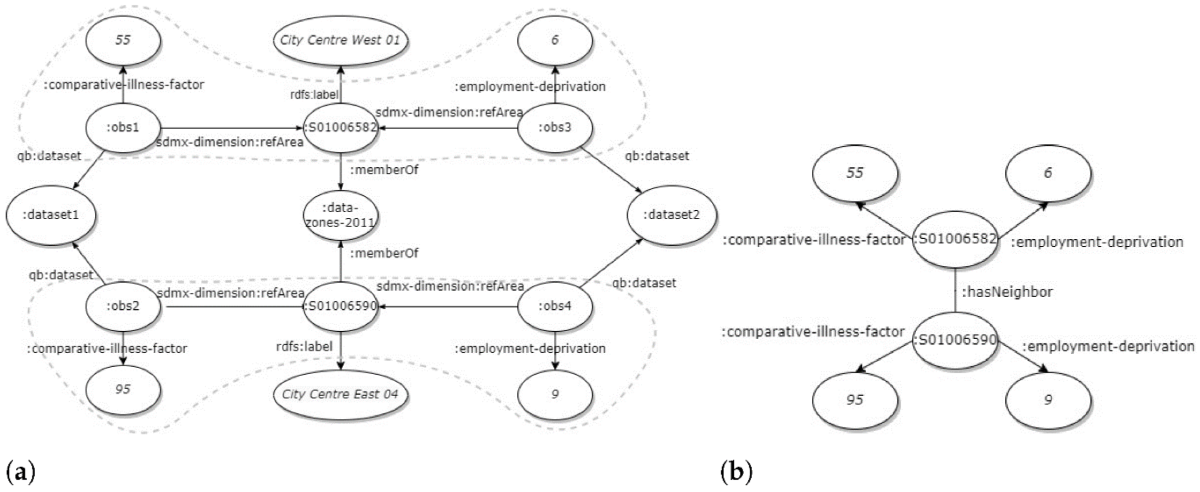


Figure 6 (**a**) The nodes of the Scottish linked statistical indicators graph that will be used for constructing the GNNs. (**b**) Part of the final graph after the transformation of the linked data graph [17].

In Figure 6 (a), the part of the geo-centric graph that will be used for constructing the GNNs is selected. Only nodes with information regarding the "2011 data zones", the value of the measure of the statistical indicator, and the values of additional properties (e.g., gender, age, etc.) are required. All other information, such as the dataset or the observations, is excluded. The final sub-graph is presented in Figure 6 (b). The edge connecting the data zone nodes represents the neighboring relationship between the data zones.

The final graph comprises a total of 6014 interconnected nodes (or data zones), with 20,521 edges representing the adjacency of the respective data zones (Figure 7). Consequently, there is a connection between each data zone and its adjacent data zones. The hue of the data zones signifies the mean house price within its locale, ranging from deep blue to intense red as the mean house price ascends, thereby offering a visual gradient of housing costs.

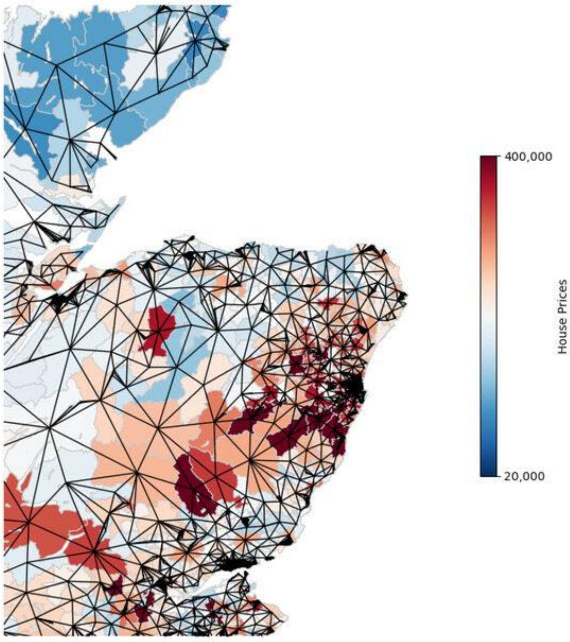


Figure 7 Subset of the final graph used for node classification using graph representation learning. Each node corresponds to a data zone, while Edges connect the centroids of adjacent regions. Blue data zones indicate regions with lower mean house prices, while red indicates higher mean house prices [17].

(3) Predict house prices.

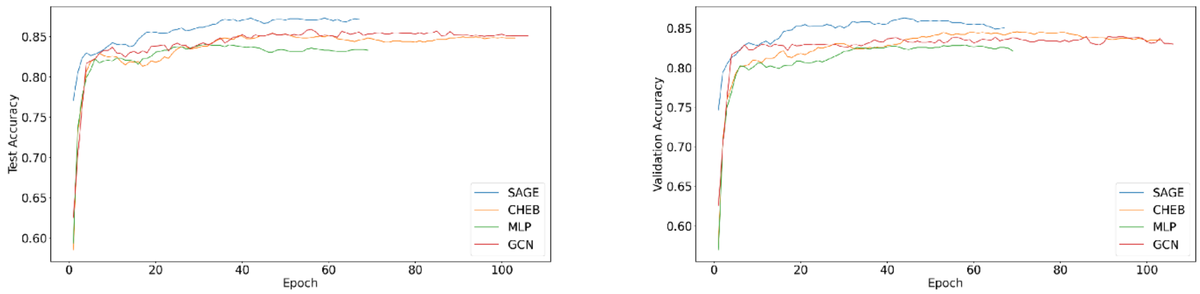
Method: This case study leverages graph modeling for the prediction of house prices. Three distinct variants of GNNs are employed, each representative of the two key methodologies concerning graph convolutions: the Chebyshev Neural Network (ChebNet), Graph Convolutional Network (GCN), and GraphSAGE. The problem is formulated as a node classification task for classifying Scottish “2011 data zones” into two categories: above the mean house price of all data zones or below the mean house price of all data zones. The training/validation/test split of this dataset is aligned with the fully supervised scenario. To this end, all labels of the training examples are used for training following the implementation of [6], with a split ratio of 0.6/0.2/0.2, respectively. The created models are evaluated and then compared with each other. They are also juxtaposed with the model created in a previous work [18], which utilized the XGBoost machine learning algorithm and a straightforward Multilayer Perceptron (MLP), based on the same dataset.

Results: The three GNN variants that were implemented namely, Graph Convolutional Network (GCN), Chebyshev Neural Network (ChebNet), and GraphSAGE are compared against an XGBoost model that has been previously tested on the same dataset and against a Multilayer Perceptron (MLP) that acts as a per-node baseline classifier that does not incorporate the underlined graph structure.

The case study utilizes an undirected, attributed graph, wherein each node incorporates a 59-feature vector. This graph comprises 6014 nodes and 20,521 unweighted edges. Prior to the implementation of the experiments, all features are normalized to a 0-1 scale to ensure consistent comparability across various feature ranges. The implementation details of the experiments are the following. All networks consist of two layers followed by a Relu non-linear activation function, and the final layer is used for binary classification followed by a logistic sigmoid activation. The learning rate of all networks is 0.01 and the dimension of hidden units is set to 32, while the dropout rate is 0. Furthermore, the Adam optimization method [21] is selected for all GNN models during training. Early stopping was implemented to mitigate the potential of model overfitting. If no improvement in validation accuracy was observed over 20 consecutive epochs, the training process was terminated. The cross-entropy loss function was adopted to assess the performance of all neural networks.

The original GCN is modified to the inductive setting following the work of [6]. On the contrary, GraphSAGE applies neighbor sampling and aggregation compared to the GCN that aggregates features from all neighbors. To this end, the mean aggregator is selected for GraphSAGE, with a neighbor sampling size set to 20 and 10 for each layer. In addition, for the Chebyshev spectral graph convolutional operator, the size of the Chebyshev filter K is set to 2.

Figure 8 shows the test and validation learning curves of the three GNN variants and MLP. Notably, GraphSAGE exhibited the highest accuracy scores, both in test and validation scenarios, from the very beginning of the training process. This suggests a promising level of learning efficiency and model stability in GraphSAGE. GraphSAGE's early high accuracy scores can be attributed to its robust learning mechanism, which effectively harnesses the features of the local neighborhood of each node. The capacity to employ sampling to aggregate information from the node's neighborhood appears to confer an initial advantage to GraphSAGE over the other architectures in the test and validation plots. Moreover, a high initial accuracy suggests that GraphSAGE requires fewer epochs to reach an optimal or near-optimal performance, leading to reduced training times and resource expenditure, while all other models are trained for longer epochs.



(a) (b)

Figure 8 Learning curves of the three different GNN variants and MLP. (a) Test learning curves. (b) Validation learning curves [17].

In addition, Figure 9 shows the training runtimes for the different approaches. All training times are comparable, with the GCN being the slowest and GraphSAGE the fastest due to its ability to sample and aggregate node information instead of aggregating from all nodes (as all other GNN variants do).

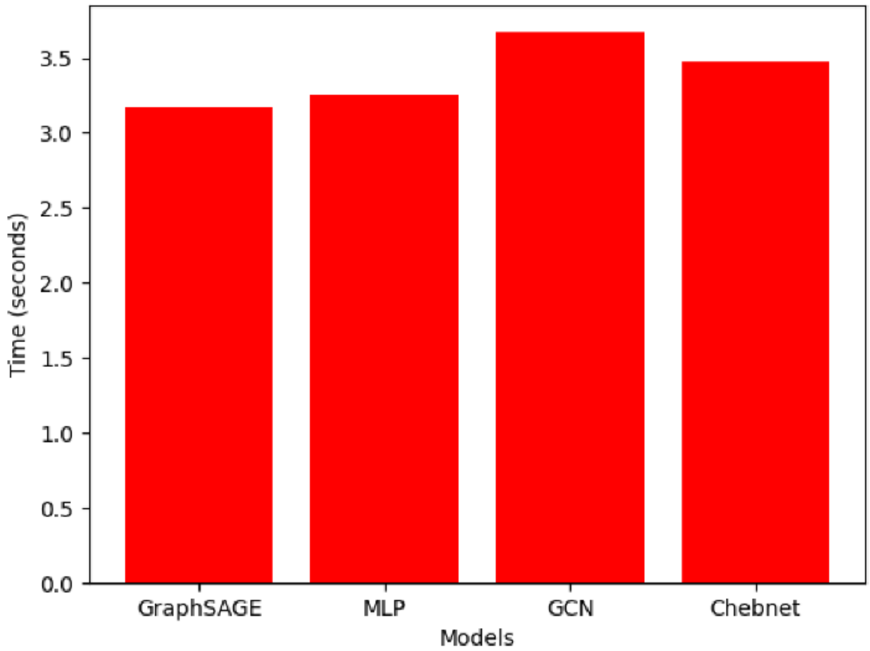


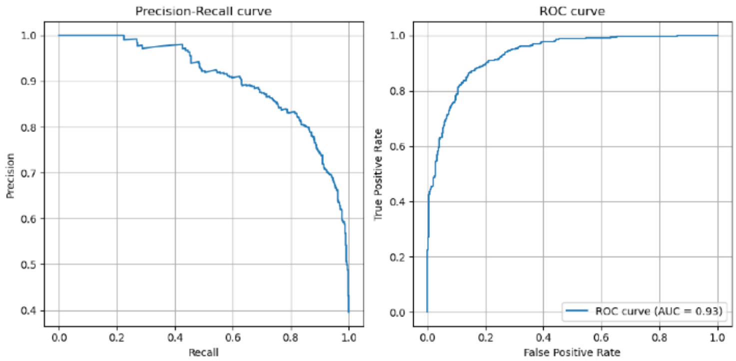
Figure 9 Training times (in seconds) of the GNN variants and MLP [17].

To validate the representational power of GNNs to incorporate the spatial dependencies among data zones, comparative evaluation results are summarized in Table 1. Moreover, to ensure the reliability and robustness of the performance comparison, each model was executed 30 times. The mean values of the performance metrics across these runs are displayed in Table 6. The evaluation involves common classification metrics, including accuracy, precision, recall, F1 score, and Area Under the ROC Curve (AUC-ROC) score. The evaluation metrics of the XGBoost model have been drawn from the prior work in [18].

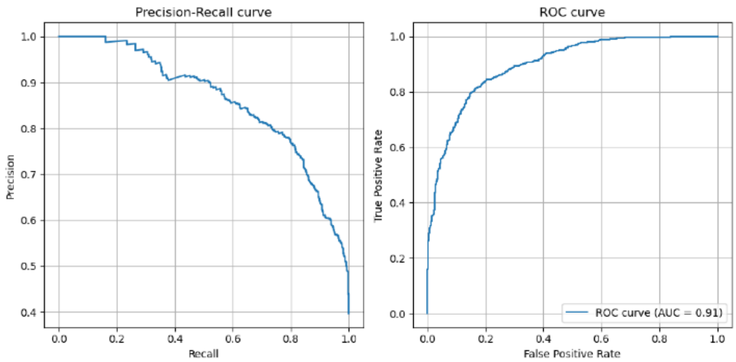
Model	Accuracy	Precision	Recall	F1	ROC-AUC	Epochs
GraphSAGE	0.876	0.876	0.876	0.876	0.93	68
GCN	0.852	0.852	0.852	0.852	0.91	112
ChebNET	0.847	0.847	0.847	0.847	0.91	103
XGBoost	0.840	0.850	0.840	0.840	0.92	-
MLP	0.827	0.832	0.827	0.827	0.90	72

Table 6 The averaged test prediction results for the supervised node classification task for the GNN variants, MLP, and XGBoost ($p < 0.05$).

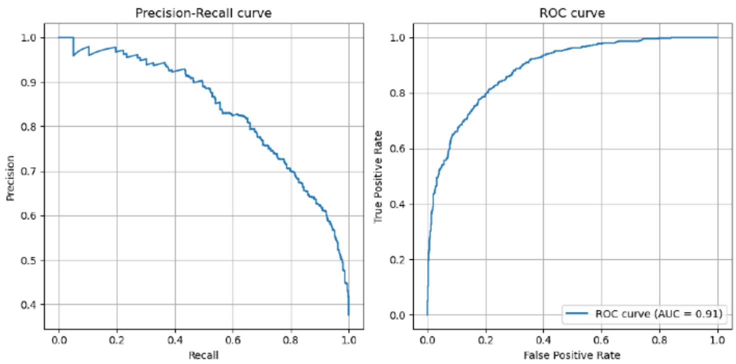
To validate the differences in performance between the models, paired t-tests were conducted between all pairs of models. This statistical significance testing was essential to confirm that the observed differences were not due to random chance but were statistically significant. The results of the paired t-tests, show that the differences in performance metrics between all models are statistically significant ($p < 0.05$). The GraphSAGE model achieves the highest accuracy, precision, recall, and F1 score among all models, coming in at 0.876 for each of these metrics, while also having an AUC-ROC value of 0.93. Additionally, GraphSAGE accomplishes these results in 68 epochs, indicating a relatively efficient learning process compared to the other graph-based models like GCN and ChebNET, which required 112 and 103 epochs, respectively. The GCN and ChebNET models perform similarly on accuracy, precision, recall, and F1 score, each achieving around 0.85, but their AUC-ROC values indicate some difference in model performance. The GCN's and ChebNet's AUC-ROC values are 0.91, indicating a slightly better performance at distinguishing between the two classes. Considering XGBoost and the MLP, both methods fall short in comparison to the graph neural network models. Although XGBoost has an AUC-ROC score of 0.92, which is higher than the GCN, its overall accuracy, precision, recall, and F1 score are lower than all three GNN models. The MLP has the lowest performance across all metrics, despite having a competitive AUC-ROC score of 0.90. Figure 10 depicts all the precision-recall and receiver operating characteristic curves for all GNN variants and the MLP [17].



(a)



(b)



(c)

Figure 10 Precision-recall and Receiver Operating Characteristic (ROC) curves of different models. (a) GraphSAGE. (b) GCN. (c) ChebNET. (d) MLP [17].

In addition, Figure 11 depicts the classification metrics further by displaying the results for each class (0 and 1) separately. In general, the GraphSAGE model outperforms other models across most of the metrics and for both classes, demonstrating its strength in the binary classification task. This is particularly noticeable in the case of accuracy, where it consistently achieves top results. Its superiority extends to the recall for class 1, F1 score for class 0, and precision for class 1. However, there are some metrics where other models exhibit stronger performance. Interestingly, the MLP model surpasses other models in terms of recall for class 0. This suggests that the MLP, while not the best overall performer, is particularly adept at identifying true positives in the instances that are actually of class 0. Similarly, XGBoost shows effectiveness in the precision metric for class 0, indicating that when it predicts an instance to be class 0, it is likely to be correct. It also achieves higher scores for F1 class 1, demonstrating its balanced performance between precision and recall for instances of class 1. The figure also reveals some of the models' weak points. For instance, the GCN appears to struggle with precision for class 0, where it ranks last among the models. This suggests that it has a higher rate of false positives when predicting class 0, which could be a significant drawback if precision in identifying this class is a priority.

Figure 11 Visual comparison of the performance metrics precision, recall, F1 score, and accuracy of the three GNN variants, XGBoost, and MLP. The metrics are evaluated separately for the entire dataset ('All'), instances above the mean house prices threshold ('Class 0') [17].

Figure 12 shows the UMAP (Uniform Manifold Approximation and Projection) projection of the node embeddings learned by the top performing model GraphSAGE into a 2D space. The UMAP visualization shows a distinct separation between the two classes, which indicates that the model has learned embeddings that can effectively distinguish between the classes.

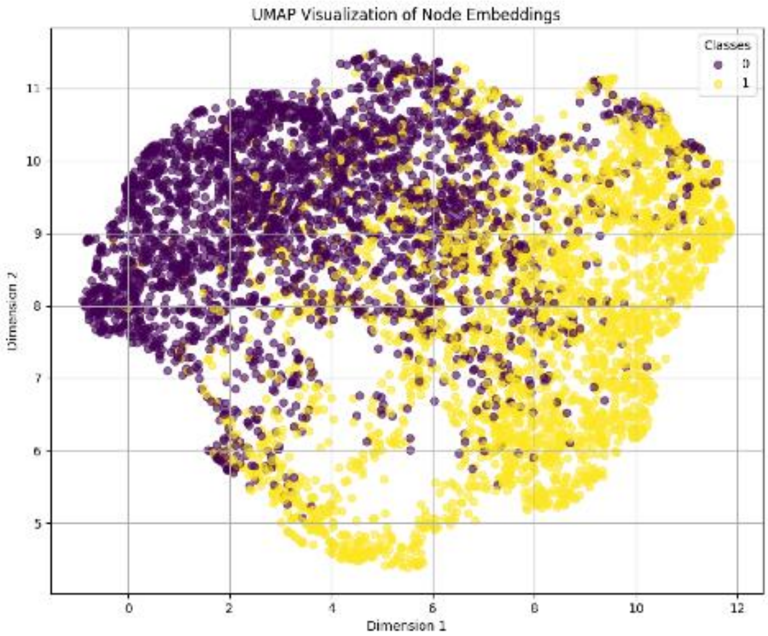


Figure 12 Node embeddings learned by GraphSAGE visualised using the UMAP method (Uniform Manifold Approximation and Projection) [17].

4.1.3.2 Ethical, Legal, and Safety Testing

This activity is operationalized in this case study through the explanation the model. To facilitate an in-depth understanding of the mechanisms behind the model's predictions, both local and global explainability are used to provide comprehensive insights into the model's behavior.

Global explainability is tackled through the implementation of a simple surrogate model using logistic regression, chosen for its interpretability and simplicity. This model is trained on the predictions of the superior performing GNN based on the results of the previous step, which is GraphSAGE. Specifically, the model is trained on the original features in combination with the predicted probabilities of the binary classification outcomes from the GraphSAGE model. The subsequent calculation of feature importance is based on the coefficients of the logistic regression. It is assumed that the magnitude of the coefficients inherently determines the relative importance of each feature for the GNN's predictions. Therefore, these coefficients form the basis for interpreting the GraphSAGE's prediction behavior on a global level.

The predicted probabilities produced by the GraphSAGE model along with the initial features are used to train the logistic regression. The higher the absolute value of the coefficient, the more significant the feature is in predicting whether a house price is above or below the mean price. Each provided coefficient represents the change in the log-odds of the target variable given a one-unit change in the predictor, all other variables being held constant. Figure 13 depicts feature importance as determined by the coefficients of the surrogate model. Each bar corresponds to a feature used in the GNN model, and the length of the bar signifies the magnitude of the feature's impact on the model's predictions. The direction of the bar indicates the polarity of the coefficient, i.e., whether the feature contributes positively or negatively to a house price being above the mean price.

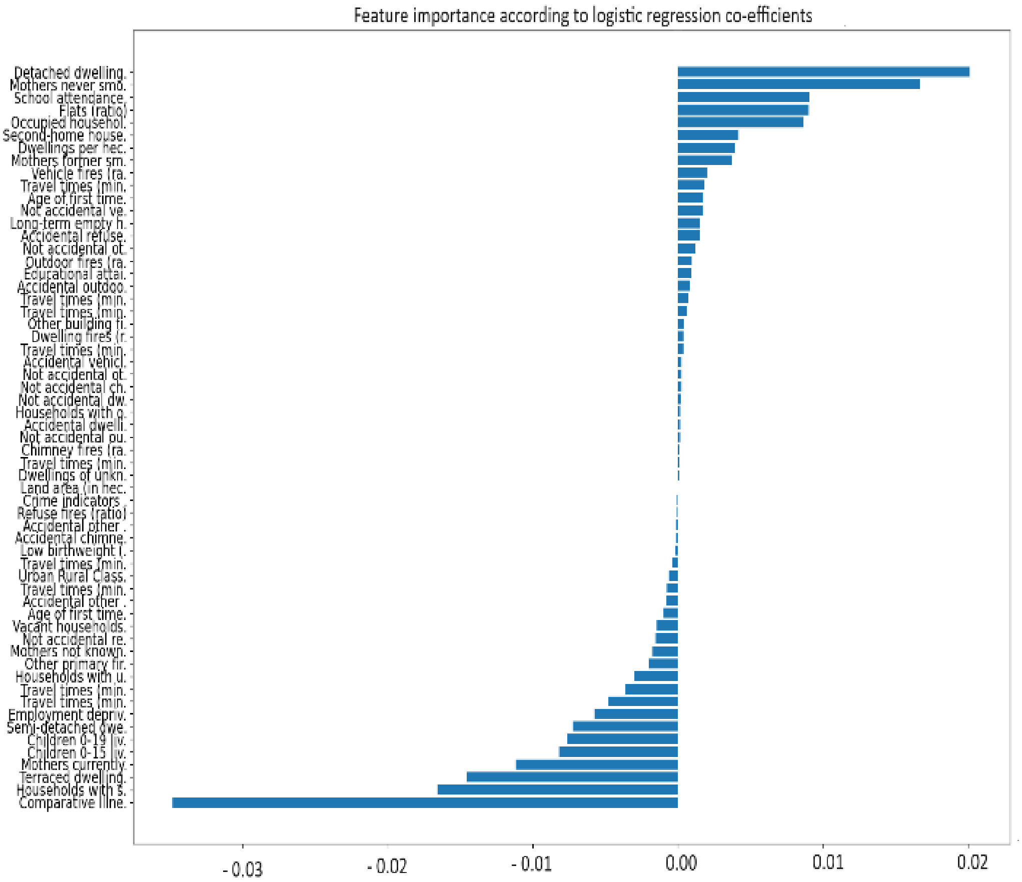


Figure 13 Global feature importance based on the logistic regression surrogate model. Each bar signifies the magnitude of the feature's coefficient in the logistic regression model, which is interpreted as its importance in the GNN's predictions.

Notably, several indicators are negatively associated with a data zone having house prices above the average. For example,"Comparative Illness Factor" (CIF) shows the largest negative coefficient, suggesting that data zones with higher values for the CIF tend to have lower mean house prices. Similar trends can be observed for features such as "Households with single adult discounts (ratio)", "Terraced dwellings", and "Mothers currently smoking (ratio)". On the other hand, factors that increase the likelihood of a data zone having house prices above the mean are also identified. The most substantial of these is the ratio of "Detached dwellings", suggesting that regions with a higher proportion of detached dwellings are more likely to have higher house prices. Other significant positive factors include the proportion of "Mothers never smoked", "Occupied households ratio", "School attendance ratio", and "Flats ratio". This could reflect the impact of education and healthier living environments on house prices. Coefficients for travel times to different services (like retail centres, schools, and GP surgeries) by car and public transport are mostly negative, suggesting that longer travel times might be associated with lower house prices. However, these coefficients are small, implying that these factors might not be influential in determining whether house prices are above or below the mean. Features related to fires show both positive and negative associations with house prices. For instance, “Accidental outdoor fires” and “Vehicle fires” have positive coefficients, indicating a slight increase in the likelihood of house prices being above the mean, while “Other primary fires” and “Accidental refuse fires” are negatively associated.

Local explainability, on the other hand, is addressed through the application of a well-established method, GNNExplainer [49]. Toward this direction, two neighboring data zones are selected, both with house prices above the mean. For each data zone, local feature importance is computed, determining the most critical features that influence the model’s prediction for the specific node. Additionally, an explanation sub-graph is visualized, highlighting the most significant nodes, edges, and features that strongly impact the prediction. Interestingly, these two data zones have previously been examined in another study [18] using the XGBoost model and the SHAP explainability framework. This previous research offers a valuable benchmark, allowing for the comparison of results, not merely on the level of predictive performance but also in terms of explainability and feature importance.

The GNNExplainer is applied to two specific nodes corresponding to data zones S01006552 or Hazlehead-06 and S01006553 or Summerhill-01. Hazlehead-06 and Summerhill-01 are adjacent data zones located within the council area of Aberdeen city, in the north-eastern region of Scotland. The average house price in these zones is GBP 257,621 and GBP 251,658, respectively, both of which are higher than the average house prices across Scotland. The GraphSAGE model correctly predicted them as having house prices above the mean.

To understand and interpret the model’s predictions for these two data zones, the explanation sub-graph is visualized. Figure 14 and Figure 15 depict the explanation graphs of Summerhill-01 and Hazlehead-06 data zones, respectively, illustrating the most influential nodes and edges contributing to this node’s prediction. Each node represented in this sub-graph is characterized by its most important feature. The sub-graph also depicts the structural interdependencies between the target node and its neighboring nodes. The edge opacity in the explanation sub-graph signifies the importance of each connection, with more opaque edges corresponding to stronger influential relationships.

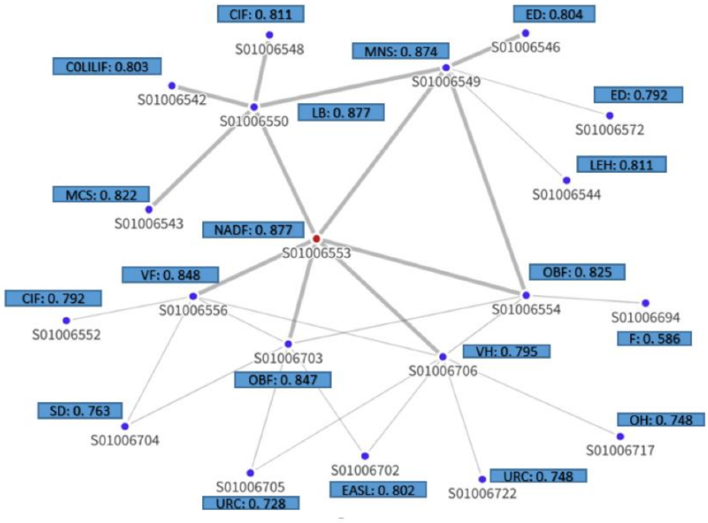


Figure 14 Visual representation of the explanation sub-graph for data zone Summerhill-01 (S01006553), highlighted in red. Each node in the sub-graph is accompanied with a feature and its importance score, determined as the most influential by GNNExplainer for predicting the class of the target node. The opacity of the edges reflects their respective importance in this prediction process.

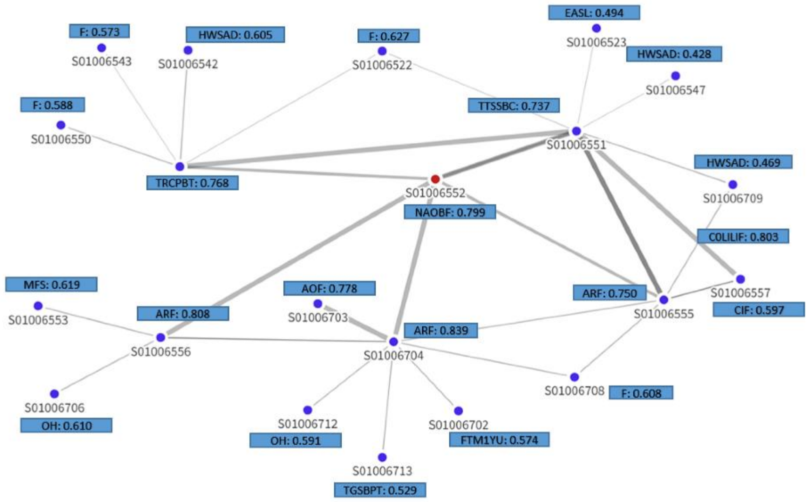


Figure 15 Visual representation of the explanation sub-graph for data zone Hazlehead-06 (S01006552), highlighted in red. Each node in the sub-graph is accompanied with a feature and its importance score, determined as the most influential by GNNExplainer for predicting the class of the target node. The opacity of the edges reflects their respective importance in this prediction process.

It is noted that all nodes within a one-hop distance from the target node have highly influential connections for both data zones, suggesting that the immediate neighborhood of a node plays a crucial role in the model's predictions. It is observed that not all regions contribute equally to the prediction outcome. In the case of Summerhill-01 (Figure 14), data zones like S01006704, S01006705, S01006702, S01006717, S01006552, and S01006722 appear to have less influence on the prediction of the target node, all of which display house prices that fall below the mean. Furthermore, the “Employment deprivation” (ED) feature appears several times across the nodes, suggesting that it is a significant predictor for the target node. Similarly, “Comparative Illness Factor” (CIF), “Other building fires” (OBF), and “Urban Rural Classification” (URC) also come up multiple times across nodes, suggesting their significant role in predicting the target variable.

Figure 16 depicts the total feature importance for data zone Summerhill-01 as the sum of the node mask values across all nodes for each feature. This plot aggregates the importance of each feature across all nodes of the explanation sub-graph, offering a comprehensive understanding of the relative importance of features. As a result, “Urban Rural Classification” is ranked first in terms of importance, followed by “Comparative Illness Factor”, “Employment Deprivation”, “Educational Attainment of School Leavers”, and “Detached Dwellings”. In this case, the local interpretation for classifying this specific region as an area likely to have house prices above the mean aligns well with the global interpretation. It is observed that features like “Urban Rural Classification” and “Comparative Illness Factor”, which hold high importance on a global scale, also play a substantial role in driving the local prediction for this particular region.

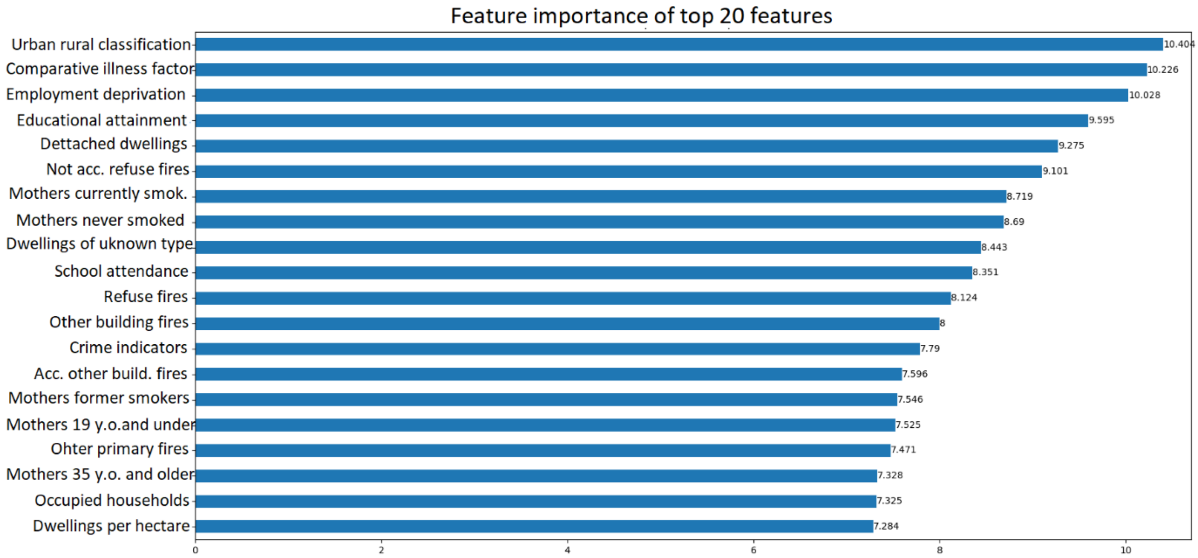


Figure 16 Feature importance of the top 20 features that play a crucial role in explaining the prediction made by GraphSAGE for the data zone Summerhill-01. The total importance score is based on the sum of the node masks (obtained during model explanation) across all nodes for each feature.

Finally, in the case of Hazlehead-06, Figure 15 depicts stronger relations of the target node with the one-hop neighbors, as well as data zones S01006703 and S01006557. “Flats” (F), “Accidental Refuse Fires” (ARF), “Occupied Households”, and “Households with single adult discounts” (HWSAD) are features that prominently appear across multiple nodes, suggesting their pivotal role in forecasting whether house prices in Hazlehead-06 will surpass the mean value.

However, as seen from the feature importance plot in Figure 17, other features like “Comparative illness factor” and “Dwelling fires” have high aggregate importance scores. While the explanation sub-graph provides valuable insights into the structural dependencies and reveals localized feature influences, the feature importance plot captures the collective importance of each feature in the model’s predictive behavior by aggregating the importance scores of all features across all nodes of the explanation graph. Furthermore, it is noteworthy that key features such as “Comparative Illness Factor”, “Flats”, “Mothers who are Former Smokers”, and “Occupied Households” appear in both global and local importance plots. The consistency indicates that despite their different methodologies and assumptions, both models perceive these particular features as having a significant influence on the prediction of house prices in Hazlehead-06.

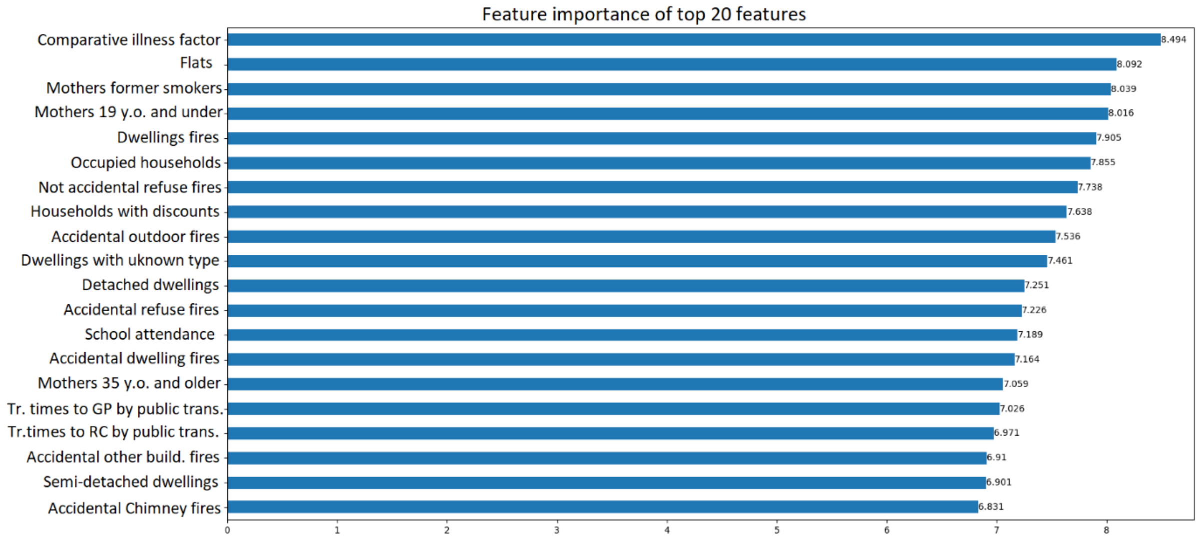


Figure 17 Feature importance of the top 20 features that play a crucial role in explaining the prediction made by GraphSAGE for the data zone Hazlehead-06.

4.1.3.3 User Testing and Feedback Integration

In contrast to the other pilot case studies, PCS1 was conducted as a research-oriented experiment and did not involve systematic user testing with public servants or policymakers. Nevertheless, elements of implicit user testing were incorporated through internal expert review and reflective assessment of the prototype’s usability, interpretability, and potential integration into public-sector workflows.

First, the explainability outputs of the models, global feature importance via the logistic regression surrogate and local explanations via GNNExplainer, were critically examined by the research team, including domain experts familiar with housing policy and socio-economic indicators. This internal review assessed whether the explanations were: (i) consistent with domain expectations (e.g., the role of dwelling types, deprivation, and health indicators), (ii) presented in formats that could plausibly be understood by non-technical decision-makers, and (iii) sufficiently granular to support diagnostic use (e.g., differentiating between neighbouring data zones). The review confirmed that the explanations broadly aligned with established knowledge about the Scottish housing context, thereby increasing confidence in the face validity of the model's behaviour. At the same time, it highlighted the need for more user-friendly visualisation and narrative explanation layers if such tools were to be used by non-technical staff.

Second, the team performed a qualitative assessment of how the predictive and explanatory outputs would fit into existing analytical practices in public administration. In particular, the learning curves, comparative performance metrics, and feature-importance plots were examined from the perspective of policy analysts: could these outputs be used to explore "what-if" questions, identify areas with atypical price dynamics, or support the design of targeted interventions? This reflection suggested that, while the current prototype is well suited for expert analytical use, additional work would be required to translate the technical artefacts into dashboards, decision-support reports, or interactive tools aligned with everyday policy workflows and time constraints.

Third, the pilot informed initial hypotheses about user requirements and organisational readiness, which are relevant to later phases of the AIGOV Transformation and Adoption Framework. For example, the complexity of GNN-based explanations underscored the importance of training and capacity building (e.g., basic concepts of graph models and explainable AI), the need for clear governance of how model outputs may be used (or not used) in sensitive policy decisions, and the necessity of embedding human-in-the-loop review whenever outputs could have redistributive or distributional consequences.

Because no formal usability testing or participatory evaluation was conducted, the KPIs related to user interpretability, decision-support suitability, and public-sector adoption remain defined but unmeasured for PCS1. However, the qualitative insights obtained from internal expert review and methodological reflection constitute an important input for future work: they specify the kinds of interfaces, documentation, and governance mechanisms that would need to accompany any future deployment of explainable GNN models for housing-market analysis in a real administrative environment.

4.1.4 Assessment of the KPIs

The tables below present the values for the KPIs related to this PCS.

KPI	Assessment	Status
Accuracy	0.876	Measured
Precision	0.876	Measured
Recall (Sensitivity)	0.876	Measured
F1-Score	0.876	Measured

Comparative Performance Score	GraphSAGE outperforms all baselines; +3.6 percentage points over XGBoost; +4.9 points over MLP	Measured
Model Robustness / Stability	Stable across 30 repeated runs	Partially assessed

Table 7 Assessment of Technical Performance KPIs for PCS1

KPI	Definition	Status
Explainability Score	Achieved: Global importance extracted via surrogate logistic regression; local explanations valid for two data zones; influential features traced	Partially assessed
Traceability Score	Strong traceability: Both global and local XAI methods successfully connected predictions to specific socio-economic indicators and spatial relations	Partially assessed
User Interpretability Index	n/a	Defined (future validation)
Documentation Completeness Indicator	Data provenance and modelling methodology fully documented; explainability documentation partially complete	Partially assessed

Table 8 Assessment of Explainability and Transparency KPIs for PCS1

KPI	Definition	Status
Decision-Support Suitability Indicator	n/a	Defined (future validation)
Fairness and Bias Assessment Score	n/a	Defined (future validation)
Reproducibility and Openness Index	High reproducibility due to open datasets, documented SPARQL queries, and standard GNN architectures	Partially assessed
Public Value Impact Potential	n/a	Defined (future phase)
Risk Assessment Rating	n/a	Defined (future phase)
Governance and Human Oversight Readiness	n/a	Defined (future phase)

Table 9 Assessment of Responsible AI and Public-Sector Applicability KPIs for PCS1

4.2 Pilot Case study 2: Evaluating Open and Proprietary Large Language Models in Law Interpretation: The Case of the EU VAT Directive

4.2.1 Phase 1: Assess Readiness and Context

4.2.1.1 Strategic Context and Public Value

It is estimated that generative Artificial Intelligence (AI) could elevate the global Gross Domestic Product (GPD) by 7% over the next decade and, at the same time, bring forth automation to approximately 300 million jobs worldwide, contributing to the global economy $2.6 trillion to $4.4 trillion annually [7]. AI has long found its way into the public sector [32], yet the impact of utilizing Large Language Models (LLMs) will likely be unprecedented. LLMs utilize the wealth of information provided by the public sector, from documents across different departments, ministries, and local authorities to Open Government Data portals, to enable, for instance, the deployment of chatbots and virtual assistants [2]; the analysis of documents for identifying key information in complex documents such as legal contracts [26]; the summarization of large volumes of text; the extraction of information from administrative documents of the public sector [50]; and the assistance in decision-making by generating reports and evaluating applications and grants [15]. The capabilities of LLMs to facilitate access to open government data has also been explored [29]. The introduction of LLMs can also lead to more proactive and data-driven public interventions, having insights in areas such as public health, urban planning, and disaster response [36, 40].

While recent studies show notable progress in using LLMs for legal interpretation, they also highlight ongoing challenges [39]. Even though capabilities have advanced considerably, research on applying LLMs to specific domains, such as the complex regulatory systems, remain an active area of research.

This pilot case study therefore builds upon this foundation by comparing the performance of LLMs in interpreting the European Union's Value Added Tax (VAT) Directive, hence, addressing the gap in the literature regarding model efficacy in specialized legal domains. Towards this direction, an exploratory case study is presented and analyzed that involves nine proprietary and open LLMs from four major families, namely Claude, GPT, Mistral, and Llama. The LLMs are tasked with answering a set of questions related to the EU VAT Directive. The questions, which are of varying complexity and of a legal nature, are provided as prompts to the LLMs, and their responses were evaluated based on their legal precision.

The public value of such experimentation lies in:

- Increasing accessibility of legal frameworks for civil servants, SMEs, and citizens;
- Supporting multilingual legal harmonisation across EU Member States;

- Enhancing evidence-based policy and regulatory compliance through consistent interpretation;
- Identifying governance requirements and safeguards before real-world application.

In line with the AIGOV Transformation and Adoption Framework, this case study represents an exploratory step toward understanding whether LLMs can responsibly contribute to legal research support, drafting assistance, and interpretive consistency, while maintaining transparency, human oversight, and legal certainty.

4.2.1.2 Stakeholders

The stakeholders for this pilot case study include actors who would either directly interact with AI-assisted legal interpretation tools or be affected by their deployment within public administration environments. As this study explores the potential use of LLMs to support legal reasoning based on the EU VAT Directive, stakeholders are mapped according to the internal and external stakeholder classification adopted in AIGOV Deliverable D1.2.

Internal Stakeholders include:

- *Public authorities and public organisations including regulatory bodies*. National tax administrations, EU institutions (e.g., DG TAXUD), and ministries responsible for financial regulation or legislative harmonisation could use such tools to support internal analysis, interpret legal provisions, or compare national implementation against EU legal baselines.
- *Policy makers and legal interpreters*. Civil servants, legal experts in public bodies, and legislative drafting units may use such systems to accelerate legal research, support analysis of amendments, or explore consistency across multilingual legal versions.
- *Public sector legal practitioners and auditors*. Professionals such as in-house legal counsel, procurement specialists, financial auditors, and VAT compliance units may interact with these systems to verify interpretation, cross-reference provisions, or support case preparation.
- *IT and data governance teams*. Technical personnel including AI engineers, cybersecurity teams, data protection officers (DPOs), and digital service designers would be responsible for evaluating integration feasibility, ensuring data protection compliance (GDPR), and maintaining accountability structures.

External Stakeholders include:

- *Businesses and economic operators*. Enterprises subject to VAT compliance including SMEs, large corporations, accountants, and tax advisors, which may benefit from improved clarity and accessibility of VAT regulations. These groups stand to gain from reduced compliance burden and more consistent interpretation.

- *Citizens and civil society organisations*. Individuals and advocacy organisations focused on tax justice, transparency, SME support, or digital rights may monitor outcomes, especially regarding fairness, accessibility, and legal interpretation risk.
- *Academic and research communities*. Researchers in AI ethics, legal informatics, computational law, and public administration may engage with the results to benchmark model behaviour and assess risks or improvements in legal-domain adaptation.
- *Technology vendors and open-source communities.* Providers of commercial and open models (e.g., OpenAI, Anthropic, Mistral, Meta) as well as open-source contributors represent stakeholders who may refine architectures or evaluate domain-specific fine-tuning.

Since this is a controlled pilot conducted in a research context rather than an operational deployment, stakeholder involvement remains conceptual rather than participatory. Nevertheless, mapping stakeholders at this stage is essential for anticipating governance requirements, accountability allocation, ethical implications, and potential real-world adoption pathways.

This structured understanding of the stakeholder ecosystem ensures alignment with the AIGOV principles of transparency, public interest, inclusivity, and responsible innovation, laying the groundwork for meaningful engagement in subsequent phases.

4.2.1.3 Data readiness

This pilot case study evaluates the performance of multiple Large Language Models (LLMs) in interpreting legal text from the EU VAT Directive. Unlike predictive modelling use cases that rely on structured data or large annotated corpora, this case study relies primarily on authoritative legal texts and manually curated evaluation prompts. The data readiness assessment therefore focuses on the quality, structure, accessibility, and legal compliance of the textual sources used to assess the models.

The primary dataset used in this study is the consolidated English-language version of the EU VAT Directive (Council Directive 2006/112/EC), obtained from official EU legal repositories. The Directive was selected as the authoritative source because of its harmonising role across Member States and its high relevance to public administration functions such as taxation, public finance, customs, compliance auditing, and economic governance.

In addition to the Directive text, the study makes use of a set of evaluation questions of varying complexity, constructed to test LLM reasoning across distinct dimensions including factual recall of specific legal provisions, ability to interpret definitions and scope, reasoning involving exceptions, exemptions, or conditional rules, multi-step legal reasoning requiring cross-article references. These questions were developed manually based on legal methodology rather than sourced automatically, ensuring relevance and meaningful complexity. Responses produced by the evaluated LLMs were subsequently benchmarked against legally accurate reference answers.

From a readiness perspective, the dataset meets key criteria of accessibility, quality, and governance:

- Authoritative Source and Reliability. The Directive text originates from EUR-Lex, ensuring accuracy, provenance, and trustworthiness.
- Structured and Machine-Readable Format. Although the Directive is written in natural language, its legal structure (e.g., articles, chapters, and hierarchical numbering) provides a logical format suitable for evaluation. However, no semantic markup (e.g., XML/ELI-compliant metadata) was used in this pilot. Future iterations may benefit from structured legal encoding to improve machine interpretability.
- Language Availability. While the Directive exists in all EU official languages, this pilot focuses exclusively on the English version. Multilingual components, relevant to EU-scale deployment, were therefore not evaluated at this stage.
- No Personal Data or Sensitive Content. The study does not process personal data and therefore does not trigger GDPR compliance risks. However, the downstream use of legal interpretations could affect regulated entities if deployed operationally, requiring future safeguards.
- Human-Curated Evaluation Dataset. The question set and reference answers were developed and validated by legal experts, supporting accuracy and methodological robustness. However, the dataset remains comparatively small and exploratory rather than exhaustive or statistically representative, consistent with the pilot nature of the study.

Overall, data readiness for this case study is assessed as high for experimental evaluation purposes, but future operationalisation would require:

- expansion of the evaluation dataset, including multilingual prompts,
- representation of edge-case scenarios and domain exceptions,
- adoption of legal AI benchmarking standards (e.g., LexGLUE, LegalBench), and
- potential integration of structured legal knowledge sources or domain ontologies.

This ensures the dataset not only supports model benchmarking, but aligns with public sector standards for accuracy, reproducibility, governance, and responsible AI deployment in legal contexts.

4.2.1.4 Governance, Ethics, and Legal Compliance

The use of LLMs in legal interpretation requires careful consideration of governance, ethical safeguards, and regulatory compliance due to the potential societal and institutional implications of automated reasoning in law. Although this pilot is exploratory and does not involve live deployment, the evaluation framework incorporates responsible AI requirements to ensure alignment with public-sector expectations and applicable regulatory frameworks.

Because the study uses publicly available legal texts rather than personal or sensitive information, the pilot does not trigger obligations under the General Data Protection Regulation (GDPR). However, the nature of the task per se (automated legal interpretation) places the use case within a high-stakes domain under the forthcoming EU Artificial Intelligence Act (AI Act). If implemented operationally, such a system could fall under classifications related to high-risk AI due to its potential impact on legal certainty, taxation outcomes, or rights of individuals and businesses. Accordingly, safeguards and governance considerations are integrated at the evaluation stage.

Ethical and Responsible AI Considerations. Key ethical risks associated with LLM-based legal reasoning include:

- Hallucination and false interpretation, where the model provides incorrect or fabricated legal information.
- Overconfidence bias, where models express incorrect answers as authoritative legal statements.
- Opacity of reasoning, making it difficult for legal experts or public servants to verify why a particular interpretation was generated.
- Risk of uneven performance across legal subtopics, which may introduce bias or inconsistency.

To address these risks, human legal expertise is explicitly required in model assessment and interpretation. The model outputs in this pilot are not used for decision-making; instead, legal experts evaluate response accuracy and classify them based on correctness and interpretative soundness.

Transparency and Explainability. Unlike traditional machine learning models, LLMs generate text autonomously, making transparency critical. In this pilot:

- All prompts, answers, and evaluation criteria are recorded for traceability.
- Responses are systematically compared to verified legal interpretations.
- The evaluation process documents variations across model families and versions, supporting reproducibility and accountability.

Even though some proprietary models restrict full insight into training data and architecture, this pilot maintains methodological transparency at the evaluation level.

Legal Compliance and Alignment with Regulatory Expectations. The study integrates ALTAI, OECD, and AIGOV responsible AI principles to ensure alignment with public-sector deployment conditions. If the system were to advance beyond research, alignment with the following AI Act requirements would be necessary (Table 10):

Expected Compliance Area	Future Deployment Implication
Human oversight	Mandatory to prevent misuse or automated legal decisions
Technical robustness	Testing against ambiguity, edge cases, and updated legislation
Data governance	Verification of model training data provenance and legal-domain relevance
Documentation and record-keeping	Full audit logs of prompts, responses, and modifications
Transparency obligations	Clear disclosure that outputs are AI-generated and not legally binding

Table 10 Necessary alignments with AI Act requirements

Accountability Structures. As this case study is conducted in a controlled research environment, formal accountability roles (such as AI system owner, risk manager, or compliance officer) are not yet established. However, the pilot identifies the governance requirements that would be necessary for operational adoption, including:

- Defined responsibility between model developers, validating legal experts, and policy end-users
- Procedures for documenting errors, corrections, or systemic model weaknesses
- Clear boundaries that prevent AI-generated legal interpretations from being used without human validation.

4.2.1.5 Organisational and Capability Capacity

This pilot case study required specialised expertise and infrastructure to conduct the evaluation of open and proprietary Large Language Models (LLMs) for legal interpretation of the EU VAT Directive. While the work was carried out in a controlled research environment rather than an operational public-sector setting, the exercise provides insight into the organisational capabilities that would be needed for potential future adoption within government contexts.

The execution of this pilot drew upon multiple knowledge domains, including:

- Legal domain expertise, specifically in EU tax law and directive interpretation, necessary to design relevant test prompts and assess the correctness and interpretive nuances of model outputs.
- Artificial intelligence methodology and evaluation expertise, required to conduct systematic benchmarking across model families, versions, and output formats.
- Prompt engineering and model interaction skills, necessary to control and standardise model inputs and ensure evaluation comparability across responses.

- Responsible AI and trustworthiness assessment competencies, including knowledge of explainability, evaluation frameworks, and risk categories relevant to legal AI systems.

These competencies were available within the project team, enabling execution of the pilot without external dependencies.

Access to a mixed ecosystem of proprietary and open-source LLMs was required, including locally deployable models and API-based commercial systems. This included the ability to:

- Run models using cloud-based inference,
- Deploy open models in controlled environments where possible,
- Record and benchmark responses systematically.

No specialised high-performance computing resources were required for this pilot, as the focus was on inference and evaluation rather than model training or fine-tuning.

If a system of this type were to be operationalised within a governmental organisation, several additional capabilities (not required at the pilot stage) would become essential. These include:

- Governance capacity: establishing oversight roles, compliance protocols, model documentation procedures, and update mechanisms aligned with regulatory frameworks such as the EU AI Act.
- Technical maintenance capacity: supporting continuous model monitoring, updating, and retraining as legal texts evolve or new versions of LLMs emerge.
- Legal interpretability assurance: ensuring that outputs are reviewed by qualified legal professionals and never treated as binding or authoritative without validation.
- Change management and digital literacy training: enabling public servants to work effectively with AI-assisted legal analysis while preserving accountability and human authority.

4.2.2 Phase 2: Design Ethical and Value-Aligned Use Cases

Phase 2 focuses on translating the strategic context and readiness assessment from Phase 1 into a coherent, responsible, and evaluable AI use case. In this pilot, the objective is not only to benchmark LLMs' capability to interpret complex EU legal texts, but also to determine whether such systems could be responsibly integrated into public administration workflows, while upholding legal certainty, transparency, and human oversight.

Following the AIGOV Transformation and Adoption Framework, Phase 2 consists of three components:

- Problem and service redesign definition
- Responsible AI use case design

- Success metrics and impact definition

4.2.2.1 Problem and service redesign definition

Public administrations routinely interpret complex EU legal instruments, including the VAT Directive, in order to ensure national transposition, compliance monitoring, taxation enforcement, and policy design. Today, these processes rely heavily on extensive manual legal research, cross-referencing national and EU legislation, and engaging specialist legal expertise.

However, as identified in Phase 1, the current service landscape presents several challenges including high complexity and volume of legal texts, making interpretation time-intensive and increasing the risk of human oversight errors, differences in linguistic expression across EU official languages, complicating consistent interpretation and sometimes leading to fragmented understanding across Member States, resource constraints, especially for smaller agencies, SMEs, and administrative units that may lack specialised legal capacity, and absence of automated tools to support exploratory legal reasoning, preliminary analysis, or rapid synthesis of legal information.

At the same time, generative AI and, particularly, LLMs have the potential to enhance accessibility, speed, and consistency in early-stage legal research. However, the risks associated with LLMs (hallucination, overconfidence, opacity, bias, inaccurate legal reasoning) make direct deployment premature without structured evaluation.

Thus, the service challenge is reframed as the need to understand whether LLMs can responsibly support and not replace public-sector legal reasoning by providing consistent preliminary interpretations, clarity, rapid information retrieval across legal provisions, and explainable, auditable outputs.

The redesigned service concept does not envision LLMs issuing legally binding conclusions. Instead, AI becomes a supporting analytical tool, generating draft interpretations that remain fully subject to human legal review.

The pilot therefore focuses on defining the conditions under which LLMs could enhance and not undermine legal certainty, transparency, and administrative efficiency.

4.2.2.2 Responsible AI Use Case Design

Based on the needs identified above, the responsible AI use case consists of evaluating open and proprietary LLMs on their ability to interpret the EU VAT Directive and generate legally accurate, consistent, and explainable responses to domain-specific questions.

The design of this use case is guided by the AIGOV Framework for Trustworthy, Fair, and Accountable AI, including requirements for transparency, explainability, human oversight, and proportionality.

1. Explainability Requirements

LLMs produce free-text answers that may be difficult to verify without structured support. To mitigate opacity:

1. All prompts, reference answers, and evaluation criteria were documented and version-controlled.
2. Outputs were compared against expert-validated interpretations.
3. Variations across models (open vs proprietary; different model families) were recorded to assess consistency.
4. Limitations, ambiguities, and deviation types (misinterpretation, omission, hallucination) were systematically annotated.

This approach ensures traceability, a key requirement for legal-domain AI systems.

2. Ethical and Legal Safeguards

Even though the pilot uses only public legal texts and does not process personal data, this domain requires high ethical vigilance due to the risk of incorrect legal interpretation, the potential real-world harm if outputs were used without verification, and known LLM tendencies toward hallucination and confident misstatements.

Safeguards embedded into the use case include, for example, to express requirement that all outputs be reviewed by legal experts, explicit prohibition of automated legal decision-making, documentation of all errors and deviations, assessment of fairness and consistency across question categories.

These safeguards align with high-risk AI considerations under the EU AI Act.

3. Human Oversight and Accountability

The designed use case positions LLMs strictly as tools for non-binding legal assistance, preserving human responsibility for legal interpretation, legal experts' authority over final conclusions, and accountability mechanisms defining the role of developers, evaluators, and potential end-users. This ensures compliance with human-in-the-loop requirements of trustworthy AI frameworks and prevents over-reliance on automated reasoning.

4. Inclusiveness and Accessibility

While multilingual testing is not included in this pilot, the use case design anticipates future scaling to all EU languages, supporting harmonisation and cross-border consistency. Additionally, outputs are structured to be readable and accessible to non-technical administrative users and documentation practices enable transferability across institutions.

5. Proportionality and Public Value Alignment

The use case avoids over-automation and instead evaluates incremental, low-risk applications such as preliminary summarisation of legal provisions, identifying relevant articles, answering clarifying questions, and supporting early-stage legal research. This ensures that the AI application remains proportionate, purpose-driven, and aligned with public-sector responsibilities.

4.2.2.3 Success Metrics and Impact Definition

To evaluate the responsible AI use case, a set of Key Performance Indicators (KPIs) was defined. They cover three dimensions, namely technical accuracy (Table 11), explainability and transparency (Table 12), and responsible use in legal contexts (Table 13).

Not all KPIs can be fully quantified in this pilot prototype; some are defined for future operational evaluation. They collectively assess whether LLMs can be trusted, even in a constrained, human-supervised setting, to support legal interpretation tasks.

KPI	Definition	Status
Legal correctness score	Degree to which answers align with the GDPR and related case law, as judged by legal experts	Partially Assessed
Consistency and reproducibility	Stability of outputs across repeated prompts and runs	Measured
Agreement metrics	Measures (e.g., expert agreement or inter-system agreement) comparing outputs across model variants or configurations	Defined (future validation)
Source citation accuracy	Extent to which retrieved legal text is relevant and correctly applied in the answer	Partially Assessed

Table 11 Technical Performance KPIs for PCS3

KPI	Definition	Status
Traceability of reasoning	Degree to which the reasoning path and sources can be followed and verified	Partially assessed
Clarity of responses	Readability, structure, and legal comprehensibility of answers	Partially assessed
Auditability	Ability to reconstruct how an answer was produced (retrieval steps, agent decisions, prompts, data versions)	Defined (future validation)

Table 12 Explainability and Transparency KPIs for PCS3

KPI	Definition	Status
Alignment with legal constraints	Compliance with core legal requirements (no hallucinated provisions, faithful use of GDPR, no misleading interpretations)	Defined (future validation)

Human oversight readiness	Ease with which experts can review, contest, or refine system outputs	Defined (future validation)
Integration potential	Suitability for embedding into real governance workflows (as a legal assistant, not decision-maker)	Partially assessed
Risk indicators	Ability to identify and flag ambiguous, low-confidence, or high-risk answers	Defined (future validation)

Table 13 Responsible AI and Public-Sector Applicability KPIs for PCS3

4.2.3 Phase 3: Build and Test

Phase 3 focuses on constructing a functional prototype of the designed AI use case and evaluating its behaviour under controlled conditions. For this pilot, the goal is to empirically assess the trustworthiness of multiple Large Language Models (LLMs) when interpreting complex legal provisions from the EU VAT Directive. This phase operationalises the responsible AI design developed in Phase 2 by applying a structured, repeatable, and ethically aligned testing strategy.

Following the AIGOV Transformation and Adoption Framework, Phase 3 involves three core activities:

1. Prototype Development and Technical Implementation (Section 4.2.3.1)
2. Ethical, Legal, and Safety Testing
3. User Testing and Feedback Integration

Given that this is a research-based pilot rather than a deployed administrative service, activities focus on model evaluation and methodological rigor rather than system integration or organisational change.

4.2.3.1 Prototype Development and Technical Implementation

To explore the truthfulness of LLMs in interpreting law to answer legal questions, we utilized an exploratory case study [50]. Since LLMs are emerging technologies across various fields, including the legal one, research on their trustworthiness is still developing. An exploratory case study is particularly suitable, as a research approach, for the in-depth exploration of the field in order to enable gaining insights, explore new ideas, and identify unknown patterns.

The study focuses on nine LLMs from four families, namely Claude, ChatGPT, Mistral, and Llama. Each LLM represents a distinct case for analysis. The LLMs are selected based on their relevance in legal research, availability, and varying architectures. The majority of the selected LLMs are proprietary (five out of nine), and the rest of them are open-weight models. Regarding the size of open LLMs, 8B, 70B, and 405B were selected for the Llama family in order to test their behavior regarding trustworthiness. The Mistral model's size is not known. An overview of the nine LLMs, their access type, and specifications is presented in Table 1.

Access Type	Company	Model Series	Model Name	Size (B)	Release Date
Proprietary	Anthropic	Claude	Claude v3 Opus	Unknown	2024-Feb-29
Proprietary	Anthropic	Claude	Claude v3 Sonnet	Unknown	2024-Feb-29
Proprietary	OpenAI	GPT	GPT 4o	Unknown	2024-May-13
Proprietary	OpenAI	GPT	GPT 4 Turbo	Unknown	2024-Apr-09
Proprietary	OpenAI	GPT	GPT 3.5 Turbo	Unknown	2024-Jan-25
Open weight	Mistral AI	Mistral	Mistral Large-2402	Unknown	2024-Feb
Open weight	Meta	Llama 3.1	Llama 3.1 8b Instruct	8	2024-Jul-23
Open weight	Meta	Llama 3.1	Llama 3.1 70b Instruct	70	2024-Jul-23
Open weight	Meta	Llama 3.1	Llama 3.1 405b Instruct	405	2024-Jul-23

Table 14 Overview of access types and model specifications for the investigated Large Language Models (LLMs).

A legal expert from the Hellenic State Legal Council, highly skilled in EU tax laws, created a set of 19 progressively complex questions covering key aspects of the EU VAT Directive. The set of questions is presented below.

- Q1. Can Member States adopt practices setting limits as regards to exercising VAT deduction, according to the EU VAT Directive? Justify the answer taking into account existing EU case law and the Advocate General’s opinion.
- Q2. Can a taxable person deduct VAT paid for purchasing goods or services, in the case this person exercises both for economic and non-economic activities, according to the EU VAT Directive?
- Q3. Can tax fraud, tax evasion or other illegal practices influence the exercise of the right to deduct VAT, according to the EU VAT Directive? Justify the answer taking into account existing EU case law and the Advocate General’s opinion.
- Q4. Clarify the VAT tax obligations for taxable persons (both natural and legal) as outlined in the EU VAT Directive. Organize these obligations into categories.
- Q5. Generate the content a model invoice relying on the elements outlined in the EU VAT Directive.

Q6. How are farmers treated by the EU VAT Directive? Identify deviations compared to other taxable persons.

Q7. How is fiscal neutrality interpreted by the EU case law and the Advocate General's opinion, and on which legal provisions of the EU VAT Directive is it based on?

Q8. Identify areas of public interest in the context of the EU VAT Directive and clarify their influence on the VAT implementation.

Q9. Identify the different treatments of small-sized enterprises for VAT reasons according to the EU VAT Directive.

Q10. Identify the place of supply of goods for the purposes of applying VAT in accordance with the EU VAT Directive. Classify the place of supply based on the criteria defined in the Directive for each of the categories of goods.

Q11. Identify the transitional VAT provisions or regimes according to the EU VAT Directive: a) specific to each Member State and b) applicable regardless of a specific Member State.

Q12. On which specific occasions does the EU VAT Directive grant Member States the discretion to establish their own deviation for VAT regulations? Identify the occasions associated with specific articles of the EU VAT Directive and provide all the relevant requirements for each one.

Q13. Provide a definition of the term 'legal certainty', exclusively based on the EU case law and the Advocate General's opinion interpreting the EU VAT Directive.

Q14. Provide specific circumstances under which taxable persons can deduct VAT that they have already paid on goods or services they have supplied, according to the EU VAT Directive.

Q15. Provide the basic principles of the EU common VAT system.

Q16. What is the impact of EU customs legislation on VAT legislation, according to the EU VAT Directive?

Q17. Which are the transactions that fall into the scope of the EU VAT Directive?

Q18. Which persons can be considered taxable taking into account the criteria and requirements defined for each category of natural or legal persons in the context of the EU VAT Directive? Identify the specific criteria for each one of the categories of taxable persons in this Directive.

Q19. Which transactions remain out of the scope of the EU VAT Directive?

Each question was given as a prompt to each of the nine LLMs in order to be answered following a zero-shot approach, i.e., using its existing knowledge. To this end, inferences from LLM were obtained via SageMaker or Bedrock Amazon Web Services (AWS). Each LLM was configured with a temperature of 0, a top-p value of 0.9, and a maximum output token limit of 512, based on its level of parameterization. This combination of low temperature and increased top-P allows for creativity and, at the same time, increased reproducibility of the responses [30]. The responses were then anonymized to prevent biases based on the LLMs.

4.2.3.2 Ethical, Legal, and Safety Testing

While accuracy is the most commonly used metric for assessing the truthfulness of LLM responses, involving humans in the evaluation process has also been recognized in literature [23]. In this context, this pilot study involved the legal expert Hellenic State Legal Council to evaluate the anonymized LLM responses with regard to truthfulness. Specifically, the expert assigned to each LLM response a score ranging from one to ten, with one representing the lowest accuracy and ten the highest accuracy. All scores were then statistically analyzed so as to understand the performance of each LLM with regard to trustworthiness. In order to compare the LLMs, the Cohen's Kappa coefficient was calculated to measure agreement between their responses.

The average truthfulness score of each LLM is presented in Figure 18. The heatmap shows a generally high evaluation of responses across different LLMs, with average truthfulness scores falling between 7.4 and 8.6, indicating overall strong performance. The top five LLMs related to truthfulness score are, in descending order, GPT-4 Turbo (8.63, 95% CI: 8.12, 9.15), GPT 4o (8.58, 95% CI: 8.14, 9.01), Llama 3.1 405B (8.45, 95% CI: 7.9, 9), Claude v3 Sonnet (8.16, 95% CI: 7.73, 8.9), and Llama 3.1 70b (8.21, 95% CI: 7.66, 8.76). Llama 3.1 8B (7.37, 95% CI: 6.43, 8.3) and Mistral Large (7.95, 95% CI: 7.3, 8.6) achieved the lowest scores. Among the evaluated LLMs with open weights, Llama 3.1 with the 405 billion parameters provided the most accurate answers.

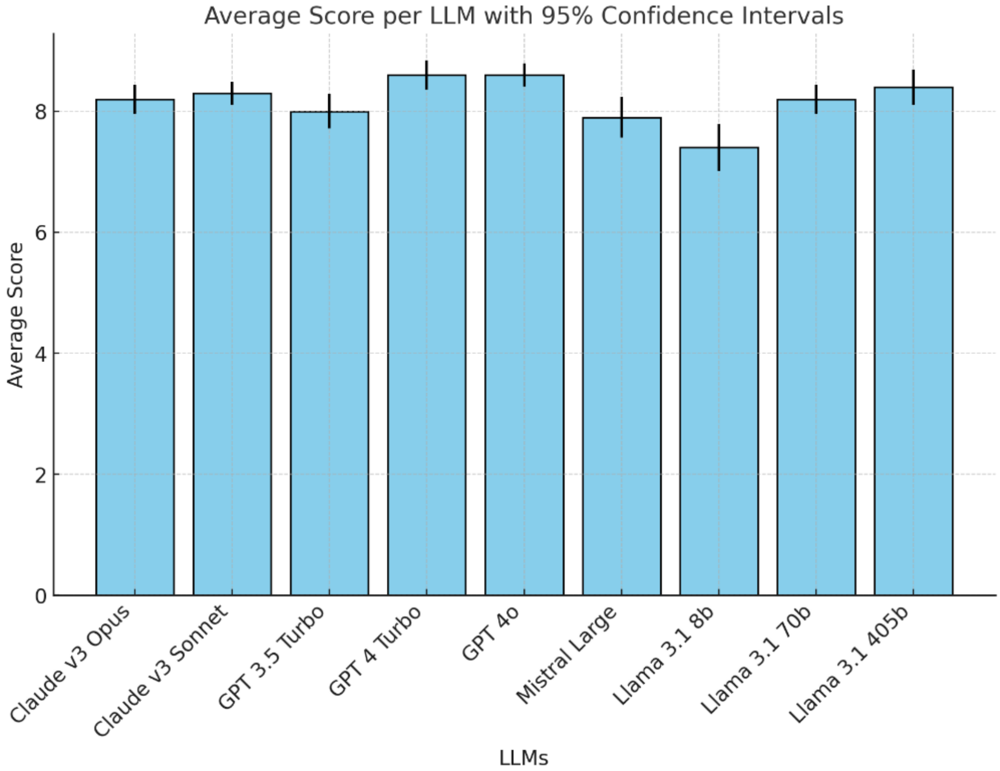
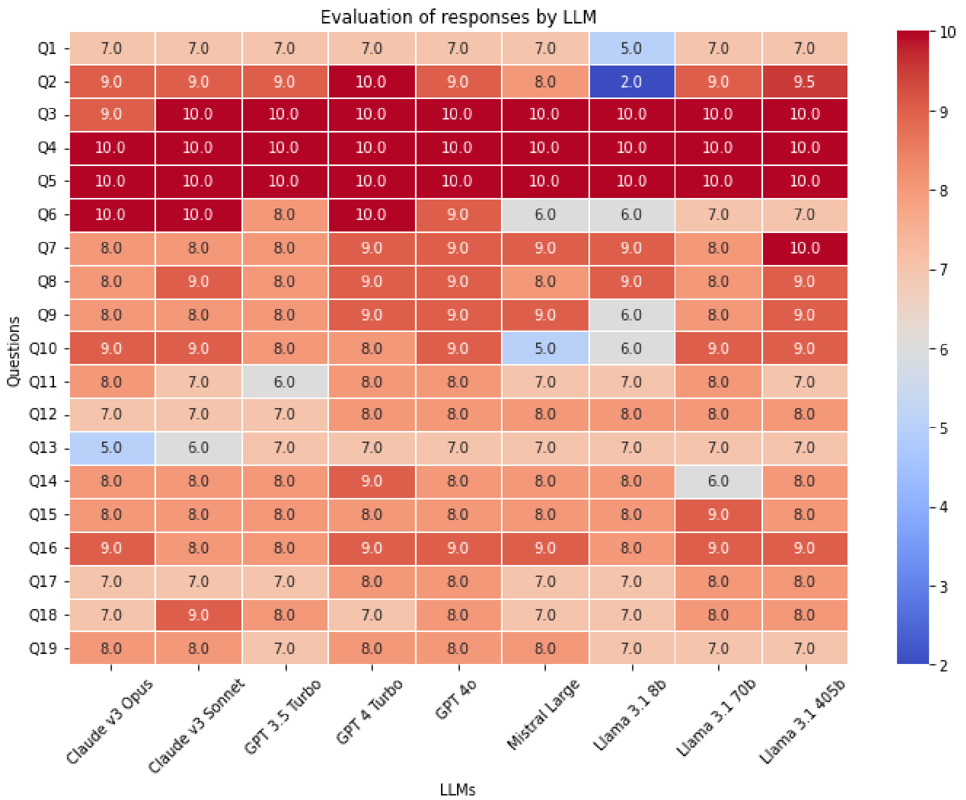


Figure 18 Truthfulness scores for all LLMs (95% confidence interval). The evaluated models include Claude v3 Opus, Claude v3 Sonnet, GPT 3.5 Turbo, GPT 4 Turbo, GPT 4o, Mistral Large, Llama 3.1 8b, Llama 3.1 70b, and Llama 3.1 405b. Results show a high evaluation of responses across different LLMs, indicating overall strong performance [20].

The individual truthfulness scores for all LLM responses are presented in Figure 19. According to the heatmap, some LLMs seem to be more stable in their truthfulness scores, while others show more variability. Specifically, the GPT 4 LLMs are the most stable related to the trustworthiness of their responses, having a few low truthfulness scores in their responses. The Claude-3 variants have in general a good performance in truthfulness, although they have some inconsistencies (e.g., Opus was rated with a 5 on Q13). Llama 3.1 8B faces an increased variation across different questions. However, its responses to some questions achieved very good scores. For example, it was evaluated with a truthfulness score of 10 for questions Q3-Q5 and with 9 for questions Q7 and Q8. This indicates that there is a potential for improvement for the specific model, for example, by fine-tuning it using domain-related data and, hence, enabling it to provide more accurate responses. At the same time, the large LLM of the family (Llama 405b) performs similarly to GPT-4 LLMs, having clearly improved and more stable responses regarding trustworthiness. This indicates that scaling significantly improves results' accuracy.



	Claude v3 Opus	Claude v3 Sonnet	GPT 3.5 Turbo	GPT 4 Turbo	GPT 4o	Mistral Large	Llama 3.1 8b	Llama 3.1 70b	Llama 3.1 405b
Q1	7.0	7.0	7.0	7.0	7.0	7.0	5.0	7.0	7.0
Q2	9.0	9.0	9.0	10.0	9.0	8.0	2.0	9.0	9.5
Q3	9.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Q4	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Q5	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
Q6	10.0	10.0	8.0	10.0	9.0	6.0	6.0	7.0	7.0
Q7	8.0	8.0	8.0	9.0	9.0	9.0	9.0	8.0	10.0
Q8	8.0	9.0	8.0	9.0	9.0	8.0	9.0	8.0	9.0
Q9	8.0	8.0	8.0	9.0	9.0	9.0	6.0	8.0	9.0
Q10	9.0	9.0	8.0	8.0	9.0	5.0	6.0	9.0	9.0
Q11	8.0	7.0	6.0	8.0	8.0	7.0	7.0	8.0	7.0
Q12	7.0	7.0	7.0	8.0	8.0	8.0	8.0	8.0	8.0
Q13	5.0	6.0	7.0	7.0	7.0	7.0	7.0	7.0	7.0
Q14	8.0	8.0	8.0	9.0	8.0	8.0	8.0	6.0	8.0
Q15	8.0	8.0	8.0	8.0	8.0	8.0	8.0	9.0	8.0
Q16	9.0	8.0	8.0	9.0	9.0	9.0	8.0	9.0	9.0
Q17	7.0	7.0	7.0	8.0	8.0	7.0	7.0	8.0	8.0
Q18	7.0	9.0	8.0	7.0	8.0	7.0	7.0	8.0	8.0
Q19	8.0	8.0	7.0	8.0	8.0	8.0	7.0	7.0	7.0

Figure 19 Evaluation of the truthfulness of LLM responses. Each cell represents the score for a given model-question pair, on a scale from 1 (least truthful) to 10 (most truthful) [20].

If we place the focus on the questions, certain questions (e.g., Q2, Q6, Q10, and Q13) show more disagreement across LLMs, suggesting that these may be harder or more ambiguous to answer. Conversely, it can be observed that there are two questions (i.e., Q3 and Q4) that achieved the highest score (10) across all LLMs. This may indicate that these questions are straightforward and, hence, less complex for LLMs to answer. Finally, Q6, which refers to the VAT treatment of farmers, is directly related to a specific exemption case and does not align well with niche tax treatment. This emphasizes the need to go deeper and assess the specific aspects of trustworthiness identified in literature, including robustness, i.e., how well the LLM responds to questions regarding exceptional cases, which is extremely important in tasks like law interpretation.

Thereafter, the agreement of the different LLMs regarding the evaluation of their responses was assessed using the Cohen's Kappa coefficient agreement scores (Figure 20). The score is a value ranging from 0 to 1. Higher values of the Cohen's Kappa coefficient indicate strong agreement between LLMs, while lower values suggest weaker agreement (or, in some cases, even no agreement). Based on the results, the highest score achieved was 0.65. This highest agreement score is observed between GP4o and Llama 3.1 405B, meaning that these two LLMs have a similar behavior in producing accurate responses. This result can be translated as that Llama 3.1 405B has made a lot of progress related to the smaller LLMs of the family and has a good improvement potential. In addition, based on the same results, LLMs from the same family have increased agreement scores when compared to the agreement between LLMs from different families. For example, the second-highest agreement score is 0.64 between GPT 4 Turbo and GPT 4o, which is followed by the agreement between the two Anhropic's LLMs (Claude v3 Opus and Claude v3 Sonnet). All remaining scores were lower, indicating a higher degree of variability in the answers each model selected.

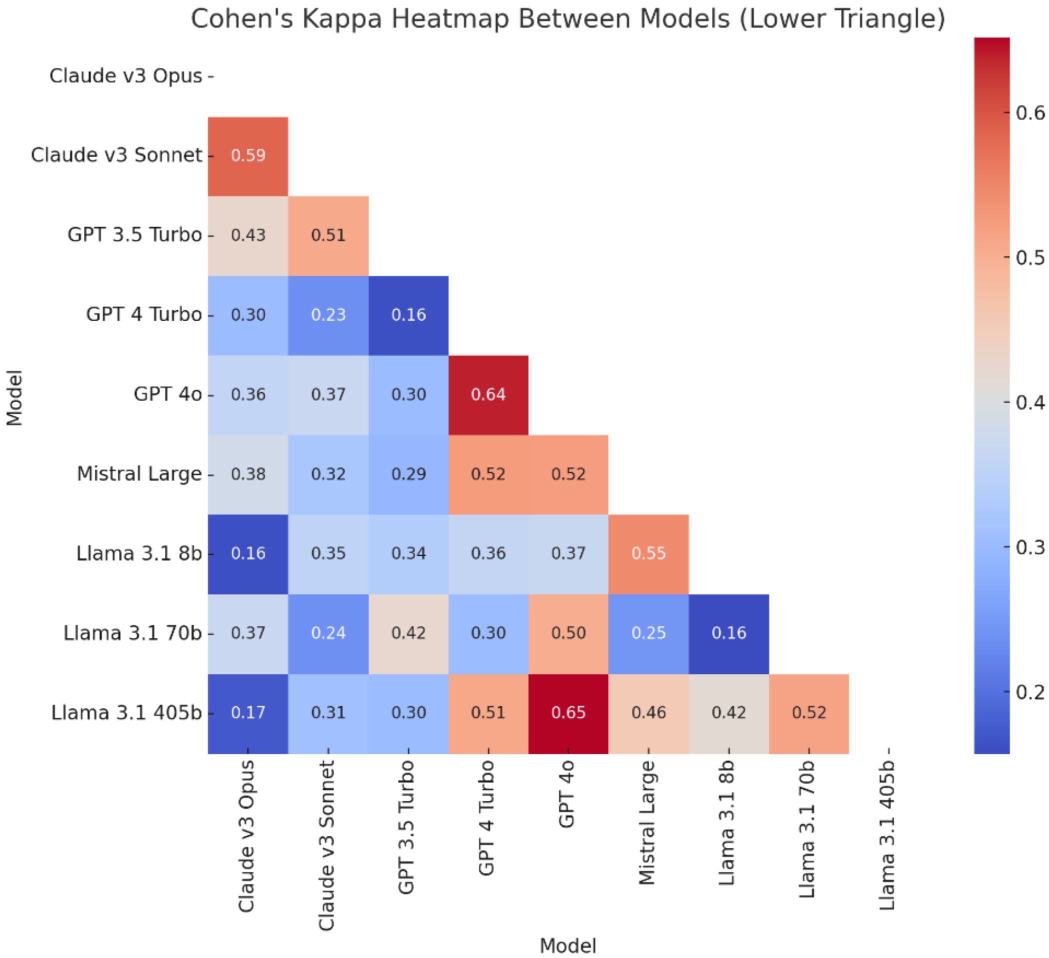


Figure 20 Evaluation of the correlation of LLMs using Cohen's Kappa coefficient agreement scores. The heatmap depicts inter-model agreement values, with higher coefficients indicating greater consistency in model scoring patterns.

The higher agreement between the LLMs from the same families possibly mirrors the anticipated impact of common architectures, training methodologies, and\or similar optimization processes. However, when comparing models from different families, such as, for example, Claude v3 Opus and GPT 4 Turbo or Claude v3 Opus and Mistral Large, the agreement scores drop significantly (0.30 and 0.38, respectively). This indicates that these LLMs are applying distinct ways for producing responses, likely due to variations in training data or optimization goals. Finally, smaller models (e.g., Llama 3.1 8b) show lower alignment with mainstream models, underscoring the impact of model size and training diversity on evaluation consistency. The above findings emphasize the need for creating and using standardized evaluation benchmarks across various LLM families. Since different LLMs are producing varying responses, human oversight and domain-specific fine-tuning remain crucial in applications where consistency is essential, such as the legal domain.

4.2.3.3 User Testing and Feedback Integration

Although this pilot case study does not involve a broad group of end users, a structured user-testing process was implemented through the involvement of the legal expert who evaluated the LLM outputs. Beyond assigning accuracy scores, the expert provided qualitative feedback that offered deeper insight into the models' interpretive behaviour, strengths, and limitations.

Through this expert review, it became evident that the clarity, coherence, and structure of the models' responses varied considerably. Some LLMs demonstrated a strong ability to produce well-organised summaries of legal provisions, yet occasionally failed to capture essential conditions or exceptions embedded in the VAT Directive. In several cases, models produced legally convincing but ultimately incorrect interpretations, a behaviour that presents significant risks in real-world administrative contexts, where outputs may be perceived as authoritative.

The expert noted that increasing question complexity revealed disparities in model performance: while some LLMs maintained logical consistency, others produced contradictory or incomplete reasoning even when temperature settings enforced deterministic responses. These inconsistencies highlighted the limitations of relying on zero-shot legal interpretation without domain-specific fine-tuning. Additionally, proprietary models generally performed better than open-weight models, although no model consistently delivered legally reliable answers across all questions.

The expert's reflections also underscored the necessity of strong human oversight. Despite their impressive linguistic fluency, the models were not capable of replicating the structured, precedent-aware reasoning that characterises legal analysis. The findings suggest that LLMs may serve as supportive tools for information retrieval or preliminary summarisation but are not suitable for unsupervised legal interpretation in public-sector decision-making.

The qualitative insights gathered during this phase contribute directly to understanding the feasibility, risks, and governance needs associated with AI-assisted legal analysis in public administration. They also form a basis for the refinement of requirements in subsequent phases of the AIGOV Transformation and Adoption Framework, particularly regarding explainability standards, safeguards, and the boundaries of responsible use.

4.2.4 Assessment of the KPIs

The tables below present the values for the KPIs related to this PCS.

KPI	Assessment	Status
Legal accuracy score	Truthfulness scores ranged 7.37-8.63 across models. Best performers: GPT-4 Turbo (8.63), GPT-4o (8.58), Llama 3.1 405B (8.45). Lowest: Llama 3.1 8B (7.37).	Measured
Completeness	Expert identified recurring omissions in conditional clauses, exceptions, and special regimes (notably VAT exemptions, discrete treatments such as farmers, place of supply). Completeness declines as question complexity increases.	Measured
Consistency	Cohen's Kappa agreement ranged 0.30-0.65. Highest agreement: GPT-4o with Llama 405B (0.65) and GPT-4 Turbo with GPT-4o (0.64). Significant inconsistencies for smaller or open models, especially Llama 8B.	Measured
Comparative Performance Score	Proprietary models (GPT, Claude) dominate top positions. Open models: Llama 405B approaches GPT-4; Llama 8B and Mistral Large underperform.	Measured

Table 15 Technical Performance KPIs for PCS2

KPI	Assessment	Status
Traceability of sources	Some models (mostly proprietary) cite article numbers; others provide general references or incorrect citations. No model systematically provides verifiable, article-level traceability.	Partially assessed
Clarity of justification	Many models produce fluent but non-transparent legal reasoning. Some explanations mix Directive-based reasoning with assumptions or generic tax logic.	Partially assessed
Absence of hallucinated content	n/a	Defined (future validation)
Ability to reproduce outputs	Temperature was fixed to 0, yet models showed variation in structure, completeness, and emphasis. GPT models most reproducible; others show variability.	Partially assessed

Table 16 Explainability and Transparency KPIs for PCS2

KPI	Definition	Status
Human-oversight viability	n/a	Defined (future validation)

Compliance and risk alignment with AI Act categories	n/a	Defined (future validation)
Public value contribution	n/a	Defined (future validation)
Multilingual scalability potential	n/a	Defined (future validation)
Governance and accountability readiness	n/a	Defined (future validation)

Table 17 Responsible AI and Public-Sector Applicability KPIs for PCS2

4.3 Pilot Case study 3: A Large Language Model Agent Based Legal Assistant for Governance Applications.

4.3.1 Phase 1: Assess Readiness and Context

4.3.1.1 Strategic context, challenges, and objective

As AI continues to advance rapidly, its accompanying tools and services are reaching greater levels of maturity, progressively infiltrating the public sector [129]. Among the array of AI applications are sophisticated tools, for instance, to facilitate law-making, enable oversight mechanisms and enhance political discourse. These are aspects of governance that fall under the EU framework of better regulation [24] that is gradually adopted by individual member states [9]. AI-based technological innovations have the capacity to transform governance institutions fundamentally, reshaping the way public services are generated and delivered [15]. A relevant research and development agenda for the introduction of such tools in the parliamentary workspace has been already outlined and may partially be used for the elaboration of an AI roadmap for governance applications [26]. Accordingly, several of the political, administrative or scientific tasks necessary to modernize governance can be linked to natural language processing (NLP) related technologies. Specifically, over the last few years, large language models have revolutionised the NLP landscape, having been established as the premier approach for a plethora of related tasks, including chatbots and virtual assistants.

The subset of machine learning models specialized in tasks that concern natural language are referred to as language models. Ever since the inception of the revolutionary attention mechanism [44], transformer based language models have proved to be efficient in understanding and generating natural language, scaling well with the volume of their training data. This breakthrough has led to the creation of large language models (LLMs) that consistently surpass state of the art performance metrics, and showcase emergent capabilities [45]. In operational contexts such as these, the potential applications of LLMs are readily apparent. Already, numerous use cases and relevant studies emanating from both executive and legislative branches are evident, and public sector bodies are already beginning to deploy generative LLMs to enhance their services; one such case being that of the Greek Government deploying “mAIgov”, a chatbot powered by OpenAI’s generative LLMs, that has been trained using open data sourced from various public entity websites.

Governance operates within its distinct linguistic domain, primarily characterized by legal language. Given its specialized nature, legal language demands dedicated resources for comprehensive analysis and application. Such resources encompass legal corpora, lexical databases, and guidelines governing grammar and style, as well as exhaustive references for acronyms, organizations, and abbreviations. In the EU context, legal resources can be openly accessed via the Publications Office of the EU, which runs the EUR-Lex platform.

Despite the opportunities, significant challenges persist. Legal interpretation carries high stakes: errors, hallucinations, or omissions could lead to misinformed policy advice, flawed administrative decisions, or diminished trust in public institutions. Moreover, LLMs trained primarily on general-purpose data often fail to grasp nuanced legal reasoning, rely on probabilistic text generation rather than deterministic logic, and exhibit variability that raises concerns regarding reliability, accountability, explainability, and compliance with emerging regulatory frameworks, including the EU AI Act.

Against this backdrop, the objective of this pilot case study is to explore how advanced LLM-based legal assistants can be integrated into governance applications in a manner that is responsible, evidence-based, and aligned with public-sector values. The pilot evaluates the capabilities of two, state-of-the-art LLMs (GPT3.5 and GPT4) combined with Retrieval-Augmented Generation (RAG) and agent-based orchestration techniques, to support queries related to EU law. A tailored legal corpus focused on the General Data Protection Regulation (GDPR) and relevant case law is used to ground the models in authoritative sources.

The aim is not to automate legal interpretation but to examine whether such systems can enhance the analytical capacity of public servants, improve access to relevant legal information, and support governance functions while maintaining transparency, explainability, and human oversight. This strategic positioning ensures that the pilot contributes to public value by exploring practical, trustworthy, and scalable avenues for integrating LLMs into future governance workflows.

4.3.1.2 Stakeholders

The stakeholders involved in this pilot case study include institutional actors, technical stakeholders, domain experts, and external beneficiaries who would either interact with or be affected by the deployment of an LLM-based legal assistant in governance contexts. As with the other pilot case studies, the stakeholder mapping follows the internal and external stakeholder categories established in Deliverable D1.2 of the AIGOV Ecosystem.

Internal Stakeholders include:

- *Public authorities and governance institutions*. These include ministries, regulatory agencies, supervisory bodies, parliamentary services, and public organisations responsible for drafting legislation, interpreting regulatory frameworks, conducting oversight, or supporting legal compliance. Such institutions would use LLM-based assistants to enhance legal research, accelerate document analysis, and support regulatory processes.
- *Public servants and legal professionals* within the administration. These are legal advisors, legislative drafters, policy analysts, GDPR officers, auditors, and administrative staff who routinely engage with legal documents. They represent the primary user group for an LLM-based legal assistant, relying on the system for rapid retrieval, summarisation, or contextual interpretation of legal provisions.
- *Regulatory and compliance authorities.* Bodies responsible for ensuring adherence to the EU AI Act, GDPR, and sector-specific regulations play a key role in assessing whether the system meets legal and ethical requirements. They would also oversee safeguards regarding accuracy, explainability, accountability, and risk management.
- *IT departments and technical units.* This group encompasses public-sector data engineers, knowledge-graph specialists, AI practitioners, cybersecurity teams, and digital service designers involved in integrating the legal assistant with existing information systems. They ensure technical feasibility, system maintenance, data governance, and secure deployment.

External Stakeholders include:

- *Legal experts and academic researchers.* Scholars in EU law, digital governance, and computational legal studies contribute domain knowledge, evaluate system performance, and identify ethical, methodological, or interpretative risks. Their role is essential for validating accuracy, ensuring doctrinal correctness, and advancing legal-AI research.
- *Citizens and businesses.* Although not direct users in this pilot, citizens and private-sector organisations may ultimately benefit from improved legal clarity, more accessible public information, and faster or more consistent administrative decisions supported by LLM tools.
- *Civil society organisations and advocacy groups.* Stakeholders such as digital rights organisations, privacy advocates, and transparency watchdogs have an interest in ensuring that AI-enabled legal interpretation remains accountable, explainable, and aligned with fundamental rights. They are key actors for monitoring fairness and preventing potential misuse.
- *Providers of LLM technologies.* Private-sector AI developers (e.g., OpenAI) and operators of cloud environments (e.g., AWS) serve as technology providers for the models and infrastructure used in the pilot. While not directly involved in governance

decisions, they influence the capabilities and limitations of the systems under evaluation.

4.3.1.3 Data Readiness

Data readiness for this pilot case study concerns the availability, quality, accessibility, and suitability of the legal and contextual data required to support an LLM-based assistant designed for governance and legal interpretation tasks. In line with the AIGOV Government Data Value Cycle and the readiness assessment criteria defined in Phase 1 of the Transformation and Adoption Framework, this subsection evaluates whether the necessary data resources are sufficiently prepared for exploratory prototyping.

The core dataset used in this case study is a curated legal corpus focusing on the General Data Protection Regulation (GDPR) and associated EU legislative and case-law materials. These legal texts were retrieved from the Publications Office of the European Union through the EUR-Lex platform, which provides machine-readable, multilingual, and versioned access to EU legislation, case law, preparatory acts, and legal summaries. As such, the primary source meets high standards of authenticity, transparency, and reliability, a prerequisite for any AI tool supporting legal interpretation.

The corpus includes:

- the full text of the GDPR (Regulation (EU) 2016/679),
- relevant recitals, definitions, and cross-referenced provisions,
- associated case law from the Court of Justice of the European Union (CJEU),
- guidance documents and interpretative communications from EU supervisory authorities, where publicly available.

Prior to integration into the LLM pipeline, documents were processed through a Retrieval-Augmented Generation (RAG) architecture, requiring segmentation, metadata structuring, and embedding generation. These steps ensure that the assistant can retrieve and reason over authoritative legal passages rather than relying solely on model-internal knowledge. The preparation process therefore supports traceability and reduces the risk of hallucinated interpretations. both essential considerations under responsible AI principles.

No personal data were included in the corpus, and therefore GDPR compliance obligations were minimal. However, copyright and licensing constraints were reviewed, ensuring lawful reuse of all legislative materials. The EUR-Lex datasets used in this pilot are openly licensed and intended for public re-use, satisfying the requirements of the EU Open Data Directive.

While the curated corpus is adequate for a research prototype, several data limitations remain relevant for future operational deployment. First, although legislative texts are authoritative, legal interpretation is often influenced by national case law, administrative guidelines, and domain-specific contextual materials that were not included in this pilot.

Second, multilingual versions of the documents exist but were not integrated into the experiment, limiting cross-linguistic robustness. Third, version control and consolidation of amendments, particularly in the case of dynamic regulatory domains, would require automated pipelines in an operational setting.

4.3.1.4 Governance, Ethics, and Legal Compliance

The integration of LLM-based legal assistants into governance applications requires rigorous attention to ethical, legal, and governance considerations, particularly because such tools interact directly with authoritative legal materials and may influence public-sector decision-making. This pilot case study operates within a controlled research environment. However, the assessment of governance and compliance follows the criteria defined in the AIGOV Framework for Trustworthy, Fair, and Accountable AI, as well as the requirements stemming from EU legislation, including the AI Act and GDPR.

From a legal compliance perspective, the pilot handles exclusively publicly available legislative texts, jurisprudence, and regulatory documents sourced from EUR-Lex and related EU repositories. No personal data were processed, eliminating the need for GDPR-related safeguards such as data minimisation, lawful basis assessment, or data-subject rights management. This positions the prototype within the low-risk category of the EU AI Act, as it does not generate legal effects or make binding assessments about individuals. Nevertheless, if the system were eventually integrated into advisory workflows or administrative procedures, its risk classification would need to be reassessed, particularly regarding potential impacts on rights, obligations, or resource allocation.

Ethically, the pilot addresses core elements of responsible AI development: transparency, explainability, fairness, and human oversight. Since generative LLMs are by nature probabilistic and prone to hallucinations, special attention was given to mitigating risks related to inaccurate or misleading legal interpretations. This is achieved through the use of Retrieval-Augmented Generation (RAG) and agent-based orchestration, which ground the model's responses in authoritative legal sources rather than relying solely on internal learned representations. All system outputs remain subject to expert human review, an essential safeguard in legal contexts where interpretive accuracy is paramount.

Accountability mechanisms were also introduced conceptually. In line with AIGOV Pillar 1 on Transparency & Accountability, the system is framed as a decision-support tool, not an automated decision-making system. Responsibility for legal interpretation remains with qualified human experts, and the AI system is intended only to enhance efficiency, retrieval capability, and analytical support. The pilot does not attempt to automate or replace legal judgment.

Additionally, transparency obligations were considered throughout system design. Documentation of data sources, model prompts, retrieval pipelines, and system limitations is maintained to ensure traceability and auditability. This aligns with requirements for technical documentation and explainability under both the AI Act and the AIGOV Framework.

From a fairness and bias perspective, risks stem primarily from model-internal biases of the underlying LLMs (GPT-3.5 and GPT-4) and from the uneven distribution of interpretative resources across topics. While the legal corpus used is authoritative and standardised, the underlying models may inadvertently introduce stylistic, contextual, or inferential biases. To limit these risks, the pilot restricts responses through grounding and employs expert evaluation to assess reliability.

Finally, future real-world adoption would require formal governance structures, including:

- clear delineation of roles and responsibilities between developers, legal experts, and administrative users,
- continuous monitoring of system outputs,
- version control and documentation of legal source updates,
- compliance with procurement and transparency obligations for AI systems in the public sector,
- human-in-the-loop escalation mechanisms for uncertain, ambiguous, or high-risk outputs.

4.3.1.5 Organisational Capability and Capacity

The organisational capability and capacity assessment for PCS3 examines the extent to which public administrations, governance institutions, and supporting technical units are prepared to adopt and operate an LLM-based legal assistant within real-world governance processes. Although this pilot remains exploratory, a readiness analysis is essential to understand the structural, institutional, and human preconditions for the integration of such technologies into public-sector workflows.

Technical and Data Capabilities. Deploying an LLM-based legal assistant requires capabilities that extend beyond standard IT operations. Public administrations would need specialised expertise in natural language processing (NLP), legal informatics, retrieval-augmented generation (RAG), and AI model orchestration. While some larger public organisations may possess in-house technical units capable of supporting database management, cybersecurity, and application hosting, few currently maintain internal competencies for configuring, fine-tuning, monitoring, and governing advanced LLM pipelines. Furthermore, the successful integration of a legal assistant depends on the ability to construct and maintain a high-quality, version-controlled corpus of legal materials, alongside metadata structures and retrieval mechanisms. These competencies are not yet standardised across most European governance institutions, indicating that technical readiness is partial and uneven.

Legal and Domain Expertise. Because LLM-based legal assistants interact directly with complex legal documents and may influence interpretative tasks, organisational capacity must include strong domain expertise. Legal advisors, legislative drafters, regulatory analysts, GDPR officers, and administrative jurists serve as essential human supervisors of AI-assisted outputs. Their role is not merely evaluative but also methodological, requiring the ability to assess the reliability, doctrinal correctness, and contextual appropriateness of AI-generated interpretations. This implies a dual literacy requirement: legal professionals must acquire foundational AI literacy, while technical teams must develop an understanding of legal reasoning and interpretative practices. At present, such interdisciplinary capability is emerging but not yet widespread within most public administrations.

Governance, Oversight, and Risk Management Structures. The introduction of LLM-based legal assistants necessitates institutional mechanisms for oversight, accountability, and compliance with evolving regulatory frameworks, particularly the EU AI Act. Public bodies would need to establish clear roles and responsibilities for AI supervision, including procedures for documenting model behaviour, validating outputs, managing risk, and escalating uncertain or ambiguous results. Many public organisations currently lack formal AI governance frameworks or internal units dedicated to AI oversight, which may hinder their ability to deploy high-risk or sensitive AI systems. Establishing such structures is therefore a prerequisite for responsible adoption.

Human-in-the-Loop Capacities. A critical aspect of organisational readiness concerns the ability to integrate human oversight into the operational workflow. Legal experts must be equipped not only to interpret AI-generated outputs but also to challenge, correct, and contextualise them within governance processes. This requires institutional recognition that AI tools cannot replace legal judgment and must operate strictly as decision-support systems. Organisational processes must also ensure that staff have adequate time, training, and procedural authority to review AI outputs rigorously—conditions that are not automatically met in resource-constrained public-sector environments.

Infrastructure and Security Capacity. Operational deployment of PCS3's legal assistant would require secure, scalable, and compliant computing infrastructure. This includes:

- secure data storage environments for legal corpora and model outputs,
- computational capacity for real-time retrieval and model inference,
- mechanisms for logging and audit trails,
- cybersecurity protections aligned with EU and national standards, and
- integration capability with existing legal databases and administrative systems.

While cloud-based AI services can provide some of these capabilities, public-sector bodies must also ensure compliance with procurement rules, data localisation requirements, and security frameworks. Infrastructure readiness is therefore variable and requires strategic planning.

Change Management and Organisational Culture. Implementing AI-assisted legal interpretation involves cultural and procedural transitions within public administrations. Staff must be open to adopting new tools, workflows must be adapted to incorporate iterative human-AI collaboration, and leadership must support innovation while maintaining a cautious approach to risk. Resistance may emerge from concerns about reliability, accountability, or potential deskilling of legal professionals. Organisational change management, including training programmes, communication strategies, and participatory design processes, is thus essential for sustainable adoption.

4.3.2 Phase 2: Design Ethical and Value-Aligned Use Cases

Phase 2 focuses on translating the strategic context and readiness assessment into a coherent and responsible AI use case. In this pilot, the goal is to explore how Large Language Model (LLM)-based legal assistants can support governance processes, with a particular focus on legal interpretation, policy analysis, and administrative decision support. The phase follows the three activities defined in the AIGOV Transformation and Adoption Framework:

- Problem and service redesign definition (Section 4.3.2.1)
- Responsible AI use case design (Section 4.3.2.2)
- Success metrics and impact definition (Section 4.3.2.3)

4.3.2.1 Problem and service redesign definition

The second case study was designed as a single exploratory case study to examine how LLM-based systems can be used as legal assistants to support policy and decision makers as well as administrators. A single exploratory case study is a useful design to gain insights about phenomena that are thus far understudied or not explored at all, and to construct a new theory or generate propositions about their understanding [50]. In doing so, the research team involved a group of legal professionals, policy making, and technical experts, whose expertise covers every facet of the experimental design, from its inception to its implementation and, finally, its evaluation.

A domain specific legal topic was selected, with relevance to all branches of governance. For this, Regulation (EU) 2016/679 entitled General Data Protection Regulation (GDPR) was considered to be a contemporary and relatively mature legal topic fit for the purpose of this research. The GDPR dictates the responsible handling of personal data by organizations and companies. It is applicable to all EU member states, establishing uniform regulations to safeguard the rights and privacy of both businesses and citizens. Some of its essential contents stipulate transparent data utilization, lawful processing, protection of individual rights, and the obligation to report data breaches promptly [41].

Legal professionals defined a reference set of twelve legal questions related to GDPR that are considered to be of high value in public administration or policy/law making. The questions were of varying levels of difficulty and were organized accordingly into three groups of four questions each.

A well-defined corpus with legal documents, directly referring to the GDPR was extracted from EUR-Lex[2]. The EUR-Lex repository is a comprehensive database that provides access to EU law. It includes EU treaties, legislative acts, international agreements, and preparatory documents. Additionally, EUR-Lex hosts case law from EU courts, national court decisions related to EU law matters, and the JURE database compiling cases on judicial cooperation. This valuable resource serves legal professionals, researchers, and anyone seeking information on EU law. Using the expert search feature of Eur-Lex, the search for legal texts was limited to the terms “General Data Protection Regulation” and “GDPR”, from the period spanning from 1.1.2015 to 15.2.2024. Only texts in English were extracted.

A proof of concept implementation was then created. It utilized the generative large language models GPT3.5 and GPT4, as well as the current state of the art embedding model (text-embedding-3-large) offered by OpenAI. A combination of prompt engineering, retrieval augmented generation architectures and agent based systems was explored in order to create a system capable of harnessing the advanced reasoning and generative capacities of the LLMs to effectively answer the questions created by the experts.

As an initial evaluation option, a qualitative approach was considered. The set of reference legal questions was posed to the system one by one. The researcher group then engaged in qualitative discussions of the outcome from a legal perspective to evaluate the quality of the responses. This internal approach allows for high flexibility and rapid development across multiple iteration cycles.

4.3.2.2 Responsible AI Use Case Design

Building on the problem definition, the experimental design was constructed to evaluate whether an LLM-based legal assistant can support governance tasks responsibly and in alignment with the AIGOV Framework for Trustworthy, Fair, and Accountable AI.

The use case incorporates the following design principles:

1. Explainability and Traceability Requirements. Because legal interpretation is highly sensitive and must remain auditable, the system was designed using Retrieval-Augmented Generation (RAG) and agent-based reasoning pipelines. These approaches enable the model to cite specific legal sources for every answer, ensuring that public servants can trace the information back to authoritative legal documents. Prompt engineering was used to encourage explicit referencing and avoid hallucinations.

[2] https://eur-lex.europa.eu/

2. Ethical and Legal Safeguards. Although GDPR documents themselves do not contain personal data, the use of LLMs in legal interpretation raises concerns about reliability, bias, and over-dependence on machine-generated text. Safeguards built into the design include:
    - strictly zero-shot prompting to avoid contamination from prior cases or fine-tuning;
    - anonymised evaluation of outputs to mitigate model- or vendor-related bias;
    - qualitative legal review by experts to detect misleading or incorrect interpretations;
    - explicit prohibition of automated decision-making, ensuring that humans retain interpretive authority.
3. Human Oversight and Accountability. The system operates exclusively as a decision-support tool. Legal experts are responsible for validating and interpreting outputs, and the model's role is limited to assistance, not determination. The design anticipates future requirements such as audit logs, model version tracking, and escalation procedures for ambiguous or inconsistent responses.
4. Corpus Quality, Transparency, and Documentation. The legal corpus extracted from EUR-Lex is openly available, citable, and maintained by the Publications Office of the EU. Its use ensures transparency regarding source material and improves the reproducibility of experiments. Detailed documentation of query processes, system prompts, hyperparameters, and evaluation protocols was maintained as part of responsible research practice.
5. Inclusion, Usability, and Accessibility. Although the pilot focuses on English-language legal texts, the architecture (RAG and agent-based reasoning) can be extended to multilingual environments, supporting broader EU governance needs. The prototype was also evaluated qualitatively for clarity, structure, and ease of interpretation by legal practitioners.

Together, these design choices ensure that the use case is grounded in responsible AI principles and is suitable for controlled testing in governance contexts where accuracy, explainability, and accountability are essential.

4.3.2.3 Success Metrics and Impact Definition

To evaluate the feasibility and public-sector relevance of the LLM-based legal assistant, a structured set of success indicators (KPIs) was defined. These metrics reflect both technical performance and the responsible AI criteria required for governance applications.

Three categories of KPIs are proposed, namely Technical Performance (Table 18), Explainability and Transparency (Table 19), and Responsible AI and Public-Sector Applicability (Table 20).

KPI	Definition	Status
Legal correctness score	Percentage of correctly classified data zones	Partially Assessed
Consistency and reproducibility	Across repeated prompts	Measured
Agreement metrics	To compare LLM outputs and detect interpretive variance	Defined (future validation)
Source citation accuracy	Whether retrieved legal text is relevant and correctly applied	Partially Assessed

Table 18 Technical Performance KPIs for PCS3

KPI	Definition	Status
Traceability of reasoning	Measured through citation completeness and internal coherence	Partially assessed
Clarity of responses	Judged by legal experts	Partially assessed
Auditability	Demonstrated through documented retrieval steps and agent reasoning paths	Defined (future validation)

Table 19 Explainability and Transparency KPIs for PCS3

KPI	Definition	Status
Alignment with legal constraints	Including avoidance of hallucinations or misleading interpretations	Defined (future validation)
Human oversight readiness	Assessed through ease of review and ability for experts to contest or refine answers	Defined (future validation)
Integration potential	Suitability for embedding into legal workflows	Partially assessed
Risk indicators	Including detection of ambiguous or low-confidence answers requiring escalation	Defined (future validation)

Table 20 Responsible AI and Public-Sector Applicability KPIs for PCS3

Not all KPIs can be fully quantified in this pilot phase, as the system is not deployed in an operational environment. However, the framework lays the foundation for future testing, monitoring, and refinement under Phases 3–5 of the AIGOV Transformation and Adoption Framework.

4.3.3 Phase 3: Build and Test

Phase 3 focuses on transforming the designed use case into a functional prototype and systematically evaluating its behaviour, limitations, and suitability for governance applications. Given that this pilot is exploratory and not intended for operational deployment, the objective of Phase 3 is to assess feasibility, identify risks, and understand how Large Language Model (LLM)-based legal assistants behave when applied to real legal questions using authoritative EU legal sources.

Consistent with the AIGOV Transformation and Adoption Framework, Phase 3 included:

(1) Prototype development and technical implementation,

(2) Ethical, legal, and safety testing, and

(3) User testing and feedback integration.

The methodology used in this pilot follows these three activities and is summarised below.

4.3.3.1 Prototype Development and Technical Implementation

4.3.3.1.1 Reference Legal Questions

A set of 12 questions related to GDPR was defined by legal professionals that have expertise in public administration and policy making. The questions were categorized into three levels of progressive difficulty, namely beginner, intermediate, and expert. Each level comprises four questions. It is important to note that the difficulty levels were assessed by human legal experts and do not necessarily reflect the system's understanding of the questions or the quality of its responses. Furthermore, even though legal experts were not expected to have fully delved in the technicalities of an LLM, the language used was clear and unambiguous, avoiding excessive legal jargon and context sensitivity. This ensured that the LLM could perform at its full potential producing useful and accurate responses.

Beginner questions (Q1-Q4) aim at extracting basic information already stated in the pertinent legal framework of GDPR. It is expected that the system generates responses that entail parts or entire provisions, eventually with the respective legal reference (Article/paragraph/etc.). There is no progressive difficulty in this set of questions.

i. What are the primary objective and the EU regulatory framework of the General Data Protection Regulation (GDPR)?
ii. How does the GDPR define "consent" in the context of data processing, and what are the requirements for obtaining valid consent?
iii. Describe the role, responsibilities, and qualifications of a Data Protection Officer (DPO) under the GDPR.
iv. Specify the framework for the protection of childs' rights under GDPR and provide the relevant provisions.

The intermediate set of questions (Q5-Q8) is more demanding and might require multiple levels of processing. There is a progressive difficulty (for a human being) in producing the requested results. The first two questions, Q5 & Q6 ask for numerical results. In the case of Q7, a further analysis is required for identifying court rulings. Similarly, Q8 asks for the identification of preliminary rulings, before classifying them further.

v. How many requests for preliminary rulings have been submitted to the Court during the period of reference?
vi. For the requests for preliminary rulings, for which a court ruling has been issued, calculate the amount of time in days between date of submission and date of issuance, and present the average time for Court response.
vii. For the entire number of preliminary rulings to the EU Court, identify the cases where a court ruling has been issued.
viii. For the period from 2015-2024, identify and classify the requests for preliminary rulings to the EU Court using the following criteria: a. member state courts, b. number of requests per year.

Q9 & Q10 ask more a more advanced processing focusing on the broader context of the legal mater. Q11 goes a step further as asks the system to combine information for identifying potential risks based on the similarity and the number of requests for preliminary rulings. Q12 is a “tricky question” provided that the requested relations is not explicitly mentioned in the GDPR but constitutes basic EU procedural law, according to which whenever national courts have doubts as regards the interpretation of EU legislation or that of domestic legislation as regards its consistency with EU legislation, they may submit question to the EU Court of Justice, which subsequently provides mandatory interpretation for all member states.

ix. For the requests for preliminary rulings submitted to the Court by Member State courts, identify overlapping areas as regards the interpretation of specific articles of GDPR, and present the most frequently addressed articles. Classify the findings by article GDPR.
x. From the requests for preliminary rulings submitted to the Court by Member State courts, derive the key principles that govern the processing of personal data.
xi. Taking into account the GDPR and the submitted requests for preliminary rulings, identify potential risks factors regarding data protection rights.
xii. What is the relation between the EU Court of Justice and the national courts as regards the interpretation of GDPR provisions.

4.3.3.1.2 Proof of Concept Implementation

The initial step of the implementation of the proof of concept system pertained to the collection of the data. Upon execution of the expert query, 428 documents were returned, whose metadata were downloaded though the platform, while the individual documents were downloaded, as HTML files, based on their unique CELEX number provided in the metadata file. The metadata that were included were: Title, CELEX number, ELI, Form, EUROVOC descriptor, ECLI identifier, Subject matter, Case-Law directory code (after Lisbon), Type of procedure, Collection of the document, Date of document, Date of publication, Author, Date of effect, Applicant/Appellant, Defendant/Other parties to the proceedings, Number of pages, and Publication Reference.

According to the EUR-Lex classification, this reference legal corpus included EU law and case-law, i.e. EU consolidated legal texts (157), EU case-law (267), as well as national case-law (4). Classified by the type of act, the corpus contained: consolidated legal texts (157), judicial information (139), Opinions of the Advocate General (64), judgments (57), orders (6), decisions by national courts in the field of European Union law (4), and information (1).

In order to answer the aforementioned twelve questions that were posed, two distinct systems were developed. The difference in the approaches used was due to the nature of the initial queries. The questions that belonged in the beginner group (Q1-Q4) required information that could be found only in the contents of the documents in order to be answered effectively. However, questions in the Intermediate and Expert question groups (Q5-Q12) did not only require knowledge extracted from the contents of the documents but also information that pertained to the document type, title, and dates associated with it. Both systems were implemented in Python, partly through the Langchain library, while the LLMs used were models gpt-3.5-turbo-0125 and gpt-4-0125-preview of OpenAI, featuring a context token window size of 16,385 and 128,000 tokens respectively, as well as the text-embedding-3-large embedding model that had an output dimension of 3,072.

RAG-Based System

Beginner questions were deemed to be able to be answered just through examination of the contents of the General Data Protection Regulation (CELEX number: 32016R0679), thus the RAG system only considered this document for the purpose. The document was split in 88 continuous chunks were each chunk corresponded to approximately 2000 tokens, with a 200 token size overlap between chunks, applied to ensure seamless continuation of meaning. Token count was estimated for model gpt-3.5-turbo-0125.

The RAG-based system was comprised of four components. The embedding model, the vector database, the retriever and the generative LLM. For each chunk of the original document, embeddings were calculated using the textembedding-3-large model. The vector representations of the chunks were stored in a ChromaDB vectorstore. Upon query execution, the query embeddings - created in the same manner - and the stored chunk embeddings were used to retrieve the four most similar chunks to the query based on the cosine similarity of their embedding vectors. All four chunks along with the original query were subsequently passed to the LLM, model gpt-3.5-turbo-0125, in order to retrieve the answer. The prompt was formulated as follows:

“Use the following pieces of context to answer the question at the end.
If you don’t know the answer, just say that you don’t know, don’t try to make

up an answer.
Context: {context}
Question: {question}
Answer:"
Where {context} and {question} were populated with the retrieved content and the query respectively. The prompt choice ensured that the model would refrain from hallucinating, and that the answers would be concise and accurate based on the provided context. The answers generated are quoted below.

Answer to Q1: The primary objective of the General Data Protection Regulation (GDPR) is to protect the fundamental rights and freedoms of natural persons with regard to the processing of personal data. The EU regulatory framework of the GDPR lays down rules relating to the protection of personal data and the free movement of such data within the Union.

Answer to Q2. The GDPR defines "consent" as a clear indication of the data subject's agreement to the processing of their personal data. In order to obtain valid consent, the controller must be able to demonstrate that the data subject has given consent, the request for consent must be presented in a distinguishable manner, the data subject must have the right to withdraw consent at any time, and consent must be freely given without any imbalance of power between the data subject and the controller. Additionally, the data subject must be informed of the identity of the controller and the purposes of the processing for which the personal data are intended. Consent should not be considered freely given if there is a clear imbalance between the data subject and the controller, or if the performance of a contract is dependent on consent.

Answer to Q3. The Data Protection Officer (DPO) is responsible for informing and advising the controller or processor and employees on their obligations under the GDPR and other data protection laws. They monitor compliance with the GDPR, provide advice on data protection impact assessments, cooperate with the supervisory authority, act as a contact point for the supervisory authority, and ensure that data subjects can contact them regarding their personal data. The DPO must have expert knowledge of data protection laws and practices, be easily accessible, and have the ability to fulfill their tasks effectively. They must also be involved in all issues related to data protection, have access to resources and personal data, and report directly to the highest management level of the controller or processor. The DPO must maintain confidentiality, avoid conflicts of interest, and have due regard to the risks associated with processing operations. They may be a staff member or work on a service contract basis.

Answer to Q4. The framework for the protection of children's rights under GDPR includes specific protections for children's personal data, especially in relation to marketing, creating profiles, and collecting data through services directly offered to children. The consent of a parent or guardian is not always necessary, particularly in the context of preventive or counseling services offered directly to a child. This is outlined in Article 8 of the GDPR.

Agent-Based System

Due to the complex nature of questions 5 through 12, an agent-based system was implemented. The implementation involved two agent sub-systems, one semiautonomous and one fully autonomous. The level of autonomy appointed to each sub-system correlated to the flexibility required for each task. Intermediate and expert level questions required both metadata information about each document as well as information contained within the document itself. For this reason it was necessary to inspect the contents of each document, extract the information -not present in the metadata file- necessary to answer the question posed, augment the metadata file with the extracted and structured information and finally have the augmented metadata file contents analyzed by the LLM.

Initially, each of questions Q5–Q12, along with the column names already present in the metadata file were passed to model gpt-4-0125-preview, that was tasked with identifying and creating sub-questions for the parts of the initial question that could not be answered directly through the metadata file. For each of the sub-questions, the LLM was also asked to provide an associated data type; this ensured to some degree a higher level of homogeneity of the LLM's responses in the next step.

Thereafter, all documents were passed to the LLM, along with orders to generate responses and structure them, for each of the sub-questions of the first step. In this second step model gpt-3.5-turbo-0125 was used, and in cases where the context window size was exceeded, chunking was employed. The structured generated responses were combined with the original metadata file.

In the second stage, a fully autonomous agent based on the gpt-4-0125- preview model was employed to analyze and generate responses for the original questions, based on the augmented metadata file. The answers generated by the autonomous agent are provided below:

Answer to Q5. There have been 209 requests for preliminary rulings submitted to the Court during the period of reference.

Answer to Q6. The average amount of time in days between the date of submission and the date of issuance for requests for preliminary rulings where a court ruling has been issued is approximately 335 days.

Answer to Q7. There are 78 cases where a court ruling has been issued for a request for a preliminary ruling to the EU Court.

Answer to Q8. For the period from 2015–2024, the requests for preliminary rulings to the EU Court from member state courts, classified by member state courts and the number of requests per year, are as follows:

- In 2016, the Corte suprema di cassazione (Court of Cassation, Italy) made 1 request.
- In 2017, Latvia, Slovak Republic each made 1 request, and the Supreme Court (Ireland) made 2 requests.
- In 2018, Belgium made 1 request.
- In 2024, Austria made 2 requests, Belgium 1 request, Bulgaria 1 request, Germany 6 requests, and Lithuania 1 request.

(Note: The list provided is a partial summary based on the observation. The full details include requests from various member state courts across the years 2016 to 2024, totaling 99 entries.)

Answer to Q9. Based on the provided data sample, there are no requests for preliminary rulings submitted to the Court by Member State courts, hence no overlapping areas regarding the interpretation of specific articles of GDPR can be identified from this dataset.

Answer to Q10. The key principles governing the processing of personal data derived from requests for preliminary rulings submitted to the Court by Member State courts include, but are not limited to, the following:

- Protection of personal data, independence of supervisory authorities
- Purpose limitation, Data minimisation
- Consent, active consent, separate consent, obligation to fully inform
- Consent, freely given consent, informed consent, burden of proof
- Protection of natural persons with regard to the processing of personal data, lawfulness of processing, protection of personal data, fundamental rights
- The right to be forgotten, processing of sensitive data, exceptions for journalistic purposes or artistic expression
- Effective and complete protection of data subjects

(Note: Output continued in the same way, until the model's output size was exceeded)

Answer to Q11. Potential risk factors regarding data protection rights in the dataset could include ongoing legal challenges or the need for clarifications in the application of data protection laws, as indicated by the 222 documents related to preliminary rulings on data protection. Additionally, the focus on data protection issues and personal data handling practices in 359 and 353 documents, respectively, highlights the importance of these areas in the context of GDPR compliance and potential areas of risk that require careful management and adherence to legal standards.

Answer to Q12. Yes, there is at least one document in the dataset that discusses the relationship between the EU Court of Justice and national courts regarding the interpretation of GDPR provisions.

4.3.3.2 Ethical, legal, and safety testing

The ethical, legal, and safety evaluation of PCS3 centres on assessing whether an LLM-based legal assistant can operate responsibly within governance contexts where accuracy, accountability, and compliance with regulatory frameworks are essential. Although the pilot remains exploratory and does not involve binding administrative decisions or the processing of personal data, the testing process evaluates the conceptual system against the standards that would apply in a real-world public-sector environment.

Legal Compliance. The system processes only publicly accessible legislative and jurisprudential materials sourced from EUR-Lex, meaning that GDPR obligations regarding personal data do not directly apply. Nevertheless, legal experts emphasised that once such a system is integrated into actual administrative workflows, its use could fall under the domain of the EU AI Act, particularly in areas involving the interpretation of legal rules, regulatory compliance, or decision support with potential legal consequences. Under the AI Act, legal-interpretation tools are likely to be classified as high-risk, requiring documented risk-management processes, technical documentation and transparency artefacts, traceability of outputs, human oversight mechanisms, and rigorous monitoring for erroneous or misleading interpretations.

The testing therefore evaluates PCS3’s conceptual design against these requirements, identifying where current capabilities align with regulatory expectations and where further safeguards would be necessary.

Ethical Considerations. LLMs present several ethical risks relevant to legal interpretation, including hallucinations, selective retrieval, inconsistency across prompts, bias in summarisation, and overconfidence in incorrect answers. The RAG architecture was tested to determine whether grounding model responses in authoritative legal texts meaningfully reduces hallucination risk. Results showed improved fidelity for fact-based questions but persistent vulnerabilities in complex reasoning tasks. Furthermore, testing highlighted the need for explicit uncertainty communication, particularly given the normative and interpretative nature of many legal queries.

Ethical assessment also focused on the risk of over-reliance by human users. In legal contexts, probabilistic text generators may inadvertently appear authoritative, potentially leading to premature acceptance of AI-generated interpretations. To counter this, the conceptual design requires that all outputs remain advisory, accompanied by accessible warning signals and human validation steps.

Safety Testing. Safety evaluation examined the robustness of the system when faced with multi-layered legal questions requiring numerical computation, cross-document referencing, or inference from case-law patterns. Testing revealed that

- the RAG system is relatively safe for extraction-based questions but limited when reasoning is required
- the agent-based system can perform complex multi-step tasks but sometimes produces incomplete or partially incorrect reasoning without signalling low confidence
- the absence of structured intermediate reasoning traces hinders error detection and contestability.

These observations indicate that safety in legal-AI systems depends not only on accuracy but also on transparent reasoning pathways, systematic constraint mechanisms, and human-in-the-loop checkpoints for ambiguous or high-impact answers.

Bias, Fairness, and Consistency. Bias testing focused on whether LLMs exhibit variability across similar legal queries, potentially generating inconsistent interpretations of equivalent provisions. Early qualitative assessments showed that differences in phrasing, order of presentation, or scenario prompting could alter the depth and structure of the answer. While these inconsistencies do not indicate systematic bias against specific groups (given the nature of the task), they do raise concerns about reproducibility, an essential criterion for fairness and legal defensibility. Mitigation strategies include stricter prompt templates, improved retrieval constraints, and mechanisms for forcing model self-verification or comparison of alternative reasoning paths.

Oversight and Accountability. The evaluation further addressed how accountability would be maintained in a real governance environment. Because public administrations require clear attribution of responsibility for legal interpretation, the system is explicitly positioned as a decision-support tool, not an autonomous legal analyst. Human oversight remains the primary safeguard, and the design anticipates future institutional requirements such as audit logging of all system interactions, version tracking of models and legal corpora, procedural escalation for ambiguous or conflicting outputs, and documentation of system limitations for end-users.

4.3.3.3 User Testing and Feedback Integration

Because PCS3 was designed as an exploratory research pilot rather than an operational deployment, no formal end-user testing sessions were conducted with public-sector legal professionals or administrative staff. However, structured internal expert review served as a functional analogue to user testing and generated valuable insights into the system’s usability, reliability, and potential alignment with governance workflows. This subsection summarises the evaluative practices adopted and the preliminary lessons drawn.

First, the legal experts involved in defining the reference questions conducted a qualitative assessment of the generated answers produced by both the RAG-based and agent-based systems. Their evaluation focused on the legal correctness of the responses, the adequacy of cited sources, the coherence and structure of the explanations, and the system's behaviour across different difficulty levels of questions.

Experts noted that the RAG-based system performed reliably for beginner questions, where the required information is explicitly stated within the GDPR text. Responses were generally accurate, well grounded in retrieved passages, and easy to verify. In contrast, the agent-based system, designed for more complex tasks, exhibited greater variability. While several outputs were insightful and aligned with expected legal reasoning patterns, others revealed challenges such as numerical inconsistencies, overgeneralisation, or incomplete legal justification. This underscores the need for systematic source verification and enhanced reasoning control before such systems can be integrated into public-sector legal tasks.

Second, the research team conducted an internal evaluation of the system's explainability and traceability features, assessing whether the outputs could be inspected and contested by human reviewers. The RAG pipeline provided clear links to authoritative legal texts, facilitating expert review and enabling transparent cross-checking. By contrast, the autonomous agent system, although capable of multi-step processing, offered limited transparency regarding intermediate reasoning paths. Legal experts emphasised that, for real-world governance applications, the ability to reconstruct how the system reached a conclusion, including intermediary assumptions and retrieved sources, would be indispensable. This insight directly informs the refinement of KPIs related to traceability, auditability, and human oversight readiness.

Third, the team reflected on the potential integration of LLM-based legal assistants into public-sector workflows, considering typical tasks such as legal research, interpretation support, GDPR compliance analysis, and case-preparation assistance. This reflection highlighted that while the prototype demonstrates clear potential for accelerating information retrieval and providing structured overviews, its outputs cannot be used operationally without robust human review, stringent safeguards, and domain-specific calibration. Participants involved in the review stressed that public-sector adoption would require additional layers of user-facing functionality, such as confidence indicators, citation completeness alerts, and risk flags, to effectively support legal workflows.

Fourth, the internal review generated insights into future user requirements, which will be relevant for subsequent phases of the AIGOV Transformation and Adoption Framework. These include the need for training in AI-assisted legal research tools, clear guidance on the appropriate use and limitations of LLM-based assistants, institutional processes for validating and archiving AI-generated outputs, governance protocols for ambiguous, conflicting, or incomplete answers, and organisational capacity to monitor and refine AI models over time.

4.3.4 Assessment of the KPIs

The tables below present the values for the KPIs related to this PCS.

KPI	Assessment	Status
Legal correctness score	For Beginner questions (Q1-Q4), the RAG-based system over the GDPR text produced answers that legal experts considered substantively correct and well grounded in the Regulation. For Intermediate and Expert questions (Q5-Q12), the agent-based system produced plausible but sometimes incomplete, approximate, or partially incorrect answers, particularly for numerical counts, time calculations, and complex aggregations over case law.	Partially Assessed
Consistency and reproducibility	For the RAG system (Q1-Q4) with fixed prompts and temperature, responses were highly stable and reproducible. For the agent-based system (Q5-Q12), outputs remained broadly similar in structure and reasoning but showed variation in phrasing and, occasionally, in specific values or emphases, especially when the agent had to reason over many documents and inferred patterns.	Measured
Agreement metrics	n/a	Defined (future validation)
Source citation accuracy	In the RAG setup, retrieved GDPR chunks were clearly relevant and answers remained closely tied to the underlying articles/recitals. In the agent-based system, references to legal bases were often implicit; the system relied more on summarised metadata and content, with less explicit, article-level citation.	Partially Assessed

Table 21 Assessment of Technical Performance KPIs for PCS3

KPI	Assessment	Status
Traceability of reasoning	For Beginner questions, traceability is strong: the RAG system retrieves specific GDPR passages and the link between context and answer is clear. For the agent-based system, the multi-step is conceptually documented, but not all intermediate reasoning steps are exposed to users, and the mapping from source to conclusion can be opaque.	Partially assessed

Clarity of responses	Legal experts reported that responses to Q1-Q4 were clear, well-structured, and close to the legal text, suitable for professional readers. For Q5-Q12, answers were linguistically clear but sometimes conceptually vague, incomplete, or over-general, especially when the system was expected to derive statistics or classify large sets of cases.	Partially assessed
Auditability	n/a	Defined (future validation)

Table 22 Assessment of Explainability and Transparency KPIs for PCS3

KPI	Assessment	Status
Alignment with legal constraints	For Q1-Q4, alignment is good: answers are grounded in the GDPR text and did not exhibit obvious hallucinations. For Q5-Q12, several answers are only approximately correct or incomplete (e.g., partial counts, vague risk descriptions, "generic" formulations not strictly tied to specific provisions or cases). The system does not always signal uncertainty.	Defined (future validation)
Human oversight readiness	n/a	Defined (future validation)
Integration potential	The RAG-based component for basic questions shows promising integration potential as a legal lookup/explanation tool. The agent-based component for complex analytics is not yet mature for integration, due to uncertainty, lack of robust metrics, and limited explainability of multi-step reasoning.	Partially assessed
Risk indicators	n/a	Defined (future validation)

Table 23 Assessment of Responsible AI and Public-Sector Applicability KPIs for PCS3

4.4 Pilot Case study 4: Fostering Multilingual Deliberation through Generative Artificial Intelligence

4.4.1 Phase 1: Assess Readiness and Context

4.4.1.1 Strategic Context and Public Value

Democracies around the globe face internal and external threats such as electoral interference, disinformation, as well as rising populism and authoritarianism. One answer to the quest for a more democratically legitimate Union and fulfilling citizens' expectations towards political institutions is the increase of participation and the broad access to the deliberative processes. Towards this end, the stringent necessity of creating a European Public Sphere before, and over, an economic union has been widely recognized [128].

However, in many cases deliberative democracy is hindered by barriers that are related to multilingualism as well as to cultural and social diversity. Political scientists know surprisingly little about how multilingualism affects politics and policy-making, even though language provides the basis for all interaction, collaboration, condensation, deliberation, and negotiation between political actors [129]. The challenges associated with comprehending public discourse underscore the complexity of adjusting to diverse linguistic and cultural environments within democratic procedures. These challenges frequently make it more difficult to collaborate, communicate, and reach consensus, which undermines the core ideas of deliberative democracy. In order to address these challenges and promote more inclusive and efficient democratic practices, a deeper comprehension of the interactions between language, culture, and politics is necessary.

In this context, the objective of this case study is to create a framework for enabling AI - supported multilingual deliberations. The framework utilises state-of-the-art generative AI technologies to address challenges related to cultural diversity and multilingualism that impede deliberative democracy. The framework is complemented by a technical implementation blueprint demonstrating how these components can interoperate within a public-sector context.

4.4.1.1.1 Multilingual Deliberation challenges

Deliberative democracy in many cases is hindered by barriers that are related to multilingualism as well as to cultural and social diversity. For example, since "Tomorrow's Europe", Europe's first transnational deliberative experiment, several pan-European initiatives have been organised to enable people from across Europe to share their ideas and help shape the EU's common future, including the European Citizens Initiative (ECI) and the Conference on the Future of Europe (CoFoE). The outcomes of these cross-border deliberative experiments have revealed various challenges that are related to deliberative democracy's scientific theories, deliberative methods and practices, as well as technological solutions employed.

The design of deliberative democracy includes aspects such as selection methods, timing, facilitation, format and structure, etc. which, when not considered, can result in unintended consequences. For example, the selection process may lead to not equal representation of different socio-cultural groups and countries, the timing may not allow diverse groups of participants to achieve common understanding, the facilitators may not be able to appreciate and interact with people who identify with cultures different from their own, etc. As a result, various deliberation designs for reconciling deliberative democracy and multilingualism should be rigorously explored and evaluated.

Moreover, online deliberations, versus face-to-face sessions, tend to disproportionately represent specific groups of people (e.g., young, male, and white users), attracting more ideologically moderate individuals, generating more negative emotions, and exhibiting a lower chance of reaching a consensus [2]. To ensure successful deliberations and fair representation of different views, effective methods and tools for integrating face-to-face with online deliberations as well as multimodal means of communication (text, audio, and video) should be considered and evaluated. These tools should be able to effectively analyse large volumes of online contributions and con-dense them in a comprehensive manner.

In multilingual deliberations, participants have the opportunity to express their arguments in their mother tongue and interpreters or translation tools are involved to facilitate the process. However, it is possible that interpreters slow down the discussion and disrupt the natural flow of dialogue, potentially hindering the dynamic ex-change of ideas [1]. On the other hand, interpreters tend to simplify, standardise, and neutralise language and thus it is believed that they can reduce the potential for conflict [35]. Moreover, many theories on deliberative democracy suggest that instead of focusing on the common language, we should move the attention towards the notion of a shared understanding [26]. As a result, existing machine and speech translation tools should be enhanced enabling interpret-like condensation that can be dynamically adapted based on the context of the deliberation (e.g., energy in the room, polarisation, fluidity and liveliness of conversations etc.) as well as to capture different cultural nuances and social codes and perceptions in language.

Finally, political scientists know surprisingly little about how multilingualism affects politics and policymaking. For example, although language barriers may lead to misunderstandings, confusion, and tension between political actors, recent studies suggest that multilingualism entails that the language(s) of EU politics tend to be utilitarian, simple, standardised, neutral, decultured, and de-ideologised. As a result, a multi-dimensional evaluation of multiple deliberative methods and design methods combined with advanced tools should be performed in order to enable understanding the impact of multilingualism in the democratic process.

4.4.1.2 Stakeholders

The case study represents a conceptual and technical exploration of how AI-based multilingual tools can support democratic participation. Accordingly, the stakeholders identified for this case study reflect the groups that would be involved in, or affected by, the deployment of such a system in real deliberative settings. These stakeholders align with the internal and external types defined in *D1.2 The AIGOV Ecosystem*, even though

Internal stakeholders consist of actors embedded within public sector institutions.

- *Public authorities and public organisations* constitute the institutional environment in which deliberation processes take place and are responsible for integrating the outputs of multilingual discussions into policy-making.
- *Public servants* support the organisation, coordination, and operational management of deliberation sessions, ensuring that participation procedures remain inclusive and functional across linguistic groups.
- *Regulators and policy makers* rely on the results of deliberative processes to inform democratic decisions while ensuring compliance with legal requirements, such as data protection and transparency obligations.
- *IT stakeholders*, including agency AI experts and developers, design, implement, and maintain the technological components of the multilingual deliberation system, including speech recognition, machine translation, summarisation, and argument mining modules.
- *Public service designers* contribute by shaping the structure and user experience of deliberation processes, ensuring accessibility, fairness, and usability for all participants.

External stakeholders represent those outside the public sector who directly participate in or are significantly affected by multilingual deliberation.

- *Citizens and residents* are the primary external stakeholder group. They participate in deliberative processes, contribute viewpoints in different languages, and benefit from improved inclusiveness and mutual understanding.
- *Customer advocacy groups* and *civil society organisations support* equitable deliberation by promoting linguistic rights, monitoring inclusiveness, and ensuring that multilingual technologies do not reinforce social or cultural inequalities.
- The *academic and scientific community* (including experts in political science, linguistics, and AI) contributes research, evaluates system performance, and informs methodological improvements.

- *Third parties*, such as independent facilitators, moderators, and professional interpreters, support deliberations operationally and may complement or benchmark AI-based translation and argumentation tools.

4.4.1.3 Data readiness

Because this pilot case study is conceptual and does not involve the processing of real deliberation data, no datasets were collected or analysed. However, the readiness assessment requires specifying the data landscape, data dependencies, and technical prerequisites that would be necessary for deploying a multilingual deliberation system in an operational public-sector environment. This analysis follows the principles of the AIGOV Government Data Value Cycle and the findings of deliverables D1.1 (State of Play) and D1.2 (AIGOV Ecosystem).

(1) Types of Data Required for Multilingual Deliberation.

A real-world AI-supported deliberation environment would rely mainly on unstructured and semi-structured multilingual data, including:

- Audio data: speech recordings from deliberation sessions (plenary sessions, working groups, citizen assemblies).
- Textual contributions: written comments, chat messages, online submissions, and forum posts in multiple languages.
- Transcripts: automatic speech recognition (ASR) outputs or human-generated transcripts, including timestamps and speaker identifiers (pseudonymised).
- Multilingual policy documents: briefing notes, consultation papers, legislative texts, and meeting minutes used for contextual grounding.
- Metadata: language labels, timestamps, session identifiers, speaking time metrics, and participation statistics.

As identified in D1.1, such multilingual unstructured data are widespread in public-sector contexts but remain underexploited due to limited tooling, inconsistent formats, and insufficient automation.

(2) Sources of Data

A fully implemented system would require the integration of data from:

- Internal public administration sources, such as parliaments, councils, ministries, and committees producing audio/video records
- Public consultation platforms that gather multilingual written input;
- Administrative documents relevant to the discussion topics.
- External or open data sources, such as multilingual corpora (EuroParl, OPUS) for pre-training or fine-tuning translation/summarisation models

- Relevant open datasets providing contextual political, legal, or socio-economic information
- Publicly available multilingual LLM resources.

Although not used in the conceptual pilot, these sources define the data ecosystem on which the framework relies.

While this case study is conceptual, an operational deployment would require high-quality multilingual datasets, including diverse language varieties and dialects, robust transcription and translation pipelines, with human-in-the-loop validation for sensitive content, standardised metadata schemas (e.g., speaker roles, language markers), consistent with European interoperability frameworks (DCAT-AP, SEMIC guidelines), secure storage compliant with GDPR, especially when deliberation data may reveal political opinions., mechanisms for bias detection, ensuring fair representation of all linguistic communities, full audit trails, documenting preprocessing, translations, model outputs, and human interventions.

As none of these pipelines were deployed in this conceptual phase, the readiness rating for data infrastructure is low, not because it is infeasible, but because it requires substantial investment, governance structures, capacity-building, and trustworthy AI safeguards.

4.4.1.4 Governance, Ethics, and Legal Compliance

Any real-world deployment of AI-supported multilingual deliberation tools would operate within a highly sensitive democratic context and therefore face substantial legal, ethical, and governance requirements. Although this pilot case study is conceptual and does not process personal data, a readiness assessment must consider the regulatory and ethical constraints that would apply in operational settings.

Multilingual deliberation systems process contributions that may include personal data, political opinions, linguistic identifiers, and audio recordings, all of which are subject to strict protection under the General Data Protection Regulation (GDPR). GDPR requirements such as data minimisation, purpose limitation, pseudonymisation, lawful basis for processing, and privacy-by-design would be central to system design. Special categories of data—particularly political opinions—trigger heightened safeguards, reinforcing the need for secure storage, strict access control, and explicit transparency measures.

In addition, the EU Artificial Intelligence Act introduces specific obligations for AI systems deployed in democratic processes. Systems that facilitate political participation, influence decision-making, or support deliberative procedures may fall under the category of high-risk AI, requiring documented risk management processes, technical robustness and security assessments, human oversight mechanisms, transparency regarding model limitations, and continuous monitoring for bias, hallucinations, or misrepresentation.

These requirements imply that AI-mediated translation, summarisation, argument extraction, or sentiment analysis must remain advisory, traceable, and under explicit human control, ensuring that AI does not autonomously shape political discourse, exclude minority voices, or distort participant intent.

Ethically, multilingual deliberation tools must safeguard principles of fairness, inclusiveness, linguistic equality, and cultural sensitivity. Automated translations must avoid reinforcing stereotypes, erasing cultural nuance, or privileging dominant languages. Similarly, summarisation and argument-mining systems must provide balanced representations of contributions to avoid unfair amplification or marginalisation of particular groups or viewpoints.

Governance structures would also need to define accountability roles, clarifying responsibilities for model performance and validation, human review of critical outputs, escalation procedures when the system behaves unexpectedly, documentation and auditability throughout the AI lifecycle.

Although the present case study does not engage with real data or operational decision-making, these governance, legal, and ethical considerations constitute essential prerequisites for any future deployment of multilingual AI systems within democratic processes. The analysis therefore informs subsequent design and evaluation phases, ensuring that conceptual development remains aligned with responsible public-sector AI practices.

4.4.1.5 Organisational and Capability Capacity

Because this pilot case study is conceptual and does not involve the deployment of AI technologies within a functioning public administration, the assessment of organisational and capability capacity focuses on identifying the types of expertise, infrastructures, and institutional readiness that would be required if an AI-supported multilingual deliberation system were to be implemented in real democratic settings. The study itself relies primarily on research-oriented competencies in generative AI, multilingual natural language processing, and deliberative process design; however, the transition from conceptual framework to operational practice would necessitate a significantly broader capability landscape.

In an applied context, public administrations would require advanced technical expertise to support the development, fine-tuning, and maintenance of the AI components underpinning multilingual deliberation. This includes specialists in natural language processing and machine learning who can adapt large language models to translation, summarisation, speech recognition, and argument mining tasks. Beyond model development, skilled data engineers and machine learning operations practitioners would be necessary to design robust pipelines for data ingestion, storage, and versioning, ensuring that multimodal inputs such as audio, transcripts, written contributions, and metadata are processed reliably and in compliance with governance requirements. Expertise in linguistics and intercultural communication would also be essential, as effective multilingual deliberation requires technologies that can convey nuance and meaning across languages and socio-cultural contexts.

The operation of such a system would also depend on strong human oversight capacities within the hosting institution. Public administrations would need established structures for monitoring, validating, and contextualising AI outputs, ensuring that translations, summaries, and argument extraction remain advisory and do not acquire unintended authoritative status. Facilitators and moderators would need to be trained to use AI-assisted tools without compromising neutrality, inclusiveness, or the natural dynamics of deliberation. Ethical and compliance officers would likewise be essential to oversee issues such as algorithmic bias, risk management, transparency obligations, and conformity with relevant legal frameworks, including the GDPR and the forthcoming EU AI Act, which places stringent requirements on systems used in democratic processes.

In addition to human capabilities, the adoption of multilingual deliberation technologies requires a secure and scalable digital infrastructure. Public administrations would need systems capable of processing multilingual audio and text data, managing large volumes of unstructured information, and ensuring traceability and accountability of all transformations applied to the data. Real-time or near-real-time processing capabilities would be required for live deliberative settings, including automatic speech recognition, language detection, translation, and summarisation. User-facing components, such as deliberation dashboards, participant interfaces, and moderation tools, would need to be accessible, multilingual, and aligned with public-sector usability standards.

Finally, institutional readiness extends beyond technical skills and infrastructures to the broader organisational environment. Effective deployment would require leadership committed to enhancing democratic participation, as well as public servants equipped with sufficient data literacy and AI literacy to interact confidently with the tools. Organisational change processes may be necessary to integrate AI-based capabilities into existing deliberation workflows, while collaboration across policy, IT, legal, and democratic engagement units would help ensure that the system is used in a coherent and responsible manner.

In summary, although the current case study does not involve an operational setting, it identifies the critical organisational and capability requirements that would need to be in place for AI-supported multilingual deliberation to function effectively within a public administration. These requirements encompass technical expertise, governance and oversight structures, secure infrastructure, and institutional readiness. The insights generated here therefore serve as an early contribution to future capability planning for administrations seeking to adopt AI tools in democratic processes.

4.4.2 Phase 2: Design Ethical and Value-Aligned Use Cases

Phase 2 focuses on translating the contextual analysis and readiness assessment from Phase 1 into a clearly defined, responsible, and evaluable AI use case that supports multilingual deliberation. While the case study is conceptual rather than operational, the aim of this phase is to articulate how generative AI technologies can be designed to address the barriers identified earlier (linguistic fragmentation, cultural diversity, unequal participation, and limited comprehension) while remaining aligned with public value, democratic principles, and governance safeguards.

Following the AIGOV Transformation and Adoption Framework, Phase 2 structures this translation process into three interconnected activities:

(1) Problem and service redesign definition (Section 4.4.2.1), which reframes deliberation challenges in terms of user needs, service delivery limitations, and opportunities for human-AI collaboration in multilingual contexts.

(2) Responsible AI use case design (Section 4.4.2.2), which converts the redesigned problem into a concrete AI-enabled approach, specifying explainability requirements, ethical safeguards, human oversight roles, and inclusiveness considerations appropriate for deliberative democratic environments.

(3) Success metrics and evaluation criteria (Section 4.2.2.3), which establish how the conceptual system should be assessed, covering transparency, accessibility, robustness, fairness, linguistic adequacy, and alignment with the AIGOV principles.

Together, these activities define the parameters of the multilingual deliberation framework and provide a clear methodological foundation for Phase 3 (Build and Test). Although no full technical implementation is carried out in this pilot, Phase 2 ensures that the conceptual design of the system is coherent, ethically grounded, and traceable to the strategic objectives established in Phase 1.

4.4.2.1 Problem and service redesign definition

The redesign of multilingual deliberation services begins by identifying the structural, linguistic, cultural, and procedural barriers that currently hinder inclusive and effective democratic participation across Europe.

Deliberative democracy in many cases is hindered by barriers that are related to multilingualism as well as to cultural and social diversity. For example, since "Tomorrow's Europe", Europe's first transnational deliberative experiment, several pan-European initiatives have been organised to enable people from across Europe to share their ideas and help shape the EU's common future, including the European Citizens Initiative (ECI) and the Conference on the Future of Europe (CoFoE). The outcomes of these cross-border deliberative experiments have revealed various challenges that are related to deliberative democracy's scientific theories, deliberative methods and practices, as well as technological solutions employed.

The design of deliberative democracy includes aspects such as selection methods, timing, facilitation, format and structure, etc. which, when not considered, can result in unintended consequences. For example, the selection process may lead to not equal representation of different socio-cultural groups and countries, the timing may not allow diverse groups of participants to achieve common understanding, the facilitators may not be able to appreciate and interact with people who identify with cultures different from their own, etc. As a result, various deliberation designs for reconciling deliberative democracy and multilingualism should be rigorously explored and evaluated.

Moreover, online deliberations, versus face-to-face sessions, tend to disproportionately represent specific groups of people (e.g., young, male, and white users), attracting more ideologically moderate individuals, generating more negative emotions, and exhibiting a lower chance of reaching a consensus [3]. To ensure successful deliberations and fair representation of different views, effective methods and tools for integrating face-to-face with online deliberations as well as multimodal means of communication (text, audio, and video) should be considered and evaluated. These tools should be able to effectively analyse large volumes of online contributions and con-dense them in a comprehensive manner.

In multilingual deliberations, participants have the opportunity to express their arguments in their mother tongue and interpreters or translation tools are involved to facilitate the process. However, it is possible that interpreters slow down the discussion and disrupt the natural flow of dialogue, potentially hindering the dynamic exchange of ideas [1]. On the other hand, interpreters tend to simplify, standardise, and neutralise language and thus it is believed that they can reduce the potential for conflict [1]. Moreover, many theories on deliberative democracy suggest that instead of focusing on the common language, we should move the attention towards the notion of a shared understanding [33]. As a result, existing machine and speech translation tools should be enhanced enabling interpret-like condensation that can be dynamically adapted based on the context of the deliberation (e.g., energy in the room, polarisation, fluidity and liveliness of conversations etc.) as well as to capture different cultural nuances and social codes and perceptions in language.

Finally, political scientists know surprisingly little about how multilingualism affects politics and policymaking. For example, although language barriers may lead to misunderstandings, confusion, and tension between political actors, recent studies suggest that multilingualism entails that the language(s) of EU politics tend to be utilitarian, simple, standardised, neutral, de-cultured, and de-ideologised. As a result, a multi-dimensional evaluation of multiple deliberative methods and design methods combined with advanced tools should be performed in order to enable understanding the impact of multilingualism in the democratic process.

Taken together, these challenges reveal a systemic gap: public administrations lack an integrated, multilingual, AI-assisted environment capable of supporting equitable, dynamic, and culturally sensitive deliberation at scale.

In response, the service challenge is reframed as the need to design an AI-supported multilingual deliberation framework that enhances comprehension, improves accessibility, and broadens participation while preserving the integrity and inclusiveness of democratic dialogue. Generative AI and LLM-based components, such as speech recognition, machine translation, summarisation, argument extraction, and semantic clustering, offer significant potential for redesigning deliberative processes, provided they remain transparent, explainable, and under robust human oversight.

The redesigned service concept therefore envisions deliberation as a collaborative human-AI process where participants contribute in their native languages, AI assists by translating, condensing, and structuring content without replacing human facilitators, cultural and linguistic nuance is preserved rather than flattened, biases and risks are monitored through governance safeguards, and public servants gain new analytical capabilities to synthesise contributions.

This reframing establishes the foundations for the Responsible AI Use Case Design (Section 4.4.2.2) and guides the selection of evaluation metrics and KPIs in Section 4.4.2.3.

4.4.2.2 Responsible AI Use Case Design

Following the reframing of the service challenge, the multilingual deliberation framework is translated into a responsible and value-aligned AI use case designed to enhance democratic dialogue while preserving the principles of transparency, fairness, and human oversight. Because deliberative processes involve sensitive political expression and highly diverse linguistic and cultural contexts, the design of this use case is guided by the requirements of the AIGOV Framework for Trustworthy, Fair, and Accountable AI.

A central design requirement concerns explainability and comprehensibility. AI systems used in democratic processes must not only provide accurate translations, summaries, or argument extractions but must do so in ways that remain interpretable to facilitators, participants, and policymakers. For this reason, the use case establishes explainability as a core requirement. Every processing step, whether translation, condensation, or argument identification, must be traceable back to the source contributions, accompanied by a rationale that can be understood by non-technical users. Confidence indicators, source links, and accessible justifications are therefore essential for ensuring that the system supports informed deliberation rather than introducing opaque or unchallengeable outputs.

Ethical and legal safeguards also play a central role in the design. Deliberation data may contain personal information, identifiable speech, political opinions, and culturally sensitive expressions. Even though the current study does not process real data, the use case is built with the assumption that an operational environment would be subject to GDPR, national data protection law, and the high-risk obligations foreseen under the EU AI Act for systems used in democratic contexts. This implies the incorporation of privacy-by-design principles, strict purpose limitation, robust logging and auditability, and mechanisms for identifying and mitigating potential biases in multilingual outputs. The broader aim is to ensure that the system does not distort, suppress, or misrepresent contributions from any linguistic or cultural group.

Human oversight is another defining component of the use case design. The system is conceived as a decision-support tool that enhances facilitation and comprehension, not as an automated mechanism that replaces human judgment. Facilitators and analysts remain responsible for interpreting outputs, validating their correctness, and contextualising them within the dynamics of a live deliberation. Clear allocations of responsibility among public authorities, technical teams, and process facilitators are assumed as part of the governance structure, together with escalation pathways when outputs appear questionable or potentially harmful. The system is therefore intentionally designed to preserve the accountability of human actors.

Inclusiveness, multilingualism, and cultural sensitivity constitute additional pillars of the design. Because the purpose of the system is to enable deliberation across languages and cultural contexts, the model must treat all languages equitably, regardless of resource availability or political prominence. Translations and summaries must respect cultural nuance, rhetorical style, and social codes embedded in speech. Outputs must also be accessible across different literacy levels and communication modalities. Although the present study remains conceptual, the design anticipates deployment scenarios in which the system dynamically adapts to the emotional tone, complexity, and intercultural character of live deliberations, supporting shared understanding without imposing a dominant linguistic frame.

Finally, the design incorporates several operational considerations that would be essential in a real deployment. These include the need for a modular architecture capable of integrating translation, summarisation, and argument-mining components; the establishment of systematic traceability mechanisms; and the provision of multilingual documentation to ensure reproducibility and transparency. The system is thus conceived not as a singular technical artefact but as an extensible and auditable framework that can evolve alongside institutional capacities and regulatory developments.

Building on these requirements, the multilingual deliberation framework is structured into two main pillars: (a) contemporary research on democratic quality, participation, misinformation mitigation, and deliberative democracy in relations to multilingualism as well as to socio-political aspects of diverse European landscape examining mechanisms, methods, and design settings (e.g., selection methods, timing, facilitation, format and structure, scoping, and settings of deliberations) to enhance multilingual and multicultural participation, to improve perceived trust and responsiveness, and to augment the quality of consensus proposals, legislative and policy recommendations and (b) state-of-the-art technological advancements related to computational linguistics, language technologies, including Large Language Models, Explainable Artificial Intelligence, Knowledge Graphs and neuro-symbolic AI architectures to develop innovative software components related to machine and speech translation as well as to multilingual argument mining and deliberation management.

Capitalizing on these pillars, the framework proposes five key offerings that enable the creation of multilingual deliberation spaces in Europe:

i. Multilingual and Multicultural Deliberation Design,
ii. Machine Translation and Interpretation for Citizen Deliberation,
iii. Multilingual Deliberation Comprehension,
iv. Online and Face-to-Face Multilingual Deliberation Support, and
v. Transparency, Trustworthiness, and Explainability in Citizen Deliberation.

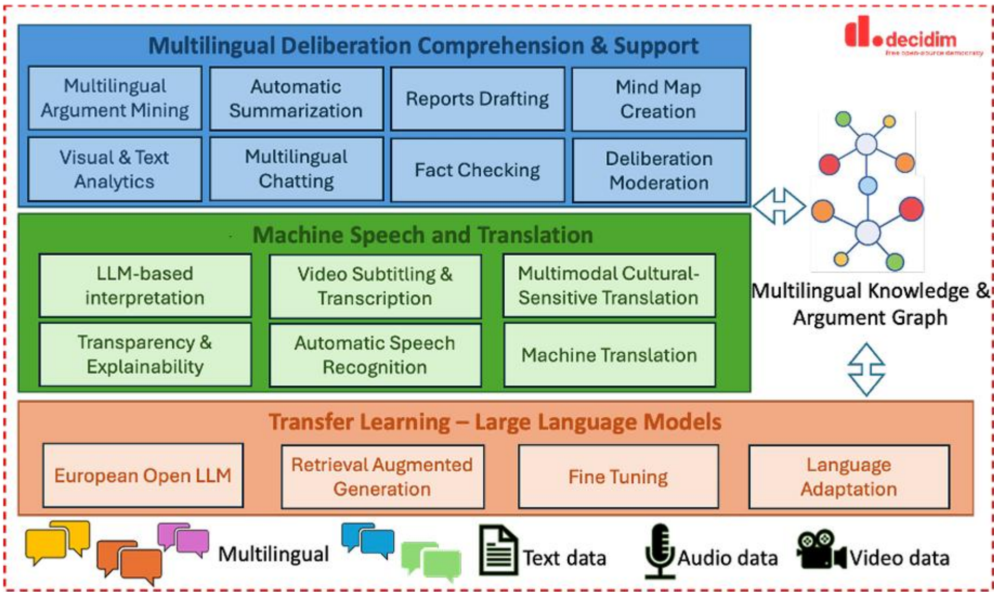


Figure 21 An Architecture for enabling Multilingual Deliberations using Generative Artificial Intelligence [16]

Multilingual and Multicultural Deliberation Design. Based on the framework, the design of multilingual deliberations should be enabled in a robust and scientifically sound manner, including aspects such as participant selection methods, timing, facilitation, format and structure, translation, scoping, processes, methods, settings, experts' involvement, etc. as well as communication channels (i.e., online, and face-to-face) and effective connection points of these channels in the case of hybrid approaches.

Machine Translation and Interpretation for Citizen Deliberation. The framework enhances existing machine and speech translation technologies so as to address the specific needs of citizen deliberations by bridging linguistic, social, and cultural divides, and handling multiple deliberation modalities including text, speech, and video in both face-to-face and online deliberation channels. To this end, this offering capitalizes on, and fine-tunes open European LLMs to enable interpreter-like machine translation that can be dynamically adapted according to the existing deliberation conditions. It enables interpret-like condensation thatcan be dynamically adapted based on the context of the deliberation (e.g., energy in the room, polarisation, etc.) and capture different cultural nuances and social codes and perceptions in language. In addition, real-time speech translation in face-to-face deliberations, video subtitling, and online text contribution translation are enhanced. End-to-end neural systems need to be employed to handle issues including synchronisation and low latency of data streams.

Multilingual Deliberation Comprehension. The interpretation, structuring, and presentation of multilingual deliberative content should be performed in a coherent and culturally aware manner. Towards this direction, LLMs and neuro-symbolic architectures can be employed to analyse deliberation content, and consequently identify and extract key components from online and face-to-face deliberations (e.g., topics, ideas, arguments), ensuring that the essence of discussions is captured irrespective of the language used. MT enhances the accuracy of argument extraction and presentation in multilingual contexts, ensuring that every voice is heard. Through culturally aware MT, deliberative comprehension, and knowledge structuring, the complete structure that incorporates all elements of the deliberation is created in the form of a multilingual Argumentation Knowledge Graph. This facilitates users in navigating through complex deliberation threads and fostering a more informed and engaged participation.

Online and Face-to-Face Multilingual Deliberation Support. Sophisticated AI tools can be harnessed to ensure that deliberations are, not only accessible to citizens with diverse linguistic and cultural backgrounds, but also substantively rich and well-organised, fostering a productive and democratic exchange of ideas. A focus on content moderation and quality control methods is required, implementing advanced AI algorithms that scrutinise deliberative content to filter out irrelevant or inappropriate material. Through intelligent argument clustering algorithms, similar arguments and ideas can be grouped and presented in a structured manner allowing participants to easily navigate through the deliberation themes and engage with content that resonates with their interests or expertise. Additionally, AI-driven fact-checking tools can verify the accuracy of statements and claims. Finally, advanced data visualisation techniques can be used to generate mindmaps and draft reports, translating complex deliberative discussions into visually appealing and easy-to-understand formats, summarising the outcomes and key points of deliberations.

Transparency, Trustworthiness, and Explainability in Citizen Deliberation. This offering goes one step beyond delivering robust, accurate, more empathetic and culture-aware translation services, by focusing on model explainability, enhancing transparency, accountability and trust in the LLM models. Explainability of LLMs is vital, as it provides insights into the translation and summarization choices made by the models, ensuring that social and cultural aspects are accurately conveyed.

Evaluation. The evaluation of a Generative AI based multilingual deliberation aims to comprehensively discern its influence on democratic quality aspects, including participation rates, deliberation quality, misinformation trends, etc.. Such an evaluation could be based on assessing dimensions like the quality of information, the quality of the deliberation, the presence of misinformation, but also the dynamics of the participation (e.g., the inclusiveness of and trust in the deliberation process), and the policy impact of the deliberation (e.g., track the shifts of public opinion among participants). Finally, various approaches like quantitative and/or LLM-based qualitative analysis can be employed to track belief changes and conduct automated multilingual content analysis of political discussions.

These five offerings together constitute the Multilingual Deliberation Framework based on Generative AI**.** They provide the conceptual backbone for the case study and define the functional capabilities that Phase 3 translates into a technical prototype architecture and testing strategy.

4.4.2.3 Success Metrics and Impact Definition

To evaluate the anticipated performance, public value, and responsible AI alignment of the multilingual deliberation framework, a structured set of Key Performance Indicators (KPIs) was defined. Unlike operational pilots where real data and user interactions allow for empirical measurement, this case study is conceptual and exploratory. As such, the KPIs serve primarily as ex-ante evaluative criteria, defining how the system would be assessed in a real-world deployment rather than reporting measurable results.

These KPIs fulfil several purposes. First, they translate the design objectives established in Phase 2 into observable outcomes that reflect technical quality, transparency, fairness, and alignment with democratic values. Second, they act as a bridge between high-level policy expectations and concrete evaluation mechanisms, ensuring that any future implementation can be assessed systematically and in accordance with the AIGOV Framework for Trustworthy, Fair, and Accountable AI. Third, they provide a roadmap for future empirical testing, making explicit which success conditions must be met for multilingual deliberation technologies to be responsibly adopted in public-sector environments.

To reflect the multidimensional nature of deliberation systems, the KPIs are organised into three complementary categories:

- Technical Performance (Table 24), capturing translation accuracy, summarisation quality, argument extraction capability, and the robustness of multilingual processing;
- Explainability and Transparency (Table 25), assessing whether AI-generated interpretations and summaries can be traced, justified, and communicated clearly to non-technical users; and
- Responsible AI and Public-Sector Applicability (Table 26), evaluating fairness across languages, compliance with legal requirements, cultural sensitivity, human oversight mechanisms, and alignment with the principles of democratic governance.

For each category, the KPIs specify the metric, its purpose, and status. Indicators that could not be empirically tested in the current conceptual setting are marked as such, but are included to ensure completeness and to guide subsequent stages of evaluation in real-world pilots.

KPI	Definition	Status
Translation Accuracy	Measures correctness and semantic fidelity of machine translation across languages.	Defined (future validation)
Semantic Preservation	Degree to which meaning, tone, cultural nuance, and argumentative intent are preserved.	Defined (future validation)
Summarisation Quality	Accuracy and coherence of AI-generated summaries of multilingual deliberation content.	Defined (future validation)
Argument Extraction Accuracy	Ability to correctly identify claims, evidence, and counterarguments.	Defined (future validation)
Latency and Responsiveness	Measures system responsiveness in real-time or near-real-time deliberation environments.	Defined (future validation)
Multilingual Robustness	Consistency of performance across high-, medium-, and low-resource languages.	Defined (future validation)

Table 24 Technical Performance KPIs for PCS4

KPI	Definition	Status
Explainability of Outputs	Clarity with which the system justifies translations, summaries, or argument structures.	Defined (future validation)
Traceability of AI Decisions	Ability to trace outputs back to source text or audio segments.	Defined (future validation)
Uncertainty Communication	Extent to which the AI indicates uncertainty or ambiguity in its outputs.	Defined (future validation)

Transparency of Model Documentation	Availability and clarity of technical and methodological documentation.	Defined (future validation)

Table 25 Explainability and Transparency KPIs for PCS4

KPI	Definition	Status
Linguistic Fairness	Ensures all languages are treated equitably without systematic bias.	Defined (future validation)
Cultural Sensitivity	Ability to respect and preserve sociocultural nuances in language.	Defined (future validation)
Compliance with GDPR & AI Act	Determines whether the system meets legal obligations for high-risk democratic applications.	Defined (future validation)
Human Oversight Effectiveness	Ensures deliberation facilitators retain interpretive control and can override AI outputs.	Defined (future validation)
Perceived Usefulness for Deliberation	Measures whether the system enhances shared understanding and reduces cognitive load.	Defined (future validation)
Inclusiveness and Accessibility	Degree to which the system supports equitable access for participants with different languages and literacy levels.	Defined (future validation)
Alignment with Public Sector Values	Measures whether the system supports transparency, legitimacy, and democratic integrity.	Defined (future validation)

Table 26 Responsible AI and Public-Sector Applicability KPIs for PCS4

4.4.3 Phase 3: Build and Test

Phase 3 focuses on translating the conceptual design developed in Phase 2 into a structured prototype architecture and testing approach. Although no operational system was implemented, this phase demonstrates how the multilingual deliberation framework can be instantiated, validated, and iteratively improved. The emphasis is placed on verifying coherence, feasibility, risks, and alignment with the AIGOV principles before any real-world deployment is considered.

The activities of this phase correspond to the AIGOV Transformation and Adoption Framework by:

(1) creating a functional conceptual prototype that integrates the key AI components (Section 4.4.3.1),
(2) conducting ethical, legal, and governance checks to ensure responsible design (Section **Error! Reference source not found.**), and
(3) outlining a testing strategy involving potential end-users (4.4.3.2), even though user trials are not executed in this pilot.

4.4.3.1 Prototype Development and Technical Implementation

The prototype design in this phase operationalises the Multilingual Deliberation Framework based on Generative AI defined in Phase 2. It follows a design science and action research orientation: design science structures the process of problem identification, objective definition, artefact design, and evaluation, while action research emphasises iterative refinement in dialogue with democratic practice and institutional constraints. In this pilot, these paradigms are instantiated through the construction of a modular technical blueprint that realises the framework's five offerings (multilingual deliberation design, machine translation and interpretation, deliberation comprehension, online and face-to-face support, and transparency and explainability) within a coherent AI architecture for multilingual deliberation.

The resulting prototype design is organised into two main families of services: (i) deliberation comprehension and support services, and (ii) machine speech and translation services, which are integrated through a common data and knowledge infrastructure.

On the machine speech and translation side, the architecture foresees components for automatic speech recognition (ASR), machine translation (MT), LLM-based interpretation, and streaming management. ASR modules transform speech contributions from deliberation sessions into multilingual text, taking into account different accents, dialects, and speaking styles. MT components, based on neural machine translation models and, potentially, fine-tuned European open LLMs, provide written translations of contributions, subtitles for video streams, and translation of online text input. An LLM-based interpretation layer performs interpreter-like condensation: it rewrites and compresses contributions while preserving their semantic content and communicative intent, adapting to deliberation context (e.g., polarisation, emotional tone, tempo of exchanges). A streaming and synchronisation component manages real-time delivery of transcriptions, translations, and captions, ensuring low latency and coherent user experience in hybrid (online and face-to-face) settings.

On the deliberation comprehension and support side, the architecture defines a suite of services that operate on multilingual textual data. Multilingual argument mining identifies topics, claims, reasons, objections, and evaluation statements within deliberation content across languages. Summarisation services generate concise overviews at different levels of granularity (e.g., per session, per topic, per stakeholder group), while report drafting and mind map generation translate complex discussions into structured, navigable artefacts. Fact-checking and content moderation services help identify misinformation, abusive content, and off-topic contributions, supporting facilitators in maintaining deliberation quality. Visual and text analytics components provide dashboards that show participation dynamics, thematic clusters, and the evolution of arguments over time. A multilingual chat interface can be layered on top of these services to support interactive engagement and on-demand clarification.

These services are underpinned by a shared knowledge and data layer based on knowledge graphs and neuro-symbolic reasoning. Contributions, arguments, topics, and relationships between them are represented in a multilingual Argumentation Knowledge Graph, using suitable vocabularies and ontologies to ensure semantic consistency. This graph supports downstream services by enabling structured queries, tracing how positions were formed, and grounding LLM outputs in explicit representations. Retrieval-Augmented Generation (RAG) mechanisms link LLMs to curated knowledge bases, ensuring that generated summaries and explanations are better anchored in factual and deliberation-specific information.

From an implementation perspective, the prototype blueprint assumes an extensible, service-oriented architecture. Individual components (ASR, MT, LLM-based interpretation, argument mining, summarisation, fact-checking, moderation, knowledge graph management, visual analytics) are deployed as interoperable services connected via APIs. This modularity allows public administrations to adopt components incrementally, to replace or upgrade models over time, and to adapt the configuration to specific deliberation formats or institutional requirements.

Throughout the design, the AIGOV principles of transparency, accountability, and responsible data management are embedded as cross-cutting constraints. Each AI component is expected to expose logging and traceability interfaces, enabling administrators and facilitators to inspect inputs, outputs, and key decisions. The knowledge graph layer provides a structured record of how content has been interpreted and aggregated. Human-in-the-loop controls are assumed at critical stages, particularly for moderation, summarisation validation, and the interpretation of analytical dashboards. Although this pilot does not implement the full architecture, the prototype design demonstrates how the conceptual framework of PCS4 can be realised technically in a manner that remains compatible with the AIGOV Transformation and Adoption Framework and the requirements of trustworthy, fair, and accountable AI in democratic processes.

4.4.3.2 Ethical, Legal, and Safety Testing

The ethical, legal, and safety testing of the multilingual deliberation prototype focuses on assessing whether the system, as conceptually designed, can operate in accordance with the stringent requirements that govern democratic processes, data protection, and the responsible use of AI in the public sector. Although the pilot does not involve real users or personal data, this phase evaluates the system’s architecture, workflows, and safeguards against the principles and obligations that would apply in a real-world deployment. The aim is to ensure that the design is not only technically coherent but also aligned with the normative expectations of transparency, fairness, legality, and human accountability.

The prototype's reliance on multimodal deliberation data, speech recordings, transcripts, written contributions, argument structures, and derived summaries, raises significant ethical and regulatory considerations. In a real deployment, much of this material would contain sensitive personal information, including political opinions, identifiable speech patterns, demographic indicators, and culturally embedded expressions. For this reason, the testing phase assesses the system's compatibility with GDPR requirements, particularly regarding data minimisation, purpose limitation, lawful processing, pseudonymisation, and privacy-by-design constraints. Even at the conceptual level, the architecture is reviewed to ensure that each processing step, ASR transcription, translation, semantic extraction, and knowledge graph construction, could operate within a framework of controlled access, auditable transformations, and clearly defined retention policies.

Because deliberation-support systems are expected to fall under the high-risk classification of the EU AI Act, ethical and safety testing also examines how the prototype anticipates the Act's key obligations. These include robustness against errors, traceability of outputs, documentation of data sources and model behaviour, and the presence of well-defined human oversight mechanisms. A central part of this assessment concerns the interpretability and contestability of AI-generated outputs. The design is therefore evaluated for its ability to provide transparent rationales for translations, summaries, and argument extractions, allowing facilitators and participants to understand how outputs were generated and to challenge them when necessary. This is particularly important in settings where linguistic nuances or cultural references might be flattened, distorted, or misinterpreted by AI models.

Bias and fairness considerations form another major dimension of testing. The multilingual nature of deliberation means that disparities in model performance across high-, medium-, and low-resource languages may lead to systemic inequalities in representation. The architecture is therefore evaluated for its potential to introduce uneven translation quality, disproportionate information loss, or misclassification of arguments based on linguistic features. Testing also considers whether cultural nuance is preserved across language pairs and whether summarisation processes risk amplifying specific voices while diminishing others. Although empirical validation is not part of this pilot, the conceptual testing establishes the methodological expectations for future bias audits, including the need for representative datasets, linguistic diversity benchmarks, and continuous performance monitoring.

Safety testing further examines the system's resilience to known AI risks relevant to deliberative environments. These include hallucinations in LLM-generated content, misaligned summaries that shift meaning or emphasis, incorrect argument mappings, and failures in real-time translation that could disrupt the flow of discussion. The architecture incorporates layered safeguards, traceability, confidence indicators, and human validation checkpoints, to mitigate these risks. The testing assesses whether these safeguards are appropriately positioned within the workflow and whether they preserve the accountability of human facilitators, ensuring that AI remains strictly advisory and cannot autonomously influence the outcome of democratic deliberation.

Finally, ethical testing reflects on the broader democratic implications of deploying such a system. Public deliberations rely on trust, legitimacy, and the perception of fairness; any AI-mediated process must therefore reinforce, rather than undermine, these foundations. The prototype is evaluated for its potential to enhance inclusiveness and shared understanding, but also for risks of over-automation, over-reliance on AI-generated interpretations, or the inadvertent centralisation of epistemic authority in technical systems. The emphasis is placed on ensuring that AI augments human judgment, increases accessibility, and clarifies complex multilingual interactions, without displacing the human deliberative process or reducing the diversity of political expression.

In sum, the ethical, legal, and safety testing conducted in Phase 3 confirms that the prototype as designed is conceptually compatible with the requirements of GDPR, the EU AI Act, and democratic values. It also highlights the conditions under which a real-world implementation would need to operate, including robust oversight, multilingual fairness auditing, transparent explainability mechanisms, and strict governance of sensitive data. These insights lay the foundation for the subsequent definition of user testing activities and future empirical validation.

4.4.3.3 User Testing and Feedback Integration

Because PCS4 was developed as a conceptual case study rather than an implemented system, no user testing activities were carried out. The case study focused on defining potential workflows, identifying design requirements and analysing how Generative AI could support multilingual and multicultural democratic consultation in principle. As a result, the evaluation relied solely on conceptual analysis and expert review of the proposed approach, without empirical testing or direct interaction with end users.

4.4.4 Assessment of KPIS

KPI	Definition	Status
Translation Accuracy	n/a	Defined (future validation)
Semantic Preservation	n/a	Defined (future validation)
Summarisation Quality	n/a	Defined (future validation)
Argument Extraction Accuracy	n/a	Defined (future validation)
Latency and Responsiveness	n/a	Defined (future validation)

Multilingual Robustness	n/a	Defined (future validation)

Table 27 Technical Performance KPIs for PCS4

KPI	Definition	Status
Explainability of Outputs	n/a	Defined (future validation)
Traceability of AI Decisions	n/a	Defined (future validation)
Uncertainty Communication	n/a	Defined (future validation)
Transparency of Model Documentation	n/a	Defined (future validation)

Table 28 Explainability and Transparency KPIs for PCS4

KPI	Definition	Status
Linguistic Fairness	n/a	Defined (future validation)
Cultural Sensitivity	n/a	Defined (future validation)
Compliance with GDPR & AI Act	n/a	Defined (future validation)
Human Oversight Effectiveness	n/a	Defined (future validation)
Perceived Usefulness for Deliberation	n/a	Defined (future validation)
Inclusiveness and Accessibility	n/a	Defined (future validation)
Alignment with Public Sector Values	n/a	Defined (future validation)

Table 29 Responsible AI and Public-Sector Applicability KPIs for PCS4

5 Lessons Learnt

The four pilot case studies of the AIGOV project, spanning predictive policy analytics (PCS1), legal interpretation using LLMs (PCS2), LLM-based legal assistants (PCS3), and AI-supported multilingual deliberation (PCS4), offer a comprehensive basis for understanding the opportunities and limitations of AI adoption in European public-sector contexts. Although each pilot addressed a distinct policy domain and deployed different technical approaches, several cross-cutting lessons emerged. These lessons represent shared methodological insights, governance challenges, and public-sector requirements that should inform future AI deployment strategies and the continued refinement of the AIGOV frameworks.

5.1 Lesson 1 - Data Quality, Granularity, and Interoperability Remain Foundational Barriers. Across pilots, the quality and structure of available datasets significantly shaped what could be built, tested, or evaluated. PCS1 benefited from well-structured statistical datasets but still faced challenges related to missing variables, spatial imbalance, and interpretability of socio-economic indicators. PCS2 and PCS3 leveraged authoritative legal corpora from EUR-Lex, yet issues of semantic ambiguity, cross-referencing complexity, and absence of multilingual integration limited further experimentation. PCS4 revealed that deliberation data including speech, text, video, and multilingual contributions, remain largely unstandardised, difficult to synchronise, and institutionally fragmented. Together, these insights confirm the centrality of the AIGOV Government Data Value Cycle, demonstrating that responsible AI deployment requires

- clear data governance,
- stable and interoperable metadata structures,
- systematic documentation of data lineage, and
- mechanisms ensuring ethical and legally compliant data processing.

Without these conditions, even advanced AI models cannot be reliably used to support public-sector decision-making.

5.2 Lesson 2 - Explainability and Traceability Are Essential for Public-Sector Legitimacy. Despite their technical differences, all pilots underscored the need for transparency mechanisms that allow public servants to understand and scrutinise AI-generated outputs. PCS1 demonstrated that prediction models, even high-performing ones, must provide intelligible explanations for spatial and socioeconomic patterns to avoid opaque decision-support. PCS2 and PCS3 showed that legal tasks demand rigorous traceability, citation integrity, and reproducibility. RAG pipelines and agent-based orchestration improved explainability but did not eliminate model variability or hallucination risks. PCS4 emphasised that, in democratic contexts, AI-mediated translations and summaries must be contestable, source-linked, and auditable to safeguard fairness and inclusiveness.

Collectively, the pilots validate the AIGOV Framework for Trustworthy, Fair, and Accountable AI, especially its pillars on explainability, transparency, human oversight, and accountability. AI tools that cannot justify their outputs remain unsuitable for high-impact public-sector domains.

5.3 Lesson 3 - Human Oversight Remains Indispensable Across All AI Maturity Levels. In every pilot, AI was effective only when embedded within robust human-in-the-loop processes. PCS1 required interpretation by domain experts to contextualise model predictions and determine their policy relevance. PCS2 and PCS3 showed that legal professionals must validate, refine, or reject AI outputs, particularly when LLMs offer ambiguous or superficially plausible answers. PCS4 revealed that facilitators must remain the authoritative decision-makers; AI tools can support comprehension and inclusiveness but must not influence political discourse autonomously. A recurring insight is that AI augments, rather than replaces, professional judgment. Public-sector workflows must therefore be redesigned to incorporate checkpoints for human review, contestability, and accountability.

5.4 Lesson 4 - Organisational Readiness and Multidisciplinary Capacity Are Uneven. All pilots highlighted capacity gaps that must be addressed for successful AI adoption. Technical readiness varies markedly: few administrations possess in-house AI engineering, NLP expertise, or capacity for model monitoring. Domain experts (legal, policy, administrative) widely lack AI literacy, hindering their ability to supervise, contest, or interpret outputs. Governance institutions often lack procedures for risk management, auditing, data lifecycle management, and validation of AI-generated content. These findings reinforce the importance of Phase 1 (Assess Readiness & Context) of the AIGOV Transformation and Adoption Framework. Successful adoption requires not only technological capabilities but also institutional structures, training programmes, cross-disciplinary teams, and leadership commitment.

5.5 Lesson 5 - Ethical, Legal, and Societal Risks Intensify with Domain Sensitivity. Risks were not uniform across pilots but increased substantially in contexts involving legal reasoning and democratic participation. PCS1 operated in a relatively low-risk environment, with ethical considerations focused on fairness and non-discrimination. PCS2 and PCS3 confronted high-stakes environments where misinterpretation of legal texts could create regulatory inconsistencies or undermine legal certainty. PCS4 sits at the highest risk tier, with potential classification as “high-risk AI” under the EU AI Act due to its impact on political participation, public discourse, and fundamental rights. The pilots show that risk is domain-dependent, and governance requirements must be proportionate to potential societal consequences. This underscores the need for continuous monitoring, documentation, and safeguards tailored to each application area.

5.6 Lesson 6 - AI Adoption Requires New Service Designs, Not Just New Tools. The pilots revealed that AI systems cannot simply be inserted into existing workflows; rather, work processes must be reconfigured to integrate AI responsibly. PCS1 required new analytical workflows and interpretation guidelines. PCS2 and PCS3 required redesigned legal research processes, including citation validation, escalation paths, and audit mechanisms. PCS4 required rethinking facilitation practices, multilingual participation methods, and interaction formats. This insight validates Phase 2 of the AIGOV Transformation and Adoption Framework: AI adoption must include deliberate redesign of services, processes, and governance ecosystems.

5.7 Lesson 7 - AI Can Generate Public Value, but Only When Properly Contextualised. Despite the challenges, the pilots demonstrated tangible benefits. PCS1 improved analytical capacity for evidence-based housing policy. PCS2 and PCS3 expanded access to legal information and provided rapid, contextualised interpretations using authoritative documents. PCS4 showed how generative AI could increase accessibility, inclusiveness, and shared understanding in multilingual democratic settings. These impacts materialise only when AI systems are embedded within responsible design principles and aligned with organisational mandates, legal frameworks, and democratic values.

5.8 Lesson 8 - The Three AIGOV Frameworks Are Mutually Reinforcing and Empirically Validated. Across pilots, the three AIGOV frameworks proved complementary. The Government Data Value Cycle clarified data governance needs and highlighted gaps. The Framework for Trustworthy, Fair, and Accountable AI guided risk assessment and ethical safeguards. The Transformation and Adoption Framework structured implementation into readiness, design, and controlled testing. All pilots relied on all three frameworks, demonstrating their applicability, completeness, and practical relevance for real public-sector AI scenarios. This confirms the conceptual decision in WP2 that trustworthy AI adoption must integrate data governance, ethical/technical requirements, and service transformation strategies into one coherent system.

5.9 Lesson 9 - Cross-Pilot Insights Indicate Pathways for Scaling and Future Research. The pilots collectively suggest several directions for further work:

- Development of shared public-sector AI testing infrastructures
- Creation of multilingual, legally compliant, domain-specific datasets
- Strengthening of AI governance capabilities within institutions
- Mechanisms for benchmarking LLM performance in sensitive contexts
- Deeper integration of explainability tools adapted to public-sector needs
- Harmonisation with the EU AI Act and sector-specific regulatory requirements

These directions should guide future iterations of the AIGOV frameworks, upcoming pilots, and policy recommendations.

6 Conclusions

The evaluation activities carried out in D3.1 provided a comprehensive assessment of the four pilot case studies and their relevance for the responsible adoption of AI in the public sector. The results demonstrate that each case study offered distinct insights into technical feasibility, interpretability, fairness and regulatory considerations, while also highlighting cross-cutting challenges that must be addressed in real governmental settings. Although PCS4 remained a conceptual case study without user testing, its analysis still contributed valuable perspectives on multilingual and multicultural engagement supported by Generative AI.

Across all case studies, several common themes emerged. High-quality and well-governed data were consistently identified as prerequisites for trustworthy AI applications. Interpretability and transparency proved essential for building confidence among public-sector stakeholders. The need for human oversight and clearly defined validation processes was evident in every domain examined. Furthermore, the work underscored that organisational readiness, including skills, workflows and governance mechanisms, is integral to effective AI adoption.

Overall, D3.1 confirms that while AI technologies hold significant promise for supporting public-sector decision-making and service provision, their deployment must be guided by careful evaluation, ethical principles and a commitment to accountability. The lessons learned from the case studies provide actionable guidance for future implementation efforts and reinforce the importance of a structured, responsible approach to integrating AI into government operations.

References

1. Addis, A. (2007) Constitutionalizing deliberative democracy in multilingual societies, Berkeley Journal of International Law 25 117.
2. Androutsopoulou, A., Karacapilidis, N., Loukis, E., Charalabidis, Y.: Transforming the communication between citizens and government through AI-guided chatbots. Government Information Quarterly 36(2), 358–367 (2019). https://doi.org/10.1016/j.giq.2018.10.001
3. Baek, Y. M., Wojcieszak, M., Carpini, M. X. D. (2012) Online versus face-to-face deliberation: Who? why? what? with what effects?, New Media & Society 14 363-383.
4. Campbell, J. Y., & Cocco, J. F. (2007). How do house prices affect consumption? Evidence from micro data. *Journal of monetary Economics*, *54*(3), 591-621. https://doi.org/10.1016/j.jmoneco.2005.10.016
5. Chami, I., Abu-El-Haija, S., Perozzi, B., Ré, C., & Murphy, K. (2022). Machine learning on graphs: A model and comprehensive taxonomy. *Journal of Machine Learning Research*, *23*(89), 1-64.
6. Chen, J.; Ma, T.; Xiao, C. FastGCN: Fast Learning with Graph Convolutional Networks via Importance Sampling. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
7. Chui, M., Hazan, E., Roberts, R., Singla, A., Smaje, K.: The economic potential of Generative AI (2023)
8. Égert, B., Mihaljek, D. Determinants of House Prices in Central and Eastern Europe. *Comp Econ Stud* **49**, 367–388 (2007). https://doi.org/10.1057/palgrave.ces.8100221
9. Fitsilis, F., Theodorakopoulos, G.: Better regulation and its evolution in the Hellenic legislative and parliamentary system. Statute Law Rev. 45(1), hmae003 (2024).
10. Hevner, A., Chatterjee, S. (2010) Design research in information systems: theory and practice, volume 22, Springer Science & Business Media.
11. Hoesli M, Malle R (2022), "Commercial real estate prices and COVID-19". *Journal of European Real Estate Research*, Vol. 15 No. 2 pp. 295–306, doi: https://doi.org/10.1108/JERER-04-2021-0024
12. Hromada, E., Čermáková, K., & Piecha, M. (2022). Determinants of house prices and housing affordability dynamics in the Czech Republic. *European Journal of Interdisciplinary Studies*, *14*(2), 119-132. https://doi.org/10.24818/ejis.2022.24
13. Jiang, W., & Luo, J. (2022). Graph neural network for traffic forecasting: A survey. Expert *systems with applications*, *207*, 117921.
14. Kalampokis, E., Tambouris, E., & Tarabanis, K. (2011). A classification scheme for open government data: towards linking decentralised data. *International Journal of Web Engineering and Technology*, *6*(3), 266-285.
15. Kalampokis, E., Karacapilidis, N., Tsakalidis, D., Tarabanis, K.: Understanding the use of emerging technologies in the public sector: A review of horizon 2020 projects. Digit. Gov.: Res. Pract. 4(1) (Apr 2023). https://doi.org/10.1145/3580603

16. Kalampokis, E., Karacapilidis, N., Karamanou, A., Tarabanis, K. (2024) Fostering Multilingual Deliberation through Generative Artificial Intelligence, IFIP EGOV-CeDEM-ePart2024 (EGOV2024), CEUR, Vol.3737.
17. Karamanou, A., Brimos, P., Kalampokis, E., & Tarabanis, K. (2024). Explainable Graph Neural Networks: An Application to Open Statistics Knowledge Graphs for Estimating House Prices. Technologies, 12(8), 128. https://doi.org/10.3390/technologies12080128
18. Karamanou, A., Kalampokis, E., & Tarabanis, K. (2022). Linked open government data to predict and explain house prices: the case of Scottish statistics portal. *Big Data Research*, *30*, 100355. https://doi.org/10.1016/j.bdr.2022.100355
19. Karamanou, A., Kalampokis, E., & Tarabanis, K. (2023). Integrated statistical indicators from Scottish linked open government data. *Data in brief*, *46*, 108779.
20. Karamanou, A., Kalampokis, E., Fitsilis, F., Theodorakopoulos, G., & Tarabanis, K. (2025, August). Evaluating Open and Proprietary Large Language Models in Law Interpretation: The Case of the EU VAT Directive. In *International Conference on Electronic Government* (pp. 380-394). Cham: Springer Nature Switzerland.
21. Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations, Banff, AB, Canada, 14-16 April 2014.
22. Kiwelekar, A. W., Mahamunkar, G. S., Netak, L. D., & Nikam, V. B. (2020). Deep learning techniques for geospatial data analysis. In *Machine learning paradigms: advances in deep* learning-*based technological applications* (pp. 63-81). Cham: Springer International Publishing.
23. Lin, S., Hilton, J., Evans, O.: Truthfulqa: Measuring how models mimic human falsehoods. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 3214–3252 (2022).
24. Listorti, G., Basyte-Ferrari, E., Acs, S., Smits, P.: Towards an evidence-based and integrated policy cycle in the EU: a review of the debate on the better regulation agenda. JCMS: J. Common Market Stud. **58**(6), 1558–1577 (2020).
25. Longo. E., The european citizens' initiative: too much democracy for eu polity?, German Law Journal 20 (2019) 181–200. https://doi.org/10.1017/glj.2019.12
26. von Lucke, J., Fitsilis, F., Etscheid, J.: Research and development agenda for the use of AI in parliaments. In: Proceedings of the 24th Annual International Conference on Digital Government Research, pp. 423–433 (2023).
27. Mamalis, M.E., Kalampokis, E., Fitsilis, F., Theodorakopoulos, G., Tarabanis, K.: A large language model agent based legal assistant for governance applications. In: International Conference on Electronic Government. pp. 286–301. Springer (2024).
28. Mbah, R. E., & Wasum, D. . (2022). Russian-Ukraine 2022 War: A Review of the Economic Impact of Russian-Ukraine Crisis on the USA, UK, Canada, and Europe. *Advances in Social Sciences Research Journal*, *9*(3), 144–153. https://doi.org/10.14738/assrj.93.12005

29. Maslaris, I., Karamanou, A., Kalampokis, E., Tarabanis, K.: Evaluating large language models in interaction with open government data. In: Proceedings of the 28th Pan-Hellenic Conference on Progress in Computing and Informatics. p. 26–33. PCI ’24, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3716554.3716558
30. Mehandru, N., Miao, B.Y., Almaraz, E.R., Sushil, M., Butte, A.J., Alaa, A.: Evaluating large language models as agents in the clinic. NPJ digital medicine 7(1), 84 (2024).
31. Morano, P., Tajani, F., Guarini, M. R., Di Liddo, F., & Anelli, D. (2019, June). A multivariate econometric analysis for the forecasting of the interdependences between the housing prices and the socio-economic factors in the city of Barcelona (Spain). In *International Conference on Computational Science and Its Applications* (pp. 13-22). Cham: Springer International Publishing.
32. van Noordt, C., Misuraca, G.: Artificial intelligence for the public sector: Results of landscaping the use of ai in government across the European Union. Government Information Quarterly 39(3), 101714 (2022). https://doi.org/10.1016/j.giq.2022.101714
33. Patten, A. (2007). Theoretical foundations of european language debates, in: D. Castiglione, C. Longman (Eds.), The Language Question in Europe and Diverse Societies: Political, Legal and Social Perspectives, Hart Publishing, London, 2007.
34. Pereira, P., Zhao, W., Symochko, L., Inacio, M., Bogunovic, I., & Barcelo, D. (2022). The Russian-Ukrainian armed conflict will push back the sustainable development goals. *Geography and Sustainability*, *3*(3), 277-287. https://doi.org/10.1016/j.geosus.2022.09.003
35. Peffers, K., Tuunanen, T. , Rothenberger, M. A., Chatterjee, S. (2007) A design science research methodology for information systems research, Journal of management information systems 24, 45-77.
36. Pulapaka, S., Godavarthi, S., Ding, S.: Empowering the public sector with generative AI (2024).
37. Rapoport, R. N. (1970) Three dilemmas in action research: with special reference to the tavistock experience, Human relations 23, 499-513.
38. Ringe, N. (2022) The Language(s) of Politics: Multilingual Policy-Making in the European Union, University of Michigan Press.
39. Siino, M., Falco, M., Croce, D., Rosso, P.: Exploring llms applications in law: A literature review on current legal nlp approaches. IEEE Access (2025).
40. Sirait, E., Zuiderwijk, A., Janssen, M.: The readiness of the public sector to implement AI: A government-specific framework. In: Janssen, M., Crompvoets, J., Gil-Garcia, J.R., Lee, H., Lindgren, I., Nikiforova, A., Viale Pereira, G. (eds.) Electronic Government. pp. 302–316. Springer Nature Switzerland, Cham (2024).
41. Solove, D.J., Schwartz, P.M.: EU Data Protection and the GDPR. Aspen Publishing (2023).

42. Song, Y., Ma, X. Exploration of intelligent housing price forecasting based on the anchoring effect. *Neural Comput & Applic* **36**, 2201–2214 (2024). https://doi.org/10.1007/s00521-023-08823-3
43. Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing price prediction via improved machine learning techniques. *Procedia Computer Science*, *174*, 433-442.
44. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
45. Wei, J., et al.: Emergent abilities of large language models. arXiv preprint arXiv:2206.07682 (2022)
46. Wu, Y., Lian, D., Xu, Y., Wu, L., & Chen, E. (2020). Graph Convolutional Networks with Markov Random Field Reasoning for Social Spammer Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(01), 1054-1061. https://doi.org/10.1609/aaai.v34i01.5455
47. Wu, S., Sun, F., Zhang, W., Xie, X. and Cui, B. (2022). Graph Neural Networks in Recommender Systems: A Survey. *ACM Comput. Surv*. 55, 5, Article 97 (May 2023), 37 pages. https://doi.org/10.1145/3535101
48. Yang, L., Chu, X., Gou, Z., Yang, H., Lu, Y., & Huang, W. (2020). Accessibility and proximity effects of bus rapid transit on housing prices: Heterogeneity across price quantiles and space. *Journal of Transport Geography*, *88*, 102850.
49. Ying, Z.; Bourgeois, D.; You, J.; Zitnik, M.; Leskovec, J. Gnnexplainer: Generating explanations for graph neural networks. Adv. Neural Inf. Process. Syst. 2019, 32, 9244–9255.
50. Yin, R.K.: Case study research: Design and methods, vol. 5. sage (2009).
51. Zeginis, D., Kalampokis, E., Tarabanis, K.: Applying an ontology-aware zero-shot LLM prompting approach for information extraction in greek: the case of diavgeia.gov gr. In: Proceedings of the 28th Pan-Hellenic Conference on Progress in Computing and Informatics. p. 324–330. PCI ’24, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3716554.3716603