



Università degli Studi di Genova
Scuola Politecnica & Facoltà di Ingegneria



Laurea Magistrale in Ingegneria Strategica

**International MSc in Engineering Technology for Strategy and Security
of Genoa University**

Trustworthy AI Decision-Support and Simulation Framework for Municipal Governance:

A Case Study on Comune di Genova

Candidate: **Nabila Rahman**

Advisor: **Prof. Engr Marco Gotelli**

Student ID / Matricola: **6814546**

Academic Year 2024
STRATEGOS 8th Edition



Acknowledgements

Alhamdulillah. All praise and gratitude belong to **Allah (SWT)**, the Most Merciful and the Most Compassionate. Throughout this journey, He has been my source of strength, patience and hope. I am grateful to Allah for guiding me through both good and difficult times, for giving me the courage to continue when things felt uncertain and for blessing me with what I needed along the way. There were moments of struggle, disappointment, and exhaustion when moving forward was not easy. Yet, through His mercy and blessings, I found the strength to carry on. Whatever I have achieved today would not have been possible without His guidance and grace.

Completing this master's thesis has been a challenging experience. The difficulties I faced along the way tested my patience and determination, sometimes more than I expected. Not every step was easy, but each experience taught me something valuable. Looking back, I realize how much this journey has shaped me, not only as a student but also as a person.

I would also like to acknowledge my thesis & Internship supervisor Marco Gotelli and Nicola Iacobone also the faculty members of my university for their academic guidance and contributions to my education and research. I would also like to pay tribute to the late Professor Agostino Bruzzone, whose vision and pioneering work helped establish the foundation of our academic discipline. His contributions have left a legacy that continues to inspire students, including myself.

My gratitude also goes to my family and specially my husband Davide Lunetta for his prayers, patience and support throughout my studies. I am thankful to those who showed me kindness, offered encouragement, or helped me during difficult times. Sometimes, even a small gesture made a meaningful difference.

Reaching this point has required patience, perseverance and the determination to keep going despite the challenges. To me, this thesis is more than an academic achievement. It represents the difficulties I have faced, the lessons I have learned and the effort it took to reach this stage of my life.



Abstract

Municipal administrations face increasing demand for timely citizen services while remaining accountable for decisions made under complex legal and procedural requirements. This thesis develops and evaluates a trustworthy artificial intelligence (AI) decision-support and simulation framework for request handling at the Comune di Genova. The approach combines a multi-agent retrieval-augmented generation pipeline with operational simulation and controls designed to support human oversight, explainability, privacy protection and auditability. Citizen requests are pseudonymised, matched to municipal regulations through hybrid retrieval and evaluated by specialised agents responsible for evidence retrieval, compliance, risk, response generation and confidence assessment. A decision gate distinguishes requests suitable for an automatic informational reply from those requiring officer review or urgent dispatch. Eligibility decisions remain under human authority. A SimPy discrete-event model represents arrivals, triage, departmental queues, misrouting, rework and service deadlines; it compares four operating policies under eight scenarios and is complemented by staffing and demand models. The evaluation uses municipal requests, independent challenge sets and historical cases, while recognising limitations in the diversity of available texts and in assumed operational parameters. On noisy requests, routing accuracy reaches 84.2%; the decision gate reduces the proportion of incorrect responses reaching citizens without human review to 3.3%, while approximately 70% of independent requests are answered automatically. Under simulated baseline conditions, average response time decreases from 1.19 to 0.28 working days. The findings indicate that AI-assisted drafting and supervised replies contribute more to operational improvements than routing automation alone. They also show the limits of automation during disasters and infrastructure incidents, when safety-critical demand still requires contingency staffing. The resulting framework provides a basis for measuring operational benefits alongside trustworthiness requirements. It is an evaluated prototype, not a production deployment: further testing with live correspondence, generative backends and municipal staff is required before operational adoption.



TABLE OF CONTENTS

Table of Contents

Acknowledgements.....	2
Abstract.....	3
TABLE OF CONTENTS.....	4
INTRODUCTION.....	9
CHAPTER 1 – Research Context, Motivation, and Objectives.....	10
1.1 <i>Public-Sector Digital Transformation in Italy and the European Union</i>	10
1.1.1 From e-Government to Digital-Era Governance	10
1.1.2 The Italian Digital Public Administration Ecosystem	10
1.1.3 Structural Bottlenecks in Municipal Request Handling.....	11
1.2 <i>The Comune di Genova: Organisation, Services and Citizen Channels</i>	12
1.2.1 Territorial and Institutional Profile.....	12
1.2.2 Departments and Services in Scope.....	12
1.2.3 Citizen Channels and the Nature of Inbound Demand.....	13
1.2.4 Evidence of Service-Level Strain in the Historical Record.....	14
1.3 <i>Regulatory Framework</i>	14
1.3.1 The EU Artificial Intelligence Act (Regulation (EU) 2024/1689)	15
1.3.2 The General Data Protection Regulation (Regulation (EU) 2016/679).....	16
1.3.3 The Italian Law on Artificial Intelligence (Law 132/2025).....	17
1.3.4 Italian Administrative Procedure Law (Law 241/1990) and Algorithmic Administration	17
1.3.5 Synthesis: Regulation as Architecture.....	18
1.4 <i>Technical Framework</i>	18
1.4.1 Large Language Models and Their Limitations in Public Administration.....	18
1.4.2 Retrieval-Augmented Generation	19
1.4.3 Multi-Agent Architectures	19
1.4.4 Discrete-Event Simulation for Strategic Decision Support	20
1.4.5 Trustworthy AI as an Engineering Discipline	20
1.5 <i>Market and State of the Art</i>	21
1.5.1 Municipal Conversational Agents.....	21



1.5.2 Decision-Support and Workflow Tools	21
1.5.3 Gap Analysis	21
1.6 Problem Statement and Requirements	22
1.6.1 Problem Statement	22
1.6.2 Data Context and Its Implications	23
1.6.3 Functional Requirements	23
1.6.4 Non-Functional Requirements	24
1.6.5 Scope and Delimitations	24
CHAPTER 2 – INTERNSHIP AT THE COMUNE DI GENOVA	25
2.1 Host Organisation and Internship Setting	25
2.2 Objectives and Methodology	25
2.3 Preparatory Study: AI in European Public Administration	27
2.4 Analysis of Administrative Processes	28
2.5 Conceptual Design of AI-Assisted Services	28
2.5.1 Administrative and Citizen-Information Chatbot	28
2.5.2 LLM-Assisted E-mail Response System	28
2.6 Business Intelligence Project for Zenia	30
2.6.1 Data and Platform	30
2.6.2 Medallion Architecture	30
2.6.3 Dashboards	31
2.6.4 Findings and Predictive Opportunities	33
2.7 Contribution of the Internship to the Thesis	34
2.8 Limitations of the Internship Work	35
CHAPTER 3 – SIMULATION MODEL	36
3.1 Strategic Engineering and Modelling & Simulation as Decision Support	36
3.2 Conceptual Model of Municipal Request Handling	37
3.3 Discrete-Event Simulator Implementation	38
3.3.1 Process-Oriented Modelling in SimPy	38
3.3.2 Arrival Process	39
3.3.3 Service-Time Distributions	39
3.3.4 Routing, Prioritisation and Rework Logic	40
3.3.5 Performance Indicators	40
3.3.6 Replications and Statistical Treatment	40
3.4 Calibration	41
3.4.1 Parameters Estimated from Data	41
3.4.2 Assumption Register	42
3.5 Scenario Generation	43
3.6 Operating Policies	43



3.7 Resource Model: Erlang-C Staffing Optimisation.....	44
3.7.1 The M/M/c Model.....	44
3.7.2 Department Load under a Policy	44
3.7.3 Optimisation Problem	44
3.8 What-If Analysis, Demand Forecasting and Risk Forecasting.....	45
3.8.1 What-If Analysis.....	45
3.8.2 Demand Forecasting	45
3.8.3 Risk Forecasting.....	45
3.9 Verification and Validation.....	46
CHAPTER 4 – TRUSTWORTHY AI DECISION-SUPPORT FRAMEWORK: DESIGN AND IMPLEMENTATION.....	48
4.1 Overall Architecture.....	48
4.2 Datasets and Data Acquisition.....	49
4.2.1 Datasets	49
4.2.2 Acquisition Connectors.....	49
4.3 Preprocessing, Privacy Protection and Data-Quality Audit.....	50
4.3.1 Cleaning and Normalisation	50
4.3.2 Language Detection.....	50
4.3.3 PII Masking and Pseudonymisation	51
4.3.4 Metadata Extraction.....	51
4.3.5 Validation and Duplicate Handling	51
4.3.6 Data-Quality Findings	52
4.4 Semantic Knowledge Construction.....	52
4.4.1 Municipal Ontology.....	52
4.4.2 Tokenisation and Chunking	52
4.4.3 Embedding Generation.....	53
4.4.4 Knowledge Graph.....	53
4.5 Municipal Knowledge Repository	53
4.5.1 Vector Database	53
4.5.2 Metadata and Document Stores	54
4.5.3 Role-Based Access Control	54
4.5.4 Encrypted PII Vault	54
4.6 Hybrid Retrieval Pipeline.....	54
4.6.1 Query Understanding.....	55
4.6.2 Query Expansion	55
4.6.3 Hybrid Candidate Generation	55
4.6.4 Re-ranking.....	56
4.6.5 Context Compression	56
4.7 Multi-Agent Decision Engine.....	56
4.7.1 Retrieval Agent	57
4.7.2 Policy and Compliance Agent.....	57
4.7.3 Risk Analysis Agent	58



4.7.4 Response Generation Agent	58
4.7.5 Confidence and Evaluation Agent	58
4.8 Governance Layer	59
4.8.1 Hallucination Detection and Faithfulness	59
4.8.2 Explainability	60
4.8.3 Fairness Monitoring and Bias Audit	60
4.8.4 Regulatory Compliance Checks	60
4.8.5 Human-in-the-Loop Review	61
4.8.6 Tamper-Evident Audit Log	61
4.9 Implementation and Technology Stack	61
CHAPTER 5 – EXPERIMENTAL EVALUATION AND RESULTS	63
5.1 Evaluation Design	63
5.1.1 Test Sets	63
5.1.2 Configuration	63
5.1.3 Metrics	63
5.2 Retrieval Performance	64
5.3 Routing and Decision-Support Quality	65
5.3.1 Routing Accuracy and Handling Decisions	65
5.3.2 Error Analysis	66
5.3.3 Ablation Study	66
5.3.4 Urgency Detection	67
5.4 Response Quality	67
5.5 Trustworthiness	68
5.5.1 Faithfulness and Hallucination Detection	68
5.5.2 Calibration	68
5.6 Explainability and Fairness	69
5.6.1 Explainability	69
5.6.2 Fairness of Automated Handling	69
5.6.3 Audit of Historical Decisions	71
5.7 Simulation Results	71
5.7.1 Policy Impact under Baseline Conditions	71
5.7.2 Scenario Analysis	72
5.7.3 What-If Analysis and Risk Forecast	73
5.7.4 Staffing Optimisation	75
5.7.5 Demand Forecast and Strategic Impact	76
5.8 System Efficiency and Scalability	76
5.9 Regulatory Compliance Assessment	77
CHAPTER 6 – STRATEGIC DECISION-SUPPORT DASHBOARD AND DEPLOYMENT	79
6.1 Requirements of Municipal Decision Makers	79



6.2 Dashboards and Views	80
6.2.1 Request Desk	80
6.2.2 Overview and Strategic Insights	81
6.2.3 Scenario Comparison, What-If and Risk	82
6.2.4 Staffing Planner	83
6.2.5 Governance, Compliance and Audit	84
6.3 Human Review and Approval Workflow	85
6.4 Strategic Outputs	87
6.4.1 Policy Recommendations	87
6.4.2 Risk Alerts	87
6.4.3 Resource Suggestions and Action Plans	87
6.5 Deployment	88
6.5.1 Deployment Architecture	88
6.5.2 Integration Interfaces	88
6.5.3 Configuration of the Generator	89
6.5.4 Roll-Out Roadmap	89
6.5.5 Operational Governance	90
CONCLUSIONS	90
TECHNICAL & SCIENTIFIC REFERENCES	91



INTRODUCTION

Digital access has made it easier to contact municipal administrations, but requests still require interpretation, assignment, regulatory checks and timely replies. When triage and information retrieval depend on manual work, rising demand can increase delays, unnecessary hand-offs and the risk of missed deadlines. These challenges are relevant to the Comune di Genova, where citizen requests reach several departments through web forms, e-mail and certified e-mail (PEC).

Artificial intelligence can support classification, regulatory retrieval and response drafting. For public administration, however, speed and fluency are not sufficient. Replies must be grounded in applicable rules, personal data must be protected, and accountable officials must retain control of consequential decisions. Operational benefits also need to be measured: faster triage alone may have little effect if departmental queues remain unchanged.

This thesis develops and evaluates a trustworthy AI decision-support and simulation framework for municipal request handling. Its bilingual pipeline combines a municipal knowledge repository, hybrid retrieval, specialised agents and a decision gate for automatic informational replies, officer review and urgent dispatch. It incorporates regulatory grounding, auditability, fairness monitoring, data protection and human oversight. Eligibility decisions remain with public officials.

A discrete-event model of the municipal workflow evaluates alternative handling policies under baseline and adverse conditions. Demand forecasts, risk analysis and Erlang-C staffing estimates link system behaviour to operational planning. The evaluation draws on municipal requests, regulations, historical cases and independent challenge sets. Unmeasured staffing and processing parameters are treated as assumptions rather than observed facts.

The research also draws on internship activities at the Comune di Genova, including analysis of administrative processes and work on AI-assisted service concepts. Chapter 1 establishes the context and research questions; Chapter 2 describes the internship; Chapters 3 and 4 present the simulation and AI architecture; Chapter 5 reports the evaluation; and Chapter 6 examines the dashboard, governance workflow and deployment requirements.



CHAPTER 1 – Research Context, Motivation, and Objectives

1.1 Public-Sector Digital Transformation in Italy and the European Union

1.1.1 From e-Government to Digital-Era Governance

European public administration has moved through three broad stages of digitalisation. The first, generally called e-Government, transferred paper-based procedures to electronic channels without fundamentally changing how administrations worked. Forms became downloadable, offices acquired e-mail addresses, and most internal processing remained manual. The second stage, which Dunleavy et al. (2006) describe as Digital-Era Governance, sought to reconnect services fragmented under New Public Management, organise provision around citizens' needs, and simplify contact with public offices through digital platforms. The current stage is increasingly data-driven and AI-assisted. Administrations are expected not only to provide digital services but also to use the resulting data to anticipate demand, allocate resources and inform decisions (Mergel, Edelmann and Haug, 2019).

This shift changes where administrative discretion is exercised. Bovens and Zouridis (2002) describe the transition from street-level bureaucracy, in which front-line officials judge individual cases, to screen-level and ultimately system-level bureaucracy, where decisions are increasingly shaped by software. While these changes can improve efficiency and consistency, they can also make responsibility harder to trace. AI-assisted municipal request handling must therefore deliver operational improvements without undermining accountability, explainability or the ability to challenge decisions. That tension informs the design developed in this thesis.

1.1.2 The Italian Digital Public Administration Ecosystem

Italy has built, over the last two decades, a dense legal and infrastructural foundation for digital administration. The *Codice dell'Amministrazione Digitale* (CAD, Legislative Decree 82/2005) establishes the citizen's right to interact with public administrations through digital means and imposes the *digital-first* principle on all administrative bodies. Several national platforms implement this principle and are directly relevant to municipal request handling:

- **PEC (Posta Elettronica Certificata)**, the certified e-mail system that gives a digital message the legal value of a registered letter with return receipt. A large share of formal citizen and business communication with municipalities arrives through PEC, and its legal weight means that every PEC submission initiates or modifies an administrative procedure with legally relevant timestamps.
- **SPID and CIE**, the public digital identity system and the electronic identity card, which authenticate citizens on municipal web portals.
- **ANPR (Anagrafe Nazionale della Popolazione Residente)**, the national resident population register, which centralises civil-registry data previously held separately by each municipality.
- **pagoPA** and the **IO** application, which provide standardised payment and notification channels between administrations and citizens.



AgID coordinates this ecosystem through the *Agenzia per l'Italia Digitale* (AgID) through the three-year plan for public-administration informatics (*Piano Triennale per l'Informatica nella Pubblica Amministrazione*). The 2024–2026 edition of the plan explicitly anticipates guidelines on the adoption of artificial intelligence in public administration; AgID opened a public consultation on the adoption component of those guidelines on 18 February 2025, closing on 20 March 2025 (Agenzia per l'Italia Digitale, 2025), with the broader guidelines on adopting, procuring and developing AI to be issued under the procedure of Article 71 of the CAD (AgID, 2025).

Mission 1, Component 1 of the Piano Nazionale di Ripresa e Resilienza (PNRR) has directed substantial funding towards public-administration digitalisation, including cloud migration and adoption of national platforms by municipalities. Yet receiving requests digitally does not necessarily mean processing them digitally. In many municipalities, an operator still reads each PEC message or web form, determines which department should handle it, and passes it to an officer who must identify the relevant regulation before replying. Digital channels have made requests easier to submit, while much of the administrative effort remains unchanged. This gap is the focus of the thesis.

1.1.3 Structural Bottlenecks in Municipal Request Handling

Four recurring bottlenecks in Italian municipal request handling shaped the framework's design.

(i) Intake and triage load. Digital channels have lowered the threshold for contacting the administration, so inbound volume grows faster than front-office capacity. Each request must be read, classified and routed before substantive work begins. When triage is manual, its cost scales linearly with volume. A queueing system in which the arrival rate λ approaches the service capacity $c\mu$ exhibits waiting times that grow without bound as the utilisation

$$\rho = \frac{\lambda}{c\mu}$$

tends to one. This situation can arise when surges such as a waste-tax billing run or a weather alert push a municipal front office.

(ii) Misrouting and inter-departmental hand-offs. Municipal competences are distributed across departments whose boundaries are not obvious to citizens. A noise complaint about a nightclub may concern the Local Police (*Polizia Locale*), the Environment department (*Ambiente*) or the Commerce office (*Commercio*), depending on whether the issue is enforcement, acoustic measurement or licensing. Every misrouted file incurs a transfer delay and duplicated reading effort. By Little's law,

$$L = \lambda W$$

any increase in the mean time in system W caused by transfers translates directly into a larger stock of open files L .

(iii) Regulatory retrieval cost. Answering a citizen correctly requires locating the applicable regulation article, the procedural deadline and the competent office. Municipal regulations are long, heterogeneous in structure and written in specialised administrative Italian. Officers repeatedly perform the same retrieval work for recurring questions, and the quality of answers varies with individual experience.

(iv) Deadline compliance. Italian administrative law imposes binding response deadlines (Section 1.3.4). When workload exceeds capacity, deadlines are breached; the breach is not



merely a service-quality problem but a legal one, exposing the administration to complaints, silence-related remedies and loss of public trust.

These problems are interconnected. Incorrect triage creates additional hand-offs; transfers consume staff time; reduced capacity lengthens queues; and longer queues raise the risk of missed deadlines. Improving one task in isolation may therefore have little effect on the service as a whole. The thesis addresses this interaction by coupling its AI decision-support engine with a discrete-event simulation of municipal request handling (Chapter 3).

1.2 The Comune di Genova: Organisation, Services and Citizen Channels

1.2.1 Territorial and Institutional Profile

Genoa is the capital of the Liguria region and one of Italy's principal port cities, with approximately 560,000 residents. Administratively, the city is divided into nine *Municipi* (decentralised district councils), which deliver part of the front-office services while central departments retain regulatory and decision-making competence. The territory is long and narrow, compressed between the Ligurian Sea and the Apennines and crossed by short, steep watercourses, most notably the Bisagno and the Polcevera. This morphology gives the city an exceptional exposure to hydrogeological risk: the flash floods of November 2011 and October 2014 in the Bisagno valley caused fatalities and extensive damage, and they shaped the municipal civil-protection planning referenced in the regulation corpus used in this work (REG-010 and REG-015). The collapse of the Morandi bridge on 14 August 2018 (Harding, 2018) further demonstrated how infrastructure incidents can abruptly reshape service demand, mobility and inter-departmental coordination across the municipality.

Genoa also has a distinctive demographic profile. Its population is among the oldest of Italy's large cities. This affects demand for social services, transport concessions for older residents and cemetery services, and makes accessible, equitable service delivery across age groups particularly important. Chapter 5 tests whether automated handling treats age groups fairly and examines historical cases for evidence of age-related disparities.

1.2.2 Departments and Services in Scope

The thesis draws on datasets covering fifteen service categories, ten municipal departments and fifteen municipal regulations comprising forty-five articles. Table 1.1 links each category to the responsible department and its governing regulation, using the knowledge-graph structure developed in Chapter 4.

Table 1.1 – Service categories, competent departments and governing regulations

Service category	Competent department	Governing regulation
Residency change	Anagrafe (Civil Registry)	REG-001 Regolamento Anagrafe
Building permit	Edilizia (Building Permits)	REG-002 Regolamento Edilizio Comunale
Waste tax (TARI)	Tributi (Taxation)	REG-003 Regolamento TARI
Parking / ZTL permit	Mobilità (Mobility & Transport)	REG-004 Regolamento Sosta e Parcheggi
Business licence	Commercio (Commerce & Licensing)	REG-005 Regolamento Attività Commerciali
Social benefit	Servizi Sociali (Social Services)	REG-006 Regolamento Servizi Sociali
Public-works complaint	Urbanistica (Urban Planning)	REG-007 Regolamento Lavori Pubblici e Segnalazioni



Service category	Competent department	Governing regulation
Environmental report	Ambiente (Environment)	REG-008 Regolamento Tutela Ambientale
Event authorisation	Polizia Locale (Local Police)	REG-009 Regolamento Manifestazioni ed Eventi Pubblici
Emergency report	Protezione Civile (Civil Protection)	REG-010 Piano di Protezione Civile Comunale
School enrolment	Servizi Sociali (Social Services)	REG-011 Regolamento Iscrizioni Scolastiche
Cemetery service	Anagrafe (Civil Registry)	REG-012 Regolamento Servizi Cimiteriali
Public-transport pass	Mobilità (Mobility & Transport)	REG-013 Regolamento Trasporto Pubblico Locale
Noise complaint	Polizia Locale (Local Police)	REG-014 Regolamento Inquinamento Acustico
Flood-risk inquiry	Protezione Civile (Civil Protection)	REG-015 Piano di Rischio Idrogeologico

The relationship between categories and departments is not one-to-one. Two departments, Social Services and the Civil Registry, each handle two categories, and several categories involve actors outside the municipal organisation: ARPAL (the regional environmental protection agency) for environmental reports, AMT (the Genoese public-transport company) for transport passes, the SUAP (*Sportello Unico per le Attività Produttive*) for business licences, and the regional administrative court (TAR) or the *difensore civico* for appeals. These external actors are represented as *Agency* nodes in the municipal ontology, because they determine procedural paths and response times.

1.2.3 Citizen Channels and the Nature of Inbound Demand

Citizens reach the Comune through three digital channels in the dataset: the municipal web portal, ordinary e-mail and PEC. The channels differ in legal weight and in the structure of what they deliver. Web-portal submissions are partially structured, because the form constrains the category. E-mail and PEC messages are free text, often mixing several issues, informal language, personal data and attachments. Figure 1.1 represents the as-is request-handling process that the framework seeks to augment.

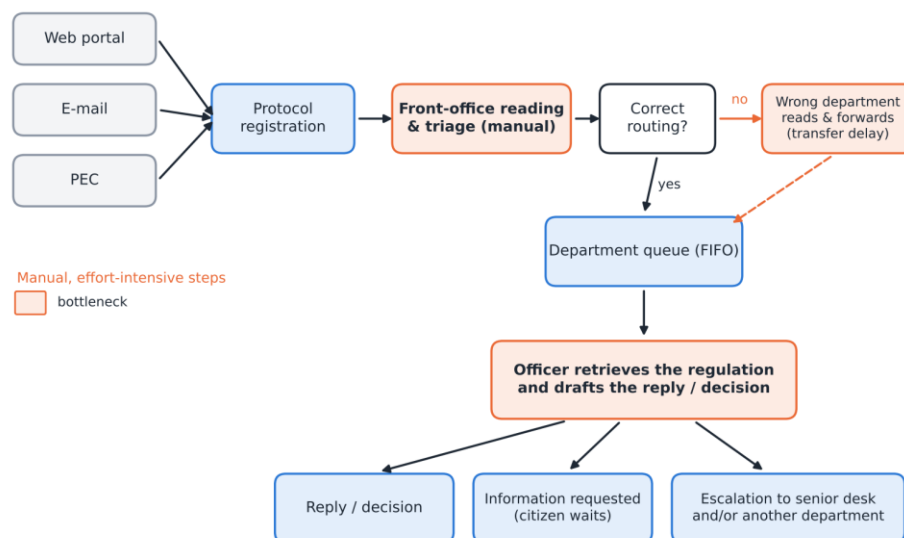


Figure 1.1: As-is request handling at the Comune di Genova (simplified); shaded boxes mark manual, effort-intensive steps



Two properties of inbound demand matter for system design. First, demand is **heterogeneous in urgency**. A question about school-enrolment dates and a report that the water of a torrent is rising in Val Bisagno enter through the same channels but require radically different response times; the civil-protection plan requires immediate activation of the municipal operations centre (*Centro Operativo Comunale*, COC) under orange or red alerts. Second, demand is **non-stationary**: it follows seasonal patterns (tourism-season peaks in events, noise and mobility requests), administrative calendars (tax billing runs, school-enrolment windows) and exogenous shocks (weather alerts, infrastructure incidents). A decision-support system must therefore support *strategic* questions (how many officers each department needs under a given scenario) as well as *operational* ones (how to answer this message).

1.2.4 Evidence of Service-Level Strain in the Historical Record

The historical case dataset (500 closed cases, 2022–2026) provides direct evidence of the strain described above. The median response time across categories is eleven days. More importantly, comparison with the deadlines stated in the municipal regulations reveals systematic breaches in time-critical services (Figure 1.2). Under the service-level targets defined in Chapter 3, every historical cemetery case exceeded the 48-hour target for urgent burials stated in REG-012. Ninety per cent of emergency reports exceeded a one-day target, 77% of public-works complaints exceeded the five-working-day inspection target for safety risks in REG-007, and 52% of environmental reports exceeded the ten-day joint-inspection target in REG-008. The remaining eleven services, whose targets range from 30 to 60 days, show no breaches.

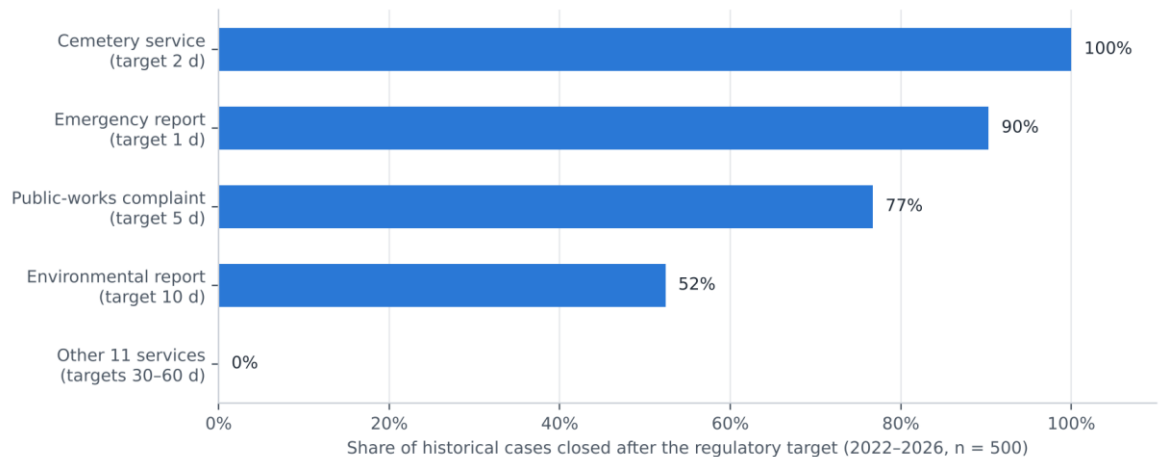


Figure 1.2: Share of historical cases closed after the regulatory target, by service (time-critical services shown individually)

These figures do not imply negligence: response time in the historical record includes waiting for citizen documents and external agencies. They show, however, that the services where delay has the highest human cost are those where the current process most often fails its own regulatory targets. This observation shapes the risk-analysis agent (Chapter 4) and the scenario design (Chapter 3).

1.3 Regulatory Framework

AI-assisted classification, routing and drafting of citizen requests must operate within a demanding legal framework. Four instruments have the most direct implications for the



proposed system: the EU Artificial Intelligence Act, the General Data Protection Regulation, Italy's AI legislation and the rules governing administrative procedure. Their requirements shape both the technical architecture and the human review process.

1.3.1 The EU Artificial Intelligence Act (Regulation (EU) 2024/1689)

The AI Act establishes a risk-based regulatory regime. It prohibits a limited set of practices (Article 5), imposes extensive obligations on *high-risk* systems (Chapter III), imposes transparency obligations on certain systems interacting with natural persons or generating content (Article 50), and sets specific duties for general-purpose AI models.

Risk classification. Under Article 6(2) and Annex III, point 5(a), AI systems intended to be used by or on behalf of public authorities *to evaluate the eligibility of natural persons for essential public assistance benefits and services*, or to grant, reduce, revoke or reclaim them, are classified as high-risk. Several categories in Table 1.1 (social benefits, waste-tax reductions, transport concessions, school-enrolment priorities) fall within this description *if* an AI system evaluates eligibility. The classification depends on the system's intended purpose, not on the technology used. A system that only routes requests, retrieves regulations and drafts *informational* replies, while the eligibility decision remains with an officer, has a materially different risk profile from one that decides. Article 6(3) excludes from high-risk status systems that perform narrow procedural tasks, improve the result of a previously completed human activity, or perform preparatory tasks, provided that they do not profile natural persons.

This thesis nevertheless adopts a deliberately conservative posture. Because the boundary between "preparatory" and "materially influencing" is contestable, and because drafts sent to citizens about eligibility-related services shape their expectations, the framework is designed to satisfy the high-risk obligations wherever its outputs touch eligibility-related services.

Obligations for high-risk systems. Articles 9 to 15 define the core technical obligations, summarised in Table 1.2 together with the design response adopted in this thesis.

Table 1.2 – AI Act high-risk obligations and corresponding design responses

Article	Obligation	Design response in the proposed framework
Art. 9	Risk-management system across the lifecycle	Per-request risk scoring and alerts; scenario-based risk forecasting by simulation
Art. 10	Data and data governance: relevance, representativeness, error and bias examination	Provenance recording; 23 automated data-quality checks; historical bias audit with multiple-testing correction
Art. 11	Technical documentation	Generated system card; documented assumptions; reproducible pipeline
Art. 12	Automatic recording of events (logging)	Hash-chained, tamper-evident audit log of every decision, access and review action
Art. 13	Transparency and information to deployers	Per-decision explanations with evidence citations and score contributions
Art. 14	Human oversight, including the ability to disregard or override outputs	Confidence- and risk-gated review queue; officers approve, edit or reject drafts; eligibility decisions reserved to officers
Art. 15	Accuracy, robustness and cybersecurity	Accuracy measured on independent and noisy test sets; hallucination detection; role-based access control

Transparency towards citizens (Article 50). Providers must ensure that natural persons are informed when they interact directly with an AI system, unless this is obvious from the context. In municipal correspondence, a citizen who receives a reply has no way of knowing whether it was drafted by a person or a machine. Every reply produced by the framework



therefore carries a fixed disclosure sentence and states that a human officer can review the request on demand.

Implementation timeline. The AI Act entered into force on 1 August 2024, with obligations phased in over several years. The timeline for high-risk obligations was revised by the Digital Omnibus on AI (Regulation (EU) 2026/1744), which entered into force on 27 July 2026. Under the amended schedule, the main obligations for stand-alone high-risk systems listed in Annex III apply from 2 December 2027, and those for high-risk systems embedded in products under Annex I legislation apply from 2 August 2028. For a municipality planning procurement and deployment over a 2026–2028 horizon, the high-risk obligations are therefore an imminent design constraint rather than a distant prospect. Building compliance into the architecture from the outset is substantially cheaper than retrofitting it after deployment.

1.3.2 The General Data Protection Regulation (Regulation (EU) 2016/679)

Citizen requests are rich in personal data: names, e-mail addresses, telephone numbers and tax codes (*codice fiscale*). They also frequently contain special-category or sensitive information, such as health conditions behind a disability parking request, economic hardship behind a social-benefit request, or a bereavement behind a burial request. Four provisions of the GDPR are decisive for the framework's design.

Article 5 – Principles. Processing must be lawful, fair and transparent, limited to specified purposes, minimised to what is necessary, accurate, retained only as long as necessary and secured. Data minimisation (Article 5(1)(c)) has a direct architectural consequence. The AI engine does not need a citizen's identity to classify a request or retrieve a regulation; it needs the *content* of the request. The framework therefore masks direct identifiers before any text reaches the retrieval and generation components.

Article 22 – Automated individual decision-making. Data subjects have the right not to be subject to a decision based *solely* on automated processing that produces legal effects concerning them or similarly significantly affects them. Decisions on social benefits, tax reductions or permits clearly produce such effects. The framework addresses Article 22 structurally: it never issues eligibility decisions, and every reply concerning an eligibility-related service states that the decision rests with a municipal officer. In the sense of Wachter, Mittelstadt and Floridi (2017), the system provides *ex ante* information about how requests are handled and *ex post* explanations of each individual output, without depending on a contested "right to explanation".

Article 25 – Data protection by design and by default. Controllers must implement appropriate technical and organisational measures, such as pseudonymisation, at the time of design. The framework applies keyed pseudonymisation (HMAC-SHA-256) to citizen identifiers, so that requests from the same person can be linked for service purposes without exposing identity to the analytical components.

Article 32 – Security of processing. Appropriate security measures include pseudonymisation and encryption, confidentiality and integrity of systems, and the ability to restore availability. The framework stores direct identifiers in a separate, encrypted vault readable only by the role of Data Protection Officer. Every access attempt, granted or denied, is written to the audit log.



1.3.3 The Italian Law on Artificial Intelligence (Law 132/2025)

Italy complemented the AI Act with a national framework law, Law no. 132 of 23 September 2025, published in the *Gazzetta Ufficiale* on 25 September 2025. Article 14 of the law addresses using AI in public administration and is particularly relevant to this thesis. It frames AI as an *instrumental* tool to increase administrative efficiency, reduce processing times and improve services to citizens and businesses, rather than as a substitute for administrative decision-making. The provision maintains that the official who adopts an act retains full responsibility for its content even where AI assisted in its preparation. It also requires that every use of AI be reconstructable and comprehensible (traceability) and that the functioning of the systems be knowable (*conoscibilità*). Finally, it requires administrations to adopt organisational and training measures, to be implemented within existing resources.

The convergence with the design of this framework is direct. The *instrumental* role corresponds to the decision-support positioning, the *reserved human decision* to the review queue and the officer-only eligibility decisions, *traceability* to the hash-chained audit log, and *knowability* to per-decision explanations and the system card.

1.3.4 Italian Administrative Procedure Law (Law 241/1990) and Algorithmic Administration

Law 241/1990 on administrative procedure and the right of access to administrative documents is the cornerstone of Italian administrative legality. Three of its principles are particularly relevant.

Binding deadlines (Article 2). Every procedure initiated by an application must be concluded with an express decision within a defined term. Where no specific term is set by law or regulation, the default is thirty days. Deadlines set by specific regulations, such as the forty-five days for residence verification under DPR 223/1989 recalled in REG-001, the sixty days for school-enrolment rankings in REG-011, or the forty-eight hours for urgent burials in REG-012, take precedence. The framework encodes these targets as service-level agreements (SLAs), which are used both by the risk-analysis agent and as the compliance metric of the simulation. SLA compliance is defined as the share of requests answered within their category-specific deadline:

$$SLA_{\text{compliance}} = \frac{1}{N} \sum_{i=1}^N \mathbf{1} [T_i \leq D_{c(i)}]$$

where T_i is the response time of request i , $c(i)$ its category, and D_c the regulatory deadline for category c .

Duty to state reasons (Article 3). Every administrative act must state the factual and legal reasons on which it rests. A system that assists in drafting replies must therefore make the legal basis of each statement explicit and verifiable. This is the legal root of the framework's citation mechanism, in which each normative sentence in a draft is linked to the regulation article from which it is taken.

Responsible official and participation (Articles 4–10). Each procedure has an identifiable *responsabile del procedimento*, and citizens have participation and access rights. AI assistance cannot dilute the identification of the responsible official. In the framework, review actions are attributed to named officers in the audit trail.



Jurisprudence on algorithmic administration. Italian administrative courts have elaborated principles for using algorithms in administrative decisions. In decisions no. 2270/2019 and no. 8472/2019, the *Consiglio di Stato* (Council of State, Section VI) held that algorithmic decisions are admissible but subject to three principles. The first is *knowability*: the algorithm and its logic must be knowable and comprehensible to the parties and the court. The second is *non-exclusivity*: there must be human contribution capable of controlling, validating or overriding the automated decision. The third is *algorithmic non-discrimination*: the controller must prevent discriminatory effects. These principles anticipate, within the Italian legal order, the requirements later codified in the AI Act and in Law 132/2025. They justify, respectively, the explanation layer, the human-in-the-loop design and the fairness audit of the framework.

1.3.5 Synthesis: Regulation as Architecture

Table 1.3 consolidates the regulatory analysis into architectural consequences. The methodological premise is that compliance should be an *architectural property* of the system, verified automatically at every run, rather than a document drafted after implementation.

Table 1.3 – Regulatory sources mapped to architectural components

Regulatory principle	Sources	Architectural component
Human decision reserved for eligibility	AI Act Art. 14; GDPR Art. 22; L. 132/2025 Art. 14; CdS 8472/2019	Decision-support positioning; officer-only eligibility decisions; review queue
Traceability and record keeping	AI Act Art. 12; L. 132/2025 Art. 14	Hash-chained audit log; per-decision trace of all agents
Explicability and duty to state reasons	AI Act Art. 13; L. 241/1990 Art. 3; CdS 2270/2019	Citation of regulation articles; structured explanations
Transparency to citizens	AI Act Art. 50	AI disclosure in every reply
Data minimisation and protection by design	GDPR Arts. 5, 25, 32	PII masking; keyed pseudonyms; encrypted vault; role-based access
Data governance and bias examination	AI Act Art. 10; CdS 8472/2019	Data-quality audit; fairness metrics; historical bias tests
Accuracy and robustness	AI Act Art. 15	Evaluation on independent and noisy test sets; hallucination detection
Timeliness of administrative action	L. 241/1990 Art. 2	Category-specific SLAs; risk scoring; simulation of SLA compliance

1.4 Technical Framework

This section describes the technologies used to build the framework and explains their relevance to the requirements identified in Sections 1.1–1.3. The mathematical treatment follows in Chapters 3 and 4.

1.4.1 Large Language Models and Their Limitations in Public Administration

Large language models (LLMs) based on the Transformer architecture generate fluent text conditioned on a prompt and have made conversational interfaces to complex domains technically accessible. Their use in administrative correspondence is nonetheless constrained by three well-documented limitations. The first is **hallucination**: models generate content that is fluent but unsupported by, or contradicts, the source information (Ji et al., 2023). In administrative communication a hallucinated deadline, amount or requirement is not a stylistic defect. It is a misstatement by the public authority, with potential legal consequences under the principle of legitimate expectations. The second is **knowledge staleness and opacity**: the knowledge encoded in model parameters cannot be



traced to a specific regulation article and is not updated when a municipal regulation changes. The third is **non-determinism and limited auditability**: the same request may receive different answers, and the reasoning behind an answer is not directly inspectable. Together, these limitations make an LLM used in isolation incompatible with the duty to state reasons (Law 241/1990, Article 3) and with the traceability requirements of the AI Act and of Law 132/2025.

1.4.2 Retrieval-Augmented Generation

Retrieval-augmented generation (RAG) addresses these limitations by conditioning generation on passages retrieved from an authoritative, updatable corpus (Lewis et al., 2020). In the municipal setting, the corpus consists of regulation articles and documents derived from historical case statistics. The generator is instructed to answer only from the retrieved passages and to cite them. RAG converts the knowledge problem from "what the model remembers" into "what the administration has published", which is precisely the shift required by the duty to state reasons. Updating the system after a regulation changes requires re-indexing the corpus, not retraining a model.

A RAG system can only perform as well as its retrieval stage. Two families of retrievers are combined in this work. **Sparse lexical retrieval**, represented by the probabilistic relevance model BM25 (Robertson and Zaragoza, 2009), scores a document d for a query q as

$$\text{BM25}(q, d) = \sum_{t \in q} \text{IDF}(t) \frac{f(t, d) (k_1 + 1)}{f(t, d) + k_1 \left(1 - b + b \frac{|d|}{\text{avgdl}}\right)}$$

where $f(t, d)$ is the frequency of term t in d , $|d|$ the document length, avgdl the average document length, and k_1 and b are saturation and length-normalisation parameters. BM25 is precise for exact administrative terms such as "TARI", "CILA" or "ZTL", which carry precise legal meaning. **Dense retrieval** represents queries and passages as vectors in a shared embedding space and ranks by cosine similarity (Karpukhin et al., 2020). It captures paraphrase and partial overlap, which matter because citizens rarely use the vocabulary of regulations. The two signals are complementary, and their rankings can be combined without score calibration by reciprocal rank fusion (Cormack, Clarke and Büttcher, 2009):

$$\text{RRF}(d) = \sum_{r \in R} \frac{1}{k + \text{rank}_r(d)}$$

where R is the set of rankers and k is a smoothing constant, conventionally 60. A subsequent stage re-ranks query–passage pairs jointly, and a context-compression stage selects the sentences that are passed to the generator. Chapter 4 details this pipeline, and Chapter 5 reports an ablation study of each component.

1.4.3 Multi-Agent Architectures

A municipal request requires several distinct competences: understanding and retrieval, legal and procedural analysis, risk assessment, drafting, and quality control. Multi-agent systems decompose such a task among autonomous components with specialised roles that interact through shared state or messages (Wooldridge, 2009). In LLM-based systems, agent decomposition has also been used to interleave reasoning and tool use (Yao et al., 2023). For public administration, the decisive advantage of decomposition is *accountability by construction*. Each agent produces an inspectable intermediate output (retrieved evidence, routing decision and rule flags, risk score and factors, draft, confidence and



verdict), so the reasoning chain behind any reply can be reconstructed step by step, as Law 132/2025 requires. Chapter 4 describes the framework's five agents and their coordination through a blackboard.

1.4.4 Discrete-Event Simulation for Strategic Decision Support

Evaluating an AI component in isolation, for instance by its classification accuracy, does not answer the question a municipal manager actually faces: *what happens to response times, deadline compliance and costs if this system is deployed, under the demand conditions the city will realistically face?* Answering it requires a model of the whole service system. Discrete-event simulation (DES) represents a system as entities (requests) flowing through activities that consume resources (officers), with state changing only at discrete event times (Law, 2015; Banks et al., 2010). DES captures exactly the features that make municipal workflows hard to analyse in closed form: priority queues for urgent requests, stochastic service times, misrouting and transfers, waiting for citizen documents, escalation to senior staff, and hand-offs between departments.

Using modelling and simulation to support decisions for strategic problems in complex socio-technical systems is a central theme of strategic engineering. This includes urban vulnerability, disaster prevention and resilience, where simulation allows decision makers to explore scenarios that cannot be experimented on in reality (Bruzzone et al., 2014; Bruzzone et al., 2017a; Bruzzone, 2018). This thesis applies that perspective to municipal request handling. The DES is implemented in SimPy (Team SimPy, 2024), a process-based discrete-event simulation library for Python. It is complemented by analytical queueing models: the Erlang-C (M/M/c) formula, long established for staffing service systems such as call centres (Gans, Koole and Mandelbaum, 2003), is used for fast staffing optimisation, while the DES captures the dynamics that the analytical model omits. Chapter 3 presents both models.

1.4.5 Trustworthy AI as an Engineering Discipline

The European approach to AI is grounded in the notion of *trustworthy AI*, articulated by the High-Level Expert Group on AI (2019) through seven requirements: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being; and accountability. The AI Act translates several of these requirements into binding obligations. The engineering challenge is to make them *measurable*. This thesis operationalises each requirement as one or more metrics that are computed automatically at every run (Table 1.4) and reported in Chapter 5.

Table 1.4 – Trustworthy-AI requirements and their operational metrics

Requirement	Operational metric(s)
Human agency and oversight	Share of decisions routed to review; eligibility decisions taken by officers; acceptance rate of drafts
Technical robustness	Routing accuracy on clean and noisy independent test sets; expected calibration error (ECE)
Privacy and data governance	Direct identifiers in the working set (target: zero); access-control denials logged
Transparency	Transparency index; citation accuracy; AI disclosure rate
Fairness	Demographic parity difference; disparate impact; equal-opportunity difference
Accountability	Audit-chain integrity verification; trace completeness
Faithfulness (generative AI)	Share of claims supported by sources; detector recall under fault injection



1.5 Market and State of the Art

1.5.1 Municipal Conversational Agents

Public administrations at all levels have deployed conversational agents, and the literature documents both their potential and their limits. Androutsopoulou et al. (2019) show that AI-guided chatbots can improve communication between citizens and government by giving round-the-clock access to information and by reducing repetitive front-office load. Wirtz, Weyerer and Geyer (2019) map the application areas of AI in the public sector and identify, alongside efficiency gains, challenges of accountability, data quality, legal compliance and public acceptance. In practice, municipal chatbots have typically taken one of three forms:

1. **Rule-based or intent-based FAQ bots**, which match a question to a predefined intent and return a scripted answer. They are predictable but brittle. They fail on paraphrases, on multi-issue messages and on any question outside the scripted catalogue, and the catalogue must be maintained by hand whenever regulations change.
2. **Generic LLM assistants on municipal websites**, which answer fluently over a broad range of questions but inherit the hallucination and traceability problems discussed in Section 1.4.1. They seldom cite the specific regulation article on which an answer rests.
3. **Search interfaces over municipal web content**, which retrieve pages rather than answers and leave the citizen to locate the relevant provision.

1.5.2 Decision-Support and Workflow Tools

Public administrations also use case-management, protocol and ticketing systems that register and track requests, and business-intelligence dashboards that report volumes and backlogs. These tools are essential for record keeping but are largely *descriptive*: they show what happened, not what would happen under alternative staffing, policies or demand scenarios. Simulation-based decision support is well established in other public-sector domains, such as emergency services, healthcare operations and disaster management. However, it is rarely coupled with the AI components that actually change the service process, so the operational impact of AI adoption is usually estimated rather than modelled.

1.5.3 Gap Analysis

Table 1.5 compares the existing approaches with the capabilities required by analysing Sections 1.1–1.4.

Table 1.5 – Gap analysis of existing municipal AI approaches

Required capability	FAQ / intent bots	Generic LLM assistants	Case management & BI	Proposed framework
Robust understanding of free-text, bilingual requests	Low	High	—	High (lexicon + hybrid retrieval)
Answers grounded in and citing specific regulation articles	Partial (scripted)	Low	—	Yes
Claim-level verification of generated text	—	No	—	Yes (hallucination detector)
Routing to the competent department with	Partial	Low	Manual	Yes



Required capability	FAQ / intent bots	Generic LLM assistants	Case management & BI	Proposed framework
explanation				
Urgency and risk assessment tied to regulatory deadlines	No	No	Partial	Yes
Human-in-the-loop review with recorded decisions	No	Rare	Yes (manual)	Yes (gated)
Tamper-evident audit trail	No	No	Partial	Yes (hash chain)
Fairness monitoring across demographic groups	No	No	No	Yes
Machine-checked AI Act / GDPR compliance	No	No	No	Yes
Simulation of operational impact and staffing under scenarios	No	No	No	Yes (DES + Erlang-C)

Four gaps emerge. **First**, existing systems separate *answering* from *governing*. Fluency and grounding are rarely combined with verification and audit. **Second**, compliance is treated as documentation rather than as a measurable property of each run. **Third**, the operational effect of AI adoption on deadlines, backlogs and staffing is not modelled. Investment decisions are therefore taken without evidence about the scenarios that matter most for a city like Genoa, such as weather alerts or infrastructure failures. **Fourth**, evaluation practice is often weak. Systems are tested on data resembling their development material, and their performance on genuinely unseen, noisy, multilingual input is unknown. This last gap proved concrete in this study: the provided request dataset contains only fifteen distinct message texts, so performance on it alone would overstate real-world capability. Chapter 5 therefore introduces independent test sets.

1.6 Problem Statement and Requirements

1.6.1 Problem Statement

Based on the preceding analysis, the problem addressed by this thesis is formulated as follows:

How can a municipal administration such as the Comune di Genova use artificial intelligence to reduce the cost and delay of handling citizen requests triage, routing, regulatory retrieval and reply drafting while (a) guaranteeing that every output is grounded in and traceable to the applicable regulation, (b) preserving human decision-making authority over outcomes with legal effect, (c) complying by design with the EU AI Act, the GDPR, Law 132/2025 and Law 241/1990, and (d) providing municipal leaders with quantitative evidence of the operational impact of adoption under realistic and adverse demand scenarios?

The study addresses three research questions:

- **RQ1 (Effectiveness).** To what extent can a multi-agent hybrid-RAG engine correctly route, ground and draft responses to bilingual, free-text citizen requests, including requests that differ from its development data?
- **RQ2 (Trustworthiness).** Can the obligations of the AI Act, the GDPR and Italian administrative law be operationalised as automatically verified properties of the system,



and how well do the resulting safeguards, such as hallucination detection, calibration and fairness monitoring, perform empirically?

- **RQ3 (Strategic impact).** What is the simulated effect of alternative AI-adoption policies on response time, deadline compliance, backlog and cost, across normal and adverse scenarios, and what staffing configurations does it imply?

1.6.2 Data Context and Its Implications

The analysis uses three datasets representing the Comune's information environment. These are 320 citizen requests across three channels (web portal, e-mail, PEC) with expected department and urgency labels, 45 articles from 15 municipal regulations, and 500 historical cases with outcomes and response times for 2022–2026. The data-quality audit performed as part of this work (Chapter 4) found characteristics that shape the requirements. All 55 requests labelled as English are in fact written in Italian, so declared metadata cannot be trusted and language must be detected from content. Seventy requests carry dates after the reference date of the study, which requires temporal validation at intake. Articles 2 and 3 of every regulation share a common template, so retrieval must favour substantive over procedural text. Finally, the small number of distinct request texts makes independent test sets indispensable for a valid evaluation.

1.6.3 Functional Requirements

Table 1.6 lists the functional requirements together with their regulatory or analytical source.

Table 1.6 – Functional requirements

ID	Requirement	Traceability
FR1	Ingest requests from web portal, e-mail and PEC, and regulations, documents and historical logs from heterogeneous sources, recording provenance	AI Act Art. 10
FR2	Detect request language from content and reply in that language (Italian, English)	Data audit (Sect. 1.6.2)
FR3	Mask direct personal identifiers before analytical processing; pseudonymise citizen identity	GDPR Arts. 5, 25
FR4	Retrieve the regulation articles and historical evidence relevant to a request using hybrid retrieval	L. 241/1990 Art. 3
FR5	Classify the service category and route to the competent department with an explanation	Sect. 1.1.3 (ii)
FR6	Identify applicable procedural obligations and rule-specific conditions (e.g., safety plan above 200 event attendees)	Municipal regulations
FR7	Assess urgency and risk with reference to regulatory deadlines and historical performance	AI Act Art. 9; L. 241/1990 Art. 2
FR8	Draft replies in which every normative statement cites its source; include AI disclosure	AI Act Arts. 13, 50
FR9	Verify each claim of a draft against sources; compute faithfulness	AI Act Art. 15
FR10	Gate outputs (auto-reply, officer review or urgent dispatch) on confidence, risk and faithfulness; never decide eligibility	AI Act Art. 14; GDPR Art. 22; L. 132/2025 Art. 14
FR11	Provide a review workflow for officers to approve, edit or reject drafts, with attribution	AI Act Art. 14; L. 241/1990
FR12	Simulate request handling under scenarios and operating policies; estimate response time, SLA compliance, backlog, utilisation and cost	RQ3
FR13	Recommend staffing per department under service targets	RQ3
FR14	Forecast demand and SLA-breach risk by department and scenario	AI Act Art. 9
FR15	Present results to municipal leaders through a dashboard with recommendations, alerts and action plans	RQ3



1.6.4 Non-Functional Requirements

Table 1.7 – Non-functional requirements

ID	Requirement	Verification (Chapter 5)
NFR1	Traceability: every decision and review action recorded in a tamper-evident log	Audit-chain integrity check
NFR2	Explicability: every decision accompanied by evidence, scores and reasons	Transparency index
NFR3	Faithfulness: drafts contain no unsupported normative claims	Faithfulness; detector recall under fault injection
NFR4	Robustness: performance maintained on unseen, paraphrased and noisy input	Accuracy on challenge and noisy sets
NFR5	Fairness: no material disparity in automated handling across age groups and channels	Disparate impact ≥ 0.8 ; parity and equal-opportunity differences
NFR6	Calibration: confidence scores informative of correctness	Expected calibration error
NFR7	Privacy and security: no direct identifiers in analytical stores; role-based access; encrypted identity vault	Automated check; access-denial tests
NFR8	Efficiency and scalability: interactive latency; graceful growth with corpus size	Latency and scalability tests
NFR9	Transparency of assumptions: every simulation parameter not measured from data explicitly declared	Assumption register in configuration
NFR10	Deployability: operable on standard municipal hardware, offline-capable, with optional external LLM	One-command launcher; offline mode

1.6.5 Scope and Delimitations

This thesis designs, implements and evaluates a decision-support framework. It does not deploy the system in production at the Comune di Genova, and it does not replace protocol or case-management systems, with which it is designed to integrate through an API. Simulation parameters that cannot be measured from the available data, such as staffing levels, handling effort and costs, are treated as explicit assumptions and examined through sensitivity and what-if analysis. Human evaluation of reply quality is prepared through annotation instruments but not conducted at scale within this work. Within these limits, the remainder of the thesis addresses RQ1–RQ3: Chapter 3 develops the simulation model, Chapter 4 the trustworthy AI framework, Chapter 5 its empirical evaluation, and Chapter 6 the decision-support dashboard and deployment path.



CHAPTER 2 – INTERNSHIP AT THE COMUNE DI GENOVA

This thesis grew out of a seven-month internship at the Comune di Genova focused on artificial intelligence in public administration. Working within a large Italian municipality provided direct insight into its administrative processes, digital initiatives and data. Those observations informed the requirements for the framework developed in Chapters 3–6. This chapter introduces the host organisation and internship setting (Section 2.1), objectives and methods (Section 2.2), and preparatory research on AI in European public administration (Section 2.3). It then examines administrative processes (Section 2.4), the proposed designs for two AI-assisted services (Section 2.5), and the Business Intelligence project involving Zenia, the municipality's virtual assistant (Section 2.6). Section 2.7 connects the internship findings to the subsequent research design, while Section 2.8 identifies the limitations of the internship work.

2.1 Host Organisation and Internship Setting

The internship was hosted by the *Scuola di Amministrazione* of the Comune di Genova, the unit responsible for training and administrative innovation and later with the IT department within the municipal organisation, as part of the requirements of the International Master's Degree in Engineering Technology for Strategy and Security (STRATEGOS, 6th edition) of the Università degli Studi di Genova. It lasted seven months and was supervised by Nicola Iacobone, Director of the *Scuola di Amministrazione*, as organisational advisor and by Professor Marco Gotelli as academic advisor. Regular meetings with both advisors defined the objectives, reviewed progress and oriented the activities towards the research questions later addressed in this thesis.

Like other large Italian municipalities, the Comune di Genova separates political direction from administrative execution. The Mayor, the municipal executive and the municipal council set priorities and policies; the general management, the directorates and their departments, technical offices and administrative units implement them, supported by digital and information-system units, front-office citizen service points and back-office service units. Some services face citizens directly, such as certificates, school services, public communication and online portals; others are internal, such as data management, administrative reporting, document flow and coordination between departments.

Understanding this organisational structure mattered because it determines where AI can responsibly be introduced. Each office has a defined remit, and named officials remain accountable for its decisions. An AI service used by citizens or staff must therefore respect departmental boundaries, draw on validated sources and procedures, and keep final responsibility with a human official. These principles shape the request-routing model in Chapter 3, the role-based access controls in Chapter 4, and the review workflow described in Chapter 6.

2.2 Objectives and Methodology

The internship combined practical work in public-sector AI with participation in the Comune di Genova's digital-transformation work. It also examined how intelligent technologies might



support more transparent, efficient and citizen-oriented administration. The specific objectives were to:

- understand the organisation and functions of the administrative structures of the Comune;
- study AI within European and local public administration, including its regulatory framework;
- identify administrative processes, in particular repetitive communication flows, that could benefit from AI assistance;
- acquire hands-on experience with Microsoft Fabric as an integrated analytics platform and with Power BI for dashboard development;
- understand how Business Intelligence and predictive analytics can support decisions, resource allocation and workload management in the municipality;
- establish a foundation for research on large language models (LLMs) in public administration.

A mixed methodology combined theoretical research with work inside the organisation. A **literature review** covered scientific publications, European Commission and OECD reports and national digital-government strategies. **Organisational observation** examined communication, reporting, information management and administrative workflows. **Weekly meetings** with supervisors and administrative staff identified practical problems and repetitive tasks, and **requirement gathering** through interviews and informal discussions turned them into documented needs. **Technical training** on Microsoft Fabric covered data integration, pipelines, lakehouses, dataflows, semantic models and Power BI. **Business Intelligence analysis** examined existing reports and data flows, and **conceptual AI system design** translated the observed needs into proposals. Activities were documented in daily reports, which supported both supervision and the traceability of the work. Table 2.1 summarises the phases of the internship.

Table 2.1 – Phases, activities and outputs of the internship

Weeks	Phase	Main activities	Outputs
1–2	Foundations	Study of AI in public administration and digital government; exploration of the Comune's organisation; study of the AI Act; survey of AI use in European municipalities	Conceptual framework; catalogue of municipal AI use cases
3–4	Administrative processes	Internal meetings; requirement gathering; workflow analysis; documentation review	List of administrative pain points and candidate processes
5–6	AI chatbot design	Needs analysis; functional requirements; safeguards	Concept of an administrative and citizen-information chatbot
7–8	E-mail response system design	Six-phase LLM pipeline; governance and safeguards	Methodology of an LLM-assisted e-mail response system (Figure 2.1)
9–12	Zenia BI project set-up	Analysis of Zenia conversation data; Microsoft Fabric workspace	Understanding of the data architecture
13–18	Data engineering	Bronze, Silver and Gold layers; semantic models	Cleaned analytical dataset (Gold layer)
19–22	Dashboard development	Language and inclusion, topic effectiveness, citizen experience dashboards	Three Power BI dashboards (Figures 2.4–2.6)
23–24	Findings	Interpretation of dashboards; predictive-analytics opportunities	Findings and recommendations



2.3 Preparatory Study: AI in European Public Administration

The first weeks were devoted to the conceptual and regulatory foundations of the work. Following Wirtz et al. (2019), public-sector AI applications were classified into four groups: *administrative* AI for routine tasks such as sorting e-mails, booking appointments and data entry; *analytical* AI for forecasting, trend analysis, fraud detection and resource planning; *interactive* AI such as chatbots and virtual assistants; and *transformative* AI, which can reshape services and forms of governance. A survey of European municipalities identified recurrent applications: citizen-information chatbots, document classification, appointment scheduling, workflow automation, predictive maintenance of infrastructure, mobility optimisation, resource allocation, urban-planning support and service analytics. A consistent observation was that cities that had already invested in Business Intelligence and data quality were better prepared to adopt AI.

European cases of failure were studied as closely as successes, because they show which safeguards are indispensable. In the Netherlands, the *System Risk Indication* (SyRI) combined data from several public sources to compute welfare-fraud risk scores; in 2020 the District Court of The Hague found that the legislation behind it violated the right to private life under Article 8 of the European Convention on Human Rights, citing the opacity of the system and its concentration on low-income neighbourhoods (Van Bekkum and Zuiderveen Borgesius, 2021). In the United Kingdom, the algorithm used in 2020 to calculate examination grades when examinations were cancelled downgraded students from historically lower-performing schools and was withdrawn after public protest (Kelly, 2021). Experimental research with Dutch civil servants further shows that officials do not simply defer to algorithmic advice but may adhere *selectively* to advice that matches existing stereotypes (Alon-Barkat and Busuioc, 2023). Table 2.2 summarises the lessons drawn from these cases.

Table 2.2 – Lessons from European cases and their design implications

Case	What went wrong or right	Implication for a municipal AI system
SyRI, Netherlands	Opaque risk scoring of citizens; disproportionate targeting; struck down by a court in 2020	Transparency of logic and data; proportionality; no risk scoring of citizens' eligibility
A-level grading, United Kingdom	Statistically consistent algorithm reproduced historical inequalities between schools	Fairness must be measured on outcomes, including on historical data, before and during use
Automation bias and selective adherence	Officials may follow algorithmic advice selectively, reinforcing bias	Human oversight needs explanations, evidence and monitoring, not only a sign-off step
Estonia (Bürokratt)	Conversational access to public services built on mature digital infrastructure and interoperable data	AI depends on data quality and interoperability; automation excluded from decisions requiring judgement

The review of the AI Act raised five practical design questions. How can staff understand an AI recommendation? Under what conditions is generated information safe to use? Who makes the final administrative decision? How are citizens told when they are interacting with AI? What prevents incorrect outputs from reaching them? Chapter 4 addresses these questions through explainability, risk gating, human oversight, disclosure and hallucination controls, respectively.



2.4 Analysis of Administrative Processes

The internship began by examining how administrative work is carried out, before proposing a solution. Requirement gathering with administrative staff focused on communication-based activities, which emerged as the largest source of repetitive effort. The recurrent problems reported were:

- the same requests for information arriving repeatedly from citizens and from other offices;
- manual searching of internal documentation to answer them;
- manual drafting of standard e-mail replies;
- difficulty in finding up-to-date procedures, and different employees giving different answers to the same question;
- delays caused by information being fragmented across documents, mailboxes and departmental resources.

Workflow analysis showed that a typical request passes through five stages: receipt, verification of the information, preparation of the response, approval and communication of the result. Repetitive communication tasks follow similar structures, which makes them suitable for AI assistance, but they still require human verification before a response is sent. Two conclusions guided the rest of the work. First, AI should be integrated into existing procedures and approval chains rather than replace them, so that institutional responsibility is preserved. Second, a successful project starts from the needs of users rather than from the capabilities of the technology. The documentation review also highlighted the importance of knowledge management: in a large administration, the problem is often not that information is missing but that it cannot be found quickly and reliably, which motivates retrieval-based assistants grounded in official documents.

2.5 Conceptual Design of AI-Assisted Services

2.5.1 Administrative and Citizen-Information Chatbot

The first concept was a chatbot for administrative support, designed as an information-retrieval and decision-support tool and not as an autonomous decision system. For citizens it would act as an *intelligent navigation assistant*, explaining in natural language how, where and when to access a service and which documents to bring, without making administrative decisions or legal interpretations. For staff it would retrieve regulations, procedures and departmental responsibilities from authorised knowledge bases. The functional requirements agreed with supervisors and staff were: natural-language interaction; secure access restricted to authorised municipal knowledge; retrieval of official documents and procedures; context-aware answers with explicit references to their sources; referral of complex cases to human staff; multilingual support; and logging of interactions for audit. Every statement produced by the assistant was required to be backed by an official document.

2.5.2 LLM-Assisted E-mail Response System

The second concept addressed e-mail, the channel through which the municipality receives large volumes of messages from citizens, businesses, tourists and other authorities on subjects ranging from taxation and permits to waste management and civil registration. Each message is currently read, sorted, forwarded and answered manually, which causes delays,



inconsistent or duplicated replies, forwarding errors and difficulty in answering in foreign languages. The proposed system processes messages in six phases (Figure 2.1):

1. **E-mail input and collection** from the municipal mailboxes (IMAP or Outlook) or from uploaded message files.
2. **Preprocessing and extraction** of sender, subject, body and attachments, with removal of signatures, HTML and noise.
3. **Language understanding with an LLM**, performing intent detection, classification into a service category, information extraction (names, dates, document identifiers), sentiment analysis and summarisation, for correspondence in Italian, English, French, Spanish and German.
4. **Decision and action generation**: a draft reply in formal, legally appropriate language, a proposed action (reply, forward, escalate), a label and a priority.
5. **Automation and execution**: routine cases are handled automatically, while legal, financial, complaint and emergency messages are sent to a municipal officer for review before any reply.
6. **Logging and learning**: every message, draft and action is stored, and officers use the logs to adjust prompts, routing rules and knowledge-base entries under supervision.

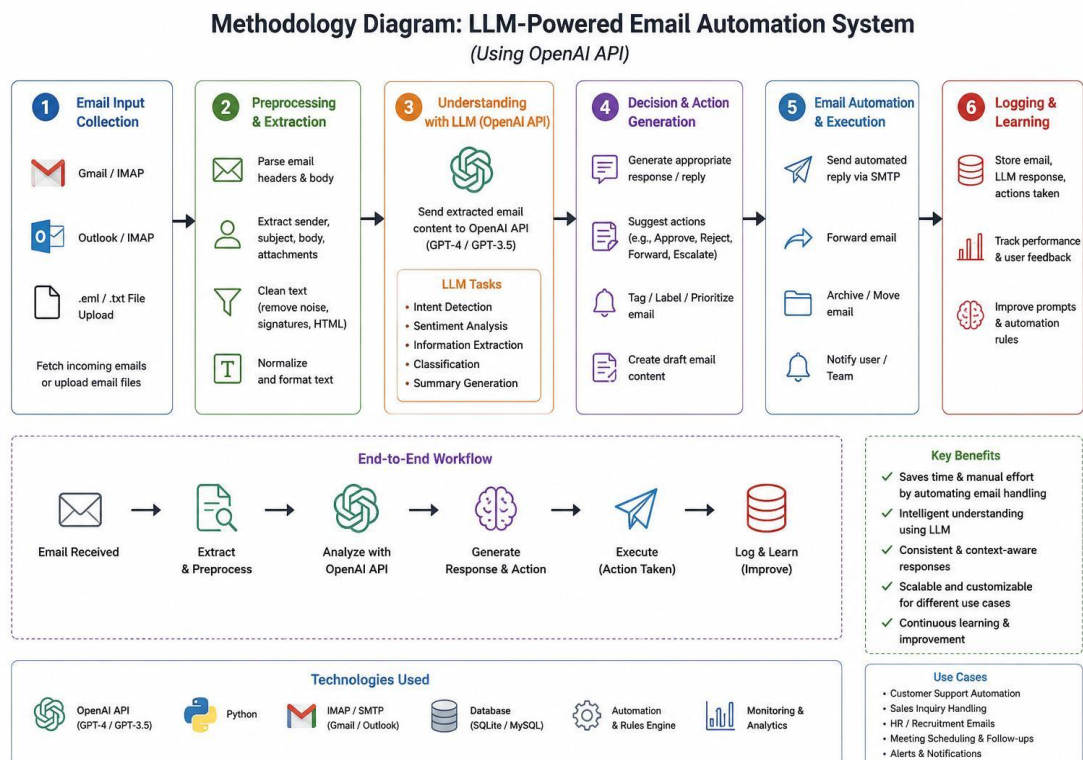


Figure 2.1: Methodology of the LLM-assisted e-mail response system designed during the internship

Because municipal e-mail contains personal data, legal requests and complaints, the design specified a set of safeguards: role-based access control, data minimisation, secure storage, audit logs, retention policies, human verification for sensitive categories, verification of answers against sources, and clearly assigned accountability for every final reply. The



system was not to make legal interpretations or final administrative decisions. The concept relied on a commercial cloud LLM interface, a dependency whose implications for data protection and vendor lock-in were noted and later addressed in the thesis (Section 2.7).

2.6 Business Intelligence Project for Zenia

In the second half of the internship the work moved from concepts to implementation, through participation in the Business Intelligence project for **Zenia**, the AI-powered virtual assistant through which the Comune di Genova answers citizens' questions on municipal services. As the number of conversations grew, the municipality needed a systematic way to monitor the assistant's performance, understand citizens' behaviour and direct improvements. The internship contributed to the full analytical workflow, from raw conversation logs to dashboards for the managers responsible for the service.

2.6.1 Data and Platform

Each Zenia conversation produces session identifiers and metadata, timestamps and duration, the detected language, the topic assigned by the assistant, completion or abandonment flags, feedback signals and escalation events. The raw records contained duplicates, missing values, inconsistent topic names and data from several sources, so data preparation, not dashboard design, was the main effort of the project. The work was conducted in Microsoft Fabric, the platform chosen by the Comune for analytics, which integrates OneLake storage, Spark notebooks for data engineering, lakehouses, SQL analytics, semantic models and Power BI in a single workspace (Figure 2.3).

2.6.2 Medallion Architecture

Data were refined through the three layers of the Medallion Architecture (Figure 2.2). The **Bronze** layer ingests conversation records unchanged, preserving an auditable copy from which later stages can be recomputed without returning to the operational systems. The **Silver** layer, implemented in a PySpark notebook, removes duplicate records, handles null values, standardises topic names, administrative categories and date-time fields, validates language fields and filters records for downstream use. The **Gold** layer produces the business-ready dataset of indicators: conversation volumes over time, average duration by topic and language, the agentic completion rate (the share of conversations resolved by Zenia without a human), the share of negative feedback by topic, language distribution, topic popularity and the short-conversation rate. Semantic models then expose these indicators in business terms (topic, language, conversation, completion rate, feedback, duration) to the Power BI reports.

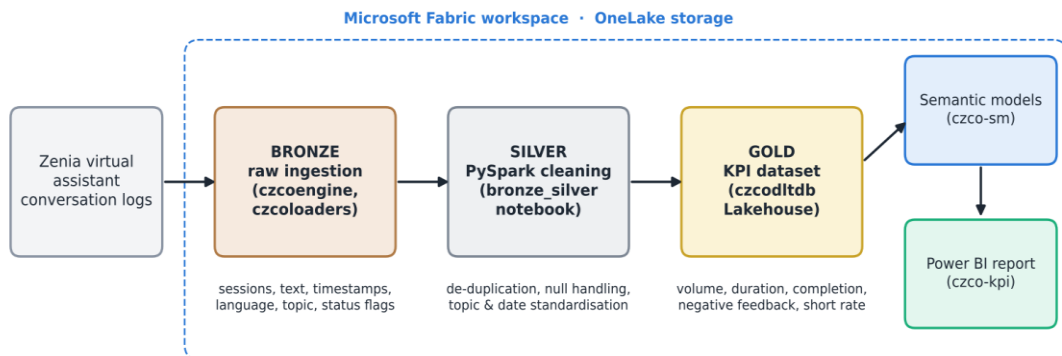


Figure 2.2: Medallion architecture of the Zenia Business Intelligence pipeline in Microsoft Fabric

Name	Status	Git status	Type	Task	Owner	Refreshed	Next refresh	Endorsement	Sensitivity	Included in app
czcoengine			Folder							
czcoloaders			Folder							
bronze_silver			Notebook		Rahman N...					
czco-kpi			Report		fbic-pers-n...	3/12/2026, 4...				No
czco-sm			Semantic ...		fbic-pers-n...	3/12/2026...	N/A			
czcodltldb			Lakehouse		Rahman N...					
etdltldb			Lakehouse		Rahman N...					
nabila_01			Semantic ...		fbic-pers-n...	4/17/2026, 9...	N/A			
snocdltldb			Lakehouse		Rahman N...					

Figure 2.3: Microsoft Fabric workspace of the Zenia project (ingestion folders, notebook, lakehouses, semantic models and report)

2.6.3 Dashboards

Three Power BI dashboards were developed. The **Language and Inclusion** dashboard (Figure 2.4) examines how citizens of different languages use Zenia through completion and short-conversation rates, a topic-by-language heatmap and duration by language. The **Topic Effectiveness** dashboard (Figure 2.5), developed in two iterations, compares service topics through a performance matrix, topic trends, an agentic-completion funnel and a dedicated analysis of conversations that Zenia could not assign to any topic (*noTopic*). The **Citizen Experience and Conversation Quality** dashboard (Figure 2.6) relates the number of messages to the duration of each conversation and analyses short conversations by topic and language. Table 2.3 reports the main indicators. Totals differ between dashboards because each uses its own filters and time window.



Language & Inclusion

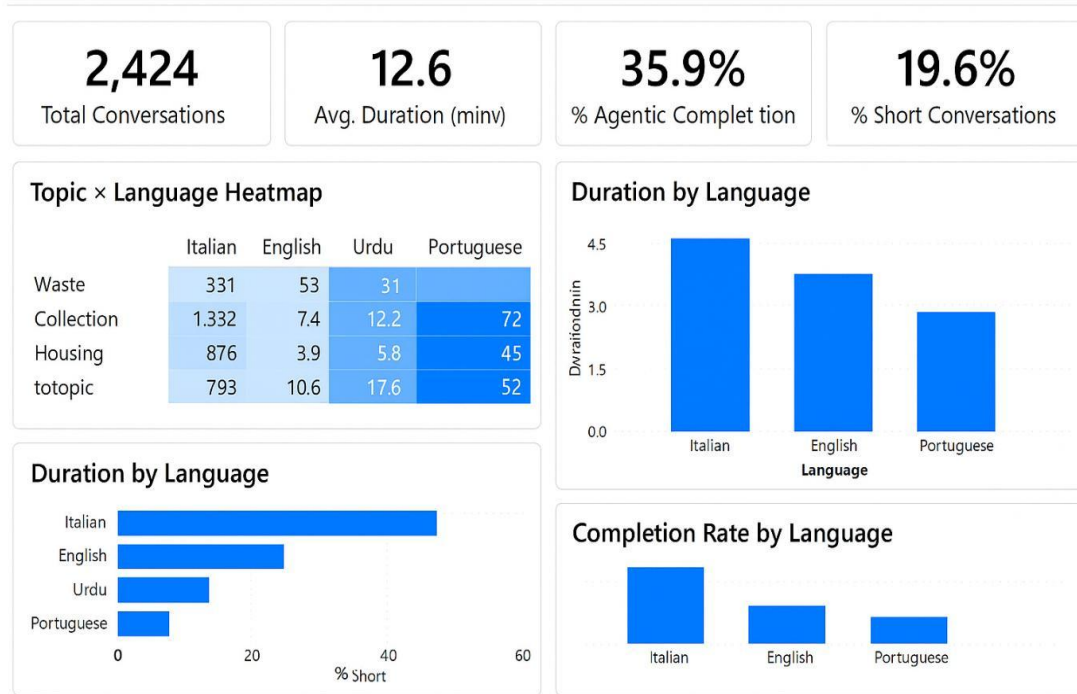


Figure 2.4: Language and Inclusion dashboard

Topic Effectiveness

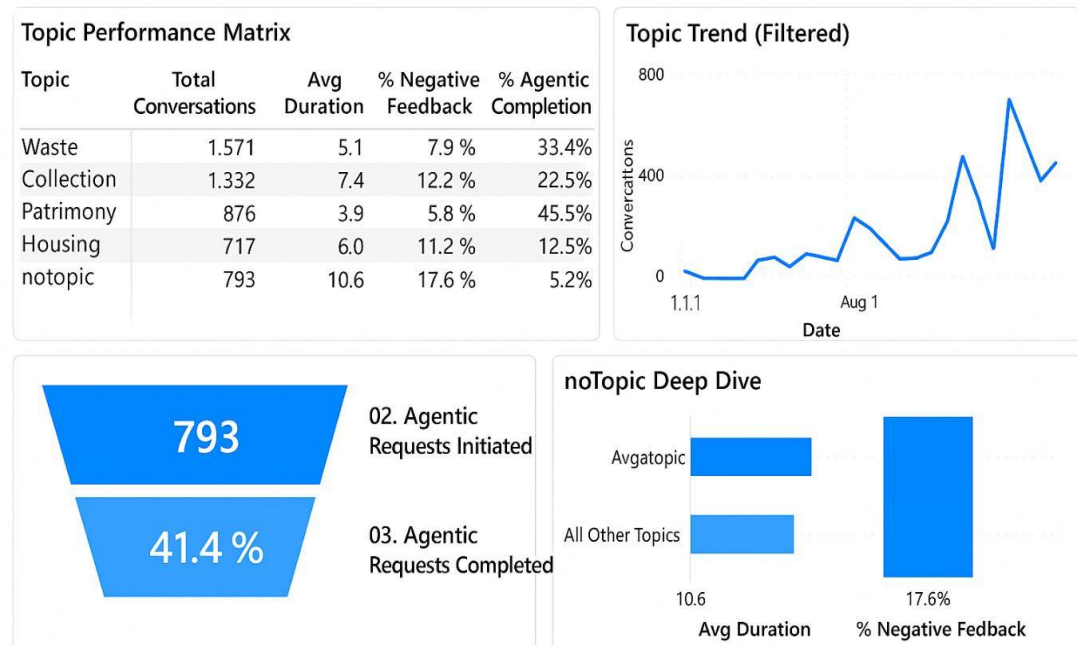


Figure 2.5: Topic Effectiveness dashboard (first iteration)



Citizen Experience & Conversation Quality

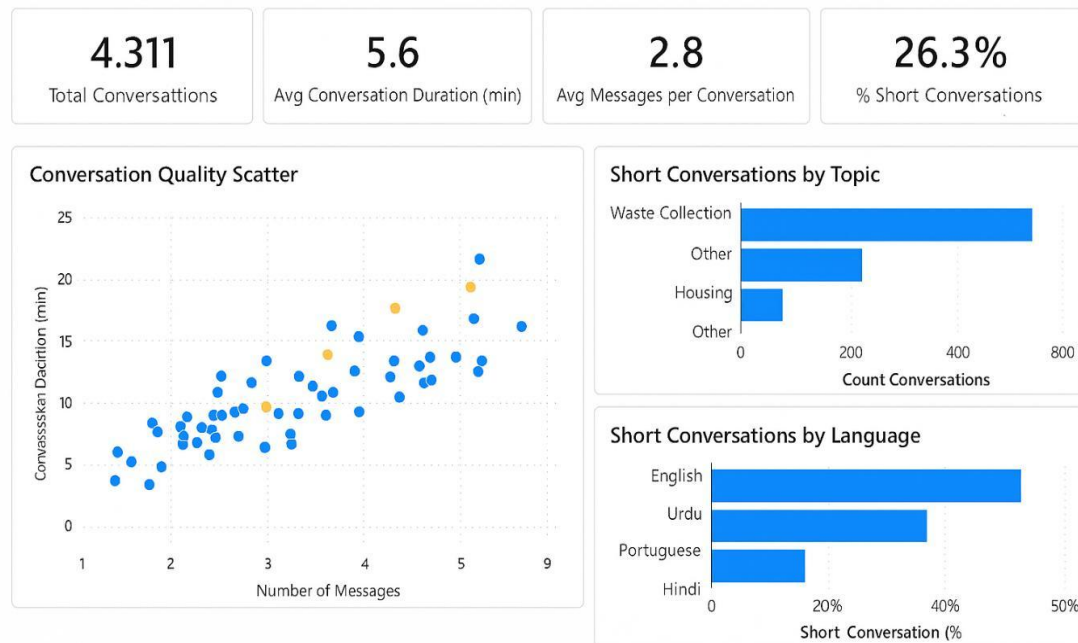


Figure 2.6: Citizen Experience and Conversation Quality dashboard

Table 2.3 – Main indicators of the Zenia dashboards

Dashboard	Indicator	Value
Language and Inclusion	Conversations in the analysed window	2,424
	Agentic completion rate	35.9%
	Short-conversation rate	19.6%
Topic Effectiveness	Waste collection: conversations / agentic completion	1,571 / 33.4%
	Patrimony: conversations / agentic completion	876 / 45.5%
	Housing: conversations / agentic completion	717 / 12.5%
	noTopic: conversations / average duration / negative feedback / completion	793 / 10.6 min / 17.6% / 5.2%
Citizen Experience	Conversations in the analysed window	4,311
	Average duration / messages per conversation	5.6 min / 2.8
	Short-conversation rate	26.3%

2.6.4 Findings and Predictive Opportunities

Five findings emerged from the dashboards. First, demand is concentrated: waste collection and housing account for most conversations in every language group. Second, the assistant's effectiveness varies widely by topic, from 45.5% autonomous completion for patrimony to 12.5% for housing. Third, conversations that Zenia cannot classify (*noTopic*) are the weakest point of the service, with the longest duration, the highest share of negative feedback and almost no autonomous completion; they are the priority for extending the knowledge base. Fourth, although most conversations are in Italian, English, Urdu and Portuguese speakers form a significant and growing share and show higher short-conversation rates, which indicates a need for better multilingual content. Fifth, the



conversation-quality scatter shows a group of conversations with long duration but few messages, suggesting citizens who struggled to find the information they sought, and topic trends show seasonal peaks of demand.

Predictive models were not built during the internship, but the Gold-layer dataset provides the basis for several applications: forecasting citizen demand by topic for staffing and resource planning; predicting periods of high escalation from the assistant to human staff; predicting conversations likely to end with negative feedback; detecting anomalies in completion rates or feedback; improving topic classification from historical interactions; and segmenting citizens by interaction behaviour. Together they would move the municipality from retrospective reporting towards anticipatory management of its services.

2.7 Contribution of the Internship to the Thesis

The internship concluded that AI in a municipality is an organisational transformation as much as a technical one: it succeeds only with reliable data, defined responsibilities, explainable outputs, human oversight and the trust of both staff and citizens. The thesis develops this conclusion into an implemented and evaluated framework. Table 2.4 traces each observation of the internship to the design decision it motivated and to the chapter in which it is addressed.

Table 2.4 – From internship observations to thesis design decisions

Internship observation	Design decision in the thesis	Where
Repetitive requests with similar structure, answered manually by e-mail	Multi-agent engine that routes requests and drafts grounded replies, with automatic replies only under oversight	Ch. 4 (4.7); simulated in Ch. 3 and 5
Information fragmented across documents; inconsistent answers between employees	Hybrid-RAG knowledge base of municipal regulations; every reply cites its sources	Ch. 4 (4.4–4.6)
Legal, financial, complaint and emergency messages must reach an officer	Risk agent and decision gate; safety-critical and eligibility matters never auto-replied	Ch. 4 (4.7.3, 4.8.5)
Unclassified (noTopic) conversations perform worst in Zenia	Confidence agent; uncertain requests are sent to officer review instead of being answered	Ch. 4 (4.7.5); Ch. 5 (5.3)
Multilingual demand; shorter conversations for non-Italian speakers	Bilingual processing (Italian and English) and fairness monitoring by language	Ch. 4 (4.3.2, 4.8.3); Ch. 5 (5.6)
Need for logging, role-based access and accountability for final replies	Hash-chained audit log, role-based access control, review queue recording the responsible officer	Ch. 4 (4.5.3, 4.8.6); Ch. 6 (6.3)
Raw data with duplicates, gaps and inconsistent labels (Bronze to Gold refinement)	Preprocessing, pseudonymisation and data-quality audit before any knowledge is built	Ch. 4 (4.3)
Opportunities for demand forecasting and workload planning	Demand forecast, discrete-event simulation of policies and Erlang-C staffing optimisation	Ch. 3; Ch. 5 (5.7)
Managers need dashboards to steer the service	Strategic decision-support dashboard with scenarios, staffing, risk and compliance views	Ch. 6
Dependency of the e-mail concept on a cloud LLM interface	Offline grounded generator by default; external LLM optional, with masked data and EU processing	Ch. 6 (6.5.3)

The internship also helped define the thesis scope. It showed that the main benefit for the Comune lies in requests handled by many offices through written channels, rather than in a single conversational interface; that a framework must be evaluated on its risks as well as on



its efficiency; and that leaders need quantitative evidence on how the adoption of AI would change response times, backlogs and staffing before committing to it. These three points correspond to the multi-agent decision engine, the trustworthiness evaluation and the simulation-based decision support that form the core of the following chapters.

2.8 Limitations of the Internship Work

Four limitations must be acknowledged. First, the chatbot and the e-mail response system remained at the conceptual stage; they were not implemented or tested with users during the internship. Second, confidentiality, privacy and organisational rules restricted access to internal administrative data and to citizen correspondence, which is why the thesis relies on the datasets made available for the project and on independently constructed test sets (Chapter 5). Third, the Zenia indicators depend on the filters, topic classification and time window of each dashboard and should be read as descriptive of the analysed periods. Fourth, any operational use of the proposed services would require user testing, legal review, a cybersecurity assessment and formal approval by the administration, steps that the roll-out plan of Chapter 6 incorporates.



CHAPTER 3 – SIMULATION MODEL

The simulation model allows the Comune di Genova to examine the operational effects of AI-assisted request handling before deployment. Section 3.1 connects the model to strategic engineering and modelling and simulation (M&S). Sections 3.2 and 3.3 explain its conceptual structure and discrete-event implementation in SimPy. Section 3.4 covers calibration, distinguishing parameters derived from data from those based on explicit assumptions. Sections 3.5 and 3.6 describe the scenarios and operating policies. The Erlang-C staffing optimisation model follows in Section 3.7, while Section 3.8 covers what-if analysis, demand forecasting and risk forecasting. Section 3.9 addresses verification and validation. The simulation results are discussed in Chapter 5.

3.1 Strategic Engineering and Modelling & Simulation as Decision Support

Strategic engineering deals with decisions whose consequences extend across an entire system and over long periods, often under uncertainty that cannot be eliminated. Testing those decisions directly may be impractical, unethical or prohibitively costly (Bruzzone, 2018). Modelling and simulation allow decision makers to compare alternatives without experimenting on the city itself. This approach has been used to study urban vulnerability and resilience, disaster prevention, and strategic decisions involving human factors (Bruzzone et al., 2014; Bruzzone et al., 2017a; Bruzzone et al., 2017b). Municipal request handling presents similar challenges: demand fluctuates unpredictably, departments share an interconnected workload, floods and infrastructure failures can abruptly change service needs, and testing AI-assisted triage on live correspondence would expose citizens to avoidable service risks.

Historical records and the AI performance measures in Chapter 5 cannot, by themselves, establish the operational consequences of a new policy. Routing accuracy shows how often a request reaches the correct department, but not how a 16-point change in misrouting affects deadlines when that department is already operating at 95% utilisation. Historical response times describe the existing workflow, not a plausible alternative. The simulation connects these levels of evidence by applying measured AI properties, including routing accuracy and the share of automatic replies, to a stochastic model of municipal request handling. Its outputs are the indicators managers need to assess policy choices: response times, deadline compliance, backlog, resource utilisation and cost.

Two complementary modelling paradigms are combined. **Discrete-event simulation (DES)** represents individual requests flowing through activities that consume resources, with state changing only at event times (Law, 2015; Banks et al., 2010; Robinson, 2014). DES captures priority disciplines, stochastic service times, routing errors, rework loops, escalations and hand-offs, which make the system analytically intractable. **Analytical queueing models**, specifically the M/M/c (Erlang-C) model (Erlang, 1917; Kleinrock, 1975; Gans, Koole and Mandelbaum, 2003), provide closed-form performance measures that can be evaluated thousands of times per second. This makes them suitable for staffing optimisation, while the DES verifies the resulting configurations under the full workflow logic. The combination follows the established practice of using analytical approximations to search the decision space and simulation to evaluate candidate solutions.



3.2 Conceptual Model of Municipal Request Handling

The conceptual model translates the existing process in Figure 1.1 into entities, resources, activities and decision points. Its scope reflects the main ways AI assistance could alter performance: triage effort, routing accuracy, handling effort, automatic replies and priority treatment of urgent cases. Other processes are included where they occupy shared resources or introduce delays, notably information requests, escalations and transfers between departments. Table 3.1 sets out the model components.

Table 3.1 – Components of the conceptual model

Component	Representation	Attributes / parameters
Entity	Citizen request	Category c (15 values), competent department $d(c)$, urgency flag, arrival time, misrouted / auto-sent / info-requested / escalated flags, completion time
Resource	Front office	6 operators performing reading, triage and routing
Resource	Department officers (10 pools)	c_d officers per department; non-preemptive priority queue (urgent first when the policy enables prioritisation)
Resource	Senior review desk	3 senior officers handling escalations
Activity	Triage and routing	Manual: lognormal effort, mean 0.25 h. AI-assisted: operator check, 0.05 h
Decision	Routing outcome	Misrouted with probability 0.12 (manual) or $1 - \text{accuracy (AI)}$
Activity	Transfer after misrouting	0.25 h reading by the wrong department, then 8 h transfer delay
Activity	Handling	Lognormal effort with category mean h_c and coefficient of variation 0.6, modified by policy
Decision	Information requested	Bernoulli(info_c) from historical outcomes; citizen wait plus rework
Decision	Escalation	Bernoulli(esc_c) from historical outcomes; senior review plus, for some categories, a hand-off to a second department
Output	Key performance indicators	Response time, SLA compliance, completion rate, backlog, utilisation, cost per completed case

The flow of an entity through the model is shown in Figure 3.1. A request arrives, is triaged, and is routed either correctly or to a wrong department, in which case it incurs reading effort and a transfer delay before joining the correct queue. In the competent department, the request is either answered automatically under human oversight (only in the policy that enables auto-replies) or handled by an officer. After handling, it may require additional information from the citizen, with a waiting period followed by rework. It may also be escalated to the senior desk, possibly followed by a hand-off to a second department. The request completes when all activities on its path have finished.

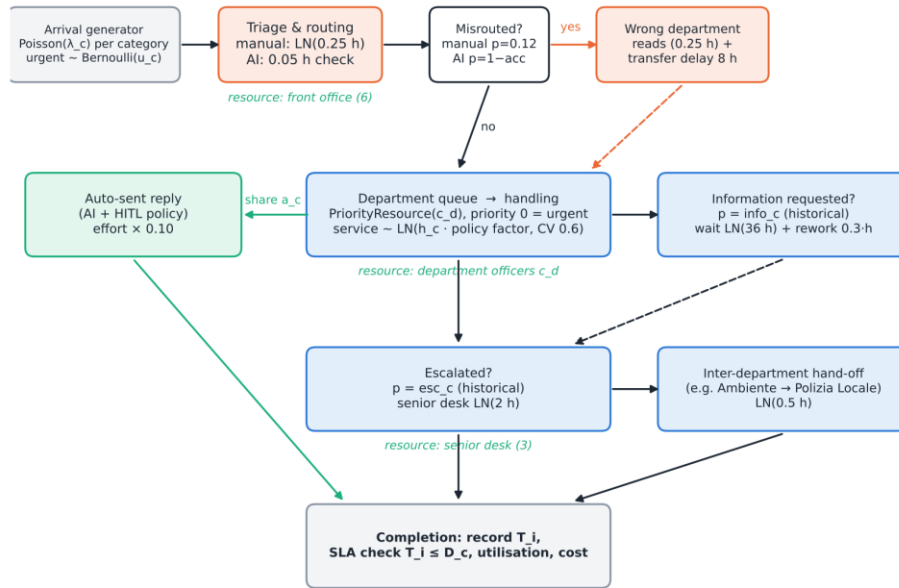


Figure 3.1: Discrete-event flow of a citizen request through the simulation model (LN = lognormal; resources in green italics)

Inter-departmental hand-offs reproduce the joint procedures stated in the municipal regulations. Environmental reports require a joint check by ARPAL and the Local Police (REG-008), so escalated environmental cases are handed to *Polizia Locale*. Residence changes require verification of habitual residence by the Local Police (REG-001). Events above 200 attendees require a safety plan (REG-009), represented as a hand-off to *Protezione Civile*. Escalated noise complaints are handed to *Ambiente* for acoustic measurement, and escalated flood-risk inquiries to *Urbanistica*.

Several simplifying assumptions define the scope of the model. First, the model is a **terminating simulation** of a 20-working-day operational window (Law, 2015) that starts with empty queues; the backlog at the end of the window is reported as an indicator of accumulated pressure. Second, time is measured in **working hours**, with 7.2 working hours per day corresponding to a 36-hour week, and response times are converted to working days for comparison with the regulatory deadlines. Third, service is **non-preemptive**: an urgent request moves ahead of non-urgent requests in the queue but does not interrupt a request already in service. Fourth, waiting for citizen documents and transfer delays are pure delays that do not occupy staff. Fifth, auto-sent replies close informational requests and therefore do not generate information requests or escalations.

3.3 Discrete-Event Simulator Implementation

3.3.1 Process-Oriented Modelling in SimPy

The simulator is implemented in Python with SimPy, a process-based discrete-event simulation library in which each entity is a generator function that yields events (timeouts and resource requests) to a shared environment. The process-interaction world view maps directly onto the conceptual model: each request is a process that requests and releases resources in the order dictated by its path. Departments and the senior desk are modelled as



SimPy *PriorityResource* objects with capacity equal to their staffing, and the front office as a standard *Resource*. Algorithm 3.1 gives the request process in pseudo-code.

Algorithm 3.1 – Request process (simplified from simulation.py)

```
process REQUEST(id, category c, urgent u):
    d ← department(c); prio ← 0 if (u and policy.priority) else 1
    if policy.ai_routing:
        use(front_office, 0.05 h)
        misrouted ← U(0,1) > acc_AI - Δ_s
    else:
        use(front_office, LN(0.25 h, 0.4))
        misrouted ← U(0,1) < 0.12
    if misrouted:
        use(random other department, 0.25 h, prio); wait(8 h)
    h ← h_c · κ_s,c # scenario multiplier
    if policy.auto_send and U(0,1) < a_c: h ← 0.10 · h; auto ← true
    elif policy.ai_drafting: h ← (1 - δ) · h
    use(department d, LN(h, 0.6), prio)
    if not auto and U(0,1) < info_c:
        wait(LN(36 h, 0.5)); use(department d, LN(0.3 · h), prio)
    if not auto and U(0,1) < esc_c:
        use(senior_desk, LN(2 h), prio)
        if c has a partner department d': use(d', LN(0.5 h), prio)
    record completion time T = now - arrival
```

3.3.2 Arrival Process

Requests arrive according to a superposition of independent Poisson processes, one per category. Under scenario s , the arrival rate of category c per working hour is

$$\lambda_c^{(s)} = \frac{V M_s \pi_c \kappa_{s,c}}{H} \quad (3.1)$$

where V is the baseline daily volume, M_s the scenario volume multiplier, π_c the category share estimated from the request dataset, $\kappa_{s,c}$ the scenario multiplier for category c (1 when not specified), and $H = 7.2$ working hours per day. The aggregate process is Poisson with rate $\Lambda = \sum_c \lambda_c$. Inter-arrival times are therefore exponential, and the category of each arrival is drawn from the categorical distribution with probabilities λ_c/Λ :

$$A_k \sim \text{Exp}(\Lambda), \quad P(c_k = c) = \frac{\lambda_c}{\Lambda} \quad (3.2)$$

Each arrival is flagged urgent with probability u_c , the share of urgent requests of category c in the data, unless the scenario overrides it (for instance, all emergency reports are urgent during a weather alert).

3.3.3 Service-Time Distributions

Handling efforts are positive, right-skewed and occasionally long, which makes the exponential distribution assumed by M/M/c models unrealistic for simulation. The lognormal distribution is used instead, parameterised by its mean m and coefficient of variation CV . The underlying normal parameters are obtained as

$$\sigma^2 = \ln(1 + CV^2), \quad \mu = \ln m - \frac{\sigma^2}{2} \quad (3.3)$$



so that $E[S] = m$ and $\text{Var}[S] = (CV \cdot m)^2$. A coefficient of variation of 0.6 is used for handling, 0.4 for manual triage and 0.5 for citizen waiting times. The mean handling effort of category c under operating policy π is

$$\bar{h}_c^{(\pi)} = h_c \kappa_{s,c}^h \times \begin{cases} 0.10 & \text{if auto-sent (AI + HITL, probability } a_c) \\ 1 - \delta & \text{if AI drafting} \\ 1 & \text{otherwise} \end{cases} \quad (3.4)$$

where h_c is the baseline effort, $\kappa_{s,c}^h$ a scenario multiplier (used by the policy-change scenario), $\delta = 0.35$ the share of handling time saved by an AI draft, and a_c the share of requests of category c that the engine auto-replies, as measured in Chapter 5.

3.3.4 Routing, Prioritisation and Rework Logic

The probability that a request is misrouted depends on the operating policy:

$$p^{\text{mis}} = \begin{cases} 0.12 & \text{manual triage} \\ 1 - \max(0, \text{acc}_{AI} - \Delta_s) & \text{AI-assisted triage} \end{cases} \quad (3.5)$$

where acc_{AI} is the department-routing accuracy of the engine and Δ_s a scenario-specific degradation (0.08 in the policy-change scenario, representing new rules not yet indexed). Prioritisation is implemented through the SimPy priority queue: requests with priority 0 (urgent, when the policy enables prioritisation) are served before requests with priority 1, and requests of equal priority are served first-come, first-served.

3.3.5 Performance Indicators

For each replication the simulator records, for every completed request i , the response time T_i in working hours and computes the following indicators over the horizon of length $T = 20 \times 7.2 = 144$ working hours. SLA compliance is computed over all arrivals, so that requests still open at the end of the horizon count as non-compliant:

$$\text{SLA} = \frac{1}{N_{\text{arr}}} \sum_{i \in \mathcal{C}} \mathbb{1} \left[\frac{T_i}{H} \leq D_{c(i)} \right] \quad (3.6)$$

where $\mathcal{C}$ is the set of completed requests and D_c the regulatory deadline of category c in days. Utilisation of department d is the busy time divided by available capacity, and cost per completed case divides the total paid staff time by completions:

$$U_d = \frac{B_d}{c_d T}, \quad \text{Cost} = \frac{(\sum_d c_d + c_{\text{senior}} + c_{\text{front}}) T w}{|\mathcal{C}|} \quad (3.7)$$

where B_d is the cumulative busy time of department d and $w = €32$ the fully loaded hourly staff cost. Completion rate is $|\mathcal{C}|/N_{\text{arr}}$, and backlog is $N_{\text{arr}} - |\mathcal{C}|$. A monitor process samples queue lengths of every department each working hour for time-series analysis.

3.3.6 Replications and Statistical Treatment

Each configuration is run for R independent replications ($R = 6$ for the policy and scenario studies, $R = 3$ for the what-if grids). Replication r uses the random seed $42 + r$, so that different policies are compared under the same seed sequence, a simple form of common random numbers that reduces the variance of differences between policies. For each scalar indicator X the mean and 95% confidence half-width are reported:



$$\bar{X} \pm 1.96 \frac{S_X}{\sqrt{R}} \quad (3.8)$$

For the baseline, the half-width of SLA compliance is below one percentage point for every policy, which is sufficient to separate the policies compared in Chapter 5.

3.4 Calibration

A simulation model is only as credible as its inputs. The calibration strategy therefore separates parameters **measured** from the project data from parameters **assumed** in the absence of data, and registers every assumption in a single configuration file so that it can be reported, challenged and varied (requirement NFR9). Figure 3.2 shows how the datasets, the AI engine and its evaluation feed the simulation.

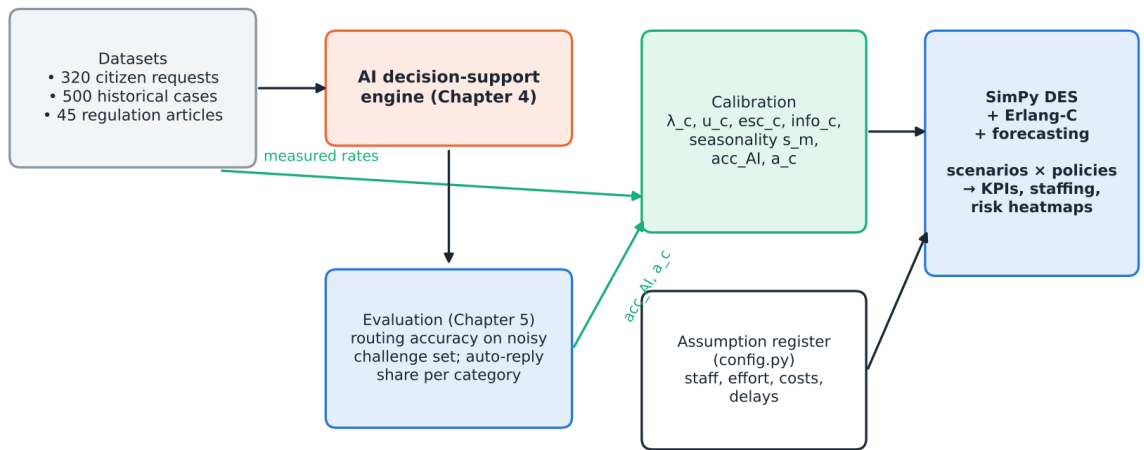


Figure 3.2: Coupling between data, AI engine, evaluation and simulation model

3.4.1 Parameters Estimated from Data

The category mix π_c is estimated from the 319 validated citizen requests. Escalation and information-request probabilities are the shares of historical cases of each category with outcome *escalated* and *info_requested*. Urgency probabilities are the share of requests flagged urgent per category. The SLA deadlines D_c are taken from the regulation texts where stated, and otherwise from the 30-day default of Law 241/1990. Table 3.2 reports the estimates.

Table 3.2 – Parameters estimated from the datasets

Category	Share π_c	Urgent u_c	Escalation esc_c	Info req. $info_c$	Deadline D_c (days)
School enrolment	0.110	0.029	0.152	0.061	60
Cemetery service	0.091	0.034	0.094	0.062	2
Waste tax (TARI)	0.085	0.111	0.143	0.107	30
Building permit	0.075	0.083	0.100	0.075	30
Emergency report	0.069	1.000	0.065	0.065	1
Environmental report	0.066	0.095	0.143	0.071	10
Social benefit	0.063	0.000	0.036	0.107	30
Parking / ZTL permit	0.063	0.050	0.200	0.160	30
Public-transport pass	0.063	0.050	0.081	0.162	30
Noise complaint	0.060	0.105	0.091	0.061	30
Event authorisation	0.056	0.056	0.103	0.103	30



Category	Share π_c	Urgent u_c	Escalation esc_c	Info req. $info_c$	Deadline D_c (days)
Residency change	0.053	0.000	0.088	0.088	45
Public-works complaint	0.053	0.235	0.233	0.100	5
Flood-risk inquiry	0.047	0.000	0.030	0.121	30
Business licence	0.047	0.000	0.086	0.143	30

Two parameters are taken from the evaluation of the AI engine rather than from the raw data. The AI routing accuracy acc_{AI} is set to 0.842, the department-routing accuracy measured on the *noisy* independent challenge set (Chapter 5). This is a deliberately conservative choice: on the provided requests the engine routes every request correctly, but that figure reflects the small number of distinct message texts and would overstate real performance. The auto-reply share a_c is the fraction of requests of each category that the engine's confidence and risk gate released without officer review. It is 1 for most informational services and 0 for cemetery services, emergency reports and public-works complaints, which are always reviewed or dispatched urgently.

3.4.2 Assumption Register

Operational parameters that are not present in the datasets are declared as assumptions (Table 3.3). Their values were chosen to be plausible for a municipal office of Genoa's size and are intended as a reference configuration, not as measurements of the Comune's actual staffing. Section 3.8 and Chapter 5 examine the sensitivity of the conclusions to the most influential of them.

Table 3.3 – Assumption register of the simulation model

Parameter	Symbol	Value	Rationale
Baseline daily volume	V	140 requests/day	Order of magnitude for written requests to a city of ~560,000 residents
Working hours per day	H	7.2 h	36-hour working week
Simulated horizon	T	20 working days	One operational month
Fully loaded staff cost	w	€32 per hour	Salary plus overheads
Manual triage effort	—	0.25 h (CV 0.4)	Reading, interpretation and protocol routing
AI triage check	—	0.05 h	Operator confirms the proposed routing
Manual misrouting rate	—	0.12	Ambiguous competences between departments
Transfer delay	—	8 h	One working day for a misrouted file
AI drafting time saving	δ	0.35	Officer edits a cited draft instead of writing from scratch
Auto-reply oversight effort	—	$0.10 \times h_c$	Sampled supervision of auto-sent replies
Citizen waiting time	—	36 h (CV 0.5)	Time to supply missing documents
Senior review effort	—	2 h	Escalation analysis
Handling effort per category	h_c	0.5–3.0 h	From 0.5 h (transport pass) to 3.0 h (building permit)
Staff per department	c_d	2–6 officers (34 in total)	Officers dedicated to written requests
Front office / senior desk	—	6 / 3	Intake and escalation capacity

The register makes explicit a central methodological point. The *direction* of the policy comparisons is robust across the plausible range of these assumptions, as the what-if analysis shows, but the *magnitude* of the gains depends materially on two of them: the



oversight effort assigned to auto-sent replies and the drafting time saving δ . Both should be re-estimated from a pilot before any operational conclusion is drawn.

3.5 Scenario Generation

Scenarios represent the demand and capacity conditions that the Comune may face. They were derived from Genoa's risk profile (Section 1.2) and from recurrent administrative events. Each scenario is a set of multipliers applied to the baseline calibration: a global volume multiplier M_s , category multipliers $\kappa_{s,c}$, staff multipliers, handling-effort multipliers, urgency overrides and an AI-accuracy degradation. Table 3.4 defines the eight scenarios.

Table 3.4 – Simulation scenarios

Scenario	Trigger	Parameterisation
Baseline	Normal operations	Category mix from data; all multipliers equal to 1
High e-mail volume	Mail/PEC surge after a TARI billing run	$M_s = 2.2$; waste-tax requests $\times 4$
Natural disaster	Orange/red weather alert on Bisagno and Polcevera	Emergency reports $\times 15$; flood-risk inquiries $\times 8$; public works $\times 3$; environmental reports $\times 2$; urgency share 1.0 / 0.6 / 0.5 for emergency / flood / public works
Tourism season	Summer peak	Events $\times 3$; noise $\times 3$; parking $\times 2$; transport passes $\times 2$; business licences $\times 1.5$
Infrastructure incident	Road collapse or lighting failure	Public works $\times 6$ with 60% urgent; parking and transport $\times 1.5$
Staff shortage	Illness or vacancies	All department staff $\times 0.7$
Policy change	New TARI and social-benefit rules	Volume $\times 1.3$ and handling effort $\times 1.4$ in both categories; AI accuracy -0.08 until re-indexing
Budget constraints	Hiring freeze, no overtime	All department staff $\times 0.85$; senior desk -1

The natural-disaster scenario is the most demanding. A fifteen-fold increase in emergency reports concentrates load on *Protezione Civile*, whose regulatory deadline (immediate activation, modelled as one day) is the tightest in the model. The infrastructure-incident scenario has a similar structure for *Urbanistica*. These two scenarios test whether AI support helps the departments that matter most when the city is under stress, rather than only those with routine informational demand.

3.6 Operating Policies

Operating policies represent alternative ways of introducing the AI engine into the workflow. They are cumulative, so each policy isolates the marginal contribution of one capability (Table 3.5).

Table 3.5 – Operating policies compared in the simulation

Policy	AI routing	AI drafting	Auto-reply with oversight	Priority for urgent	Interpretation
Manual	—	—	—	—	Current process
AI routing	✓	—	—	—	Engine proposes the department; operator confirms
AI routing + drafting	✓	✓	—	—	Officers receive a cited draft for every request
AI + human oversight (HITL)	✓	✓	✓	✓	Low-risk, high-confidence informational replies sent automatically and sampled; all others reviewed; urgent requests prioritised



The policy definitions respect the governance constraints of Chapter 1. In no policy does the AI take an eligibility decision. Auto-replies are limited to the requests that the engine's confidence and risk gate releases (Chapter 4), which excludes, by construction, cemetery services, emergency reports and public-works complaints with safety implications. Prioritisation of urgent requests is introduced together with AI support because the engine's urgency detection is what makes reliable prioritisation feasible at intake.

3.7 Resource Model: Erlang-C Staffing Optimisation

3.7.1 The M/M/c Model

For staffing decisions, each department is approximated as an M/M/c queue with Poisson arrivals at rate λ_d , exponential service at rate $\mu_d = 1/E[S_d]$ and c identical officers. With offered load $a = \lambda/\mu$ (in Erlangs) and utilisation $\rho = a/c < 1$, the probability that an arriving request has to wait is given by the Erlang-C formula (Erlang, 1917; Kleinrock, 1975):

$$C(c, a) = \frac{\frac{a^c}{c!} \frac{c}{c-a}}{\sum_{k=0}^{c-1} \frac{a^k}{k!} + \frac{a^c}{c!} \frac{c}{c-a}} \quad (3.9)$$

and the expected waiting time in queue is

$$W_q(c) = \frac{C(c, a)}{c\mu - \lambda} \quad (3.10)$$

which diverges as c approaches a from above. When $c \leq a$ the queue is unstable and W_q is infinite.

3.7.2 Department Load under a Policy

The arrival rate and mean service time of department d aggregate the categories it handles and account for the policy and for information-request rework:

$$\lambda_d = \sum_{c \in d} \lambda_c, \quad \tilde{h}_c = [a_c 0.10 + (1 - a_c)(1 - \delta)]h_c(1 + (1 - a_c) 0.3 \text{info}_c) \quad (3.11)$$

$$E[S_d] = \frac{1}{\lambda_d} \sum_{c \in d} \lambda_c \tilde{h}_c \quad (3.12)$$

Under the manual policy $a_c = 0$ and $\delta = 0$, and the bracket in Equation (3.11) reduces to 1.

3.7.3 Optimisation Problem

The minimum staffing of department d that satisfies a utilisation cap $\rho_{\max}$ and a waiting-time target w^* is

$$c_d^* = \min \left\{ c \in \mathbb{N} : \frac{\lambda_d}{c\mu_d} \leq \rho_{\max}, W_q(c) \leq w^* \right\} \quad (3.13)$$

which is found by incrementing c from $\lceil a_d/\rho_{\max} \rceil$. Default values are $\rho_{\max} = 0.85$ and $w^* = 4$ working hours. When the total headcount N is fixed, as under the Comune's personnel



constraints, staff are redistributed greedily using the marginal reduction in the expected number of waiting requests, $L_q = \lambda W_q$:

$$\Delta_d(c) = \lambda_d [W_q(c) - W_q(c + 1)] \quad (3.14)$$

While the allocation exceeds N , an officer is removed from the department with the smallest loss $\Delta_d(c - 1)$; while it falls short, one is added to the department with the largest gain $\Delta_d(c)$. Because W_q is convex in c for the M/M/c queue (Dyer and Proll, 1977), this marginal-allocation procedure yields an allocation that minimises the total expected queue length for the given headcount. The analytical model omits misrouting, escalation hand-offs and the lognormal service variability, so its recommendations are verified in the DES before they are reported (Section 3.9).

3.8 What-If Analysis, Demand Forecasting and Risk Forecasting

3.8.1 What-If Analysis

What-if analysis explores the response of the system to combined changes in demand and capacity. A full-factorial grid crosses five volume multipliers (0.75, 1.0, 1.25, 1.5, 2.0) with five staffing multipliers (0.7, 0.85, 1.0, 1.15, 1.3). For each cell the DES is replicated and SLA compliance, average response time, maximum utilisation, backlog and cost per case are recorded. The grid is evaluated for both the manual and the AI + HITL policies, so that the two response surfaces can be compared directly (Chapter 5). Policy impact analysis complements the grid by evaluating all four policies of Table 3.5 under the same scenario and reporting the differences regarding the manual baseline.

3.8.2 Demand Forecasting

Strategic planning requires an estimate of future demand. For each category c , the yearly count of historical cases $y_{c,t}$ is modelled with a Poisson generalised linear model with a log-linear trend (McCullagh and Nelder, 1989):

$$y_{c,t} \sim \text{Poisson}(\mu_{c,t}), \quad \log \mu_{c,t} = \beta_{0,c} + \beta_{1,c} (t - \bar{t}) \quad (3.15)$$

estimated by iteratively reweighted least squares, $\beta \leftarrow (X^T W X)^{-1} X^T W z$, with $W = \text{diag}(\mu)$ and working response $z = X\beta + (y - \mu)/\mu$. The annual growth rate of category c is $\exp(\beta_{1,c}) - 1$. Monthly seasonality is estimated from the request dates as the index $s_m = n_m / \bar{n}$, where n_m is the number of requests in calendar month m and $\bar{n}$ the monthly mean. The expected number of requests of category c in a future month m is then

$$\hat{\lambda}_{c,m} = 21 V \pi_c \frac{\hat{\mu}_c(t_m)}{\bar{y}_c} s_m \quad (3.16)$$

where 21 is the number of working days per month and $\bar{y}_c$ the mean yearly count. An 80% prediction interval is taken from the 10th and 90th percentiles of the Poisson distribution with mean $\lambda_{c,m}$. The seasonal index is notably above one in February (1.32) and November (1.28) and below one in January (0.75) and October (0.72). These months indicate when temporary capacity is most useful.

3.8.3 Risk Forecasting

Risk forecasting estimates, for each department and scenario, the probability that the department fails to answer at least 90% of its requests within the deadline. It is estimated by Monte Carlo over the R replications:



$$\hat{P}_{d,s} = \frac{1}{R} \sum_{r=1}^R \mathbb{1} \left[\text{SLA}_d^{(s,r)} < 0.9 \right] \quad (3.17)$$

A risk heatmap displays the resulting probabilities by department and scenario in the decision-support dashboard (Chapter 6). This presents the simulation output into the language of risk management required by Article 9 of the AI Act and familiar to municipal civil-protection planning.

3.9 Verification and Validation

Verification establishes that the model is implemented correctly, and validation that it represents the real system with sufficient accuracy for its intended purpose (Balci, 1997; Sargent, 2013). The intended purpose here is the *comparison* of operating policies and scenarios, not the point prediction of response times. The validation effort is therefore focused on relative behaviour and on the plausibility of mechanisms. Table 3.6 summarises the techniques applied.

Table 3.6 – Verification and validation techniques applied

Technique	Application	Outcome
Code verification and unit tests	Automated tests of the simulator, Erlang-C bounds and calibration	All tests pass; Erlang-C returns 1 for unstable queues and values in (0,1) otherwise
Conservation check	Arrivals = completions + open requests in every replication	Satisfied by construction and checked in the metrics
Analytical comparison	DES utilisation vs. Erlang-C utilisation per department (Table 3.7)	Consistent; DES slightly higher for the reasons explained below
Extreme-condition tests	Natural-disaster and infrastructure-incident scenarios	Unstable departments ($\rho > 1$) show growing backlogs, as theory predicts
Monotonicity (face validity)	What-if grid under the manual policy	SLA compliance falls with volume and rises with staffing, apart from differences within Monte Carlo noise
Statistical precision	Six replications with common random numbers	95% CI half-width of SLA compliance below one percentage point in the baseline
Input validation	Rates estimated from data; assumptions registered	Every non-measured parameter is declared in the assumption register

Table 3.7 compares, for the AI + HITL policy in the baseline scenario, the use of each department predicted by the analytical load model (Equations 3.11–3.12, utilisation λ_d $E[S_d]/c_d$) with the utilisation observed in the DES.

Table 3.7 – Analytical (Erlang-C load model) vs. simulated department utilisation, AI + HITL, baseline

Department	Officers	Analytical ρ	Simulated U_d	Difference
Urbanistica (Urban Planning)	2	0.416	0.450	+0.034
Protezione Civile (Civil Protection)	3	0.321	0.348	+0.027
Anagrafe (Civil Registry)	5	0.255	0.263	+0.008
Commercio (Commerce & Licensing)	2	0.114	0.149	+0.035
Edilizia (Building Permits)	4	0.110	0.127	+0.017
Ambiente (Environment)	2	0.096	0.132	+0.036
Polizia Locale (Local Police)	4	0.084	0.103	+0.019
Servizi Sociali (Social Services)	6	0.069	0.080	+0.011
Tributi (Taxation)	3	0.055	0.081	+0.026
Mobilità (Mobility & Transport)	3	0.045	0.070	+0.025



The two models agree on the ranking of departments by load and on the magnitudes, which supports the correctness of both implementations. The simulated utilisation is systematically higher by 0.01–0.04. This difference is expected and explained by mechanisms present only in the DES: the reading effort imposed on wrong departments by misrouted requests and the inter-departmental hand-offs that follow escalations. The departments with the largest positive differences (*Ambiente, Commercio*) are precisely those that receive hand-offs or a disproportionate share of misrouted files relative to their size. The analytical model is therefore slightly optimistic, which is why staffing recommendations are always re-evaluated in the DES.

The extreme-condition tests behave as queueing theory predicts. In the natural-disaster scenario the offered load of *Protezione Civile* rises to about 13.9 Erlangs against three officers ($\rho \approx 4.6$), and the Erlang-C optimiser requires 17 officers to stabilise the queue. In the simulation, the department accumulates a backlog of several thousand requests over the horizon. The model thus reproduces the qualitative behaviour expected of an overloaded civil-protection office during an alert.

Validation against historical response times is limited by a structural difference between the two quantities. Historical response times (median eleven days) include waiting for external agencies, inspections and document exchanges that are not resources of the request-handling system. Simulated response times measure the time a request spends in the municipal handling system. The simulated values are therefore expected to be substantially shorter, and they are used only for the relative comparison of policies and scenarios. Calibrating absolute response times would require workload logs from the Comune's protocol system (effort per file, actual staffing and queueing times), which were not available to this study. This is the principal limitation of the model and the first item of the future developments.

With these qualifications, the model is considered fit for its intended purpose: the comparative assessment of AI-adoption policies and contingency staffing under the scenarios of Table 3.4. Chapter 4 now presents the AI decision-support framework whose measured properties feed the model, and Chapter 5 reports the results of both.



CHAPTER 4 – TRUSTWORTHY AI DECISION-SUPPORT FRAMEWORK: DESIGN AND IMPLEMENTATION

This chapter presents the design and implementation of the trustworthy AI decision-support framework. The architecture and its ten functional blocks are introduced in Section 4.1. Sections 4.2–4.5 cover data acquisition, preprocessing and privacy protection, semantic knowledge construction, and storage. The hybrid retrieval pipeline is examined in Section 4.6, followed by the multi-agent decision engine in Section 4.7. Section 4.8 explains how the regulatory requirements identified in Chapter 1 are implemented as executable governance controls. Section 4.9 documents the technology stack. Throughout, design choices are linked to the functional and non-functional requirements in Section 1.6.

4.1 Overall Architecture

The framework is organised in ten functional blocks that follow the research methodology shown in Figure 4.1. Blocks 1 to 4 form the **data and knowledge layer**: they acquire multi-source municipal data, clean and anonymise it, convert regulations and historical records into indexed knowledge, and store that knowledge with access control. Blocks 5 and 6 form the **reasoning layer**: hybrid retrieval supplies evidence to a multi-agent engine that routes requests, assesses risk, drafts replies and decides how each request is handled. Block 7 is the **governance layer**, which operates across all others: it verifies outputs, explains them, monitors fairness, enforces human oversight and records every action. Blocks 8 to 10 form the **strategic layer**: the simulation environment (Chapter 3), the evaluation framework (Chapter 5) and the dashboard for municipal leaders (Chapter 6).

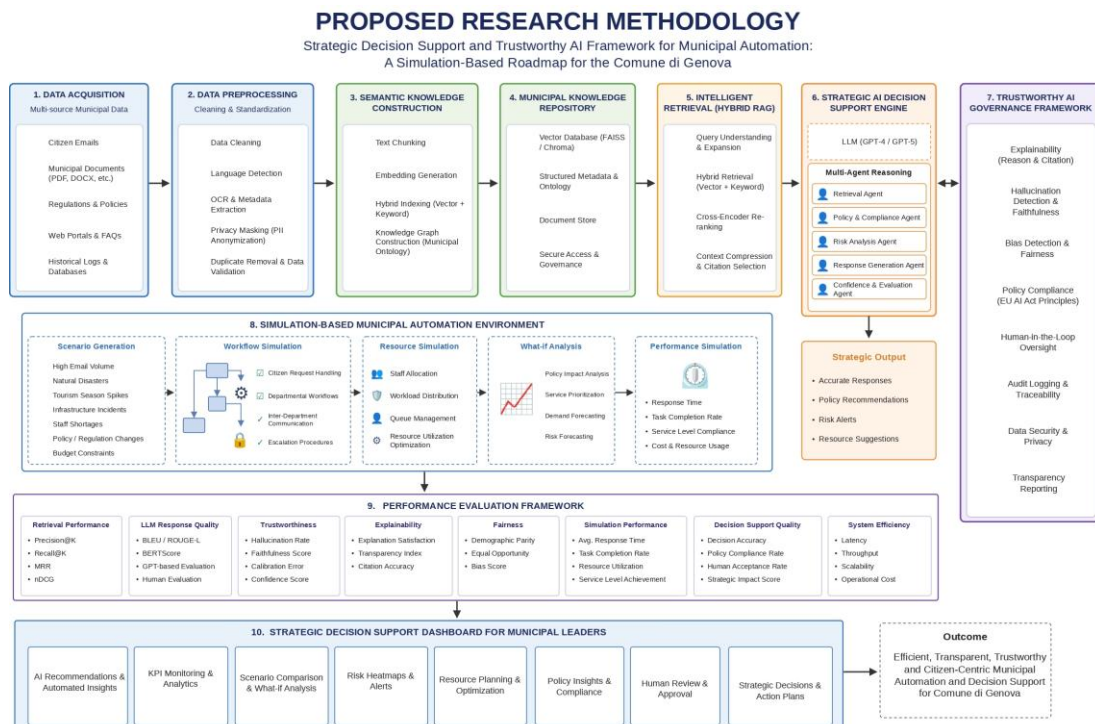


Figure 4.1: Proposed research methodology – The Strategic Decision Support and Trustworthy AI Framework for the Comune di Genova



Three architectural principles guide the design. The first is **governance by construction**: controls required by regulation, such as PII masking, citation, claim verification, review gating and audit logging, are pipeline stages that every request passes through rather than optional add-ons. The second is **inspectable intermediate state**: every agent writes its output to a shared blackboard, so the full reasoning chain behind any reply can be reconstructed. This satisfies the traceability requirement of Law 132/2025 and Article 12 of the AI Act. The third is **graceful degradation**: optional components (a GPT or Claude backend, multilingual neural embeddings, a neural cross-encoder, OCR) are loaded only if available, and the system falls back to deterministic offline components otherwise. The framework therefore runs on standard municipal hardware without external services (NFR10), and its behaviour in the offline configuration is a reproducible lower bound.

4.2 Datasets and Data Acquisition

4.2.1 Datasets

Three datasets represent the information environment of the Comune di Genova (Table 4.1). They are complementary: the requests are the operational input, the regulations are the authoritative knowledge, and the historical cases provide performance evidence about the current process. Because the citizen correspondence and case records of the Comune could not be released for research (Section 2.8), all three datasets were constructed synthetically with Python tools, as described in Section 4.2.2.

Table 4.1 – Datasets used in this work

Dataset	Records	Main fields	Role in the framework
Citizen requests	320 (319 after validation)	request ID, citizen name, e-mail, phone, age group, declared language, category, expected department, channel (web portal, e-mail, PEC), subject, body, urgency flag, date	Operational input; ground truth for routing evaluation; calibration of arrival mix and urgency
Municipal regulations	45 articles from 15 regulations	article ID, regulation, title, article number, category, department, text, effective date, source, language	Authoritative knowledge base for retrieval, citation and rule checks
Historical cases	500 (2022–2026)	case ID, category, department, age group, outcome, response time (days), year	Historical statistics documents; risk factors; bias audit; simulation calibration

4.2.2 Acquisition Connectors

Synthetic data generation. The datasets used in this work were produced with Python scripts based on pandas and NumPy, modelled on the departments, services, citizen channels and public regulatory framework of the Comune di Genova. Each generator draws records from explicit distributions rather than copying real cases, so that the data reproduce the structure of the municipal environment without containing any information about real people:

- **Citizen requests** (320 records dated between January and December 2026) combine a category and its competent department, a channel (web portal 119, PEC 103, e-mail 98), a language (Italian 265, English 55), an age group, an urgency flag and a subject and body drawn from a set of fifteen message templates. Names, e-mail addresses and telephone numbers are fictitious values created by the generator, so the masking and pseudonymisation stages of Section 4.3 can be tested on realistic identifiers without processing personal data.



- **Municipal regulations** (45 articles from 15 regulations) are written in the structure of municipal regulatory texts, with regulation identifier, article number, title, category, competent department, effective date and text, including the deadlines and obligations that the risk and policy agents check.
- **Historical cases** (500 records, 2022–2026) assign to each case a category, department, age group, outcome and response time in days, sampled from category-specific distributions so that the simulation model of Chapter 3 can be calibrated and the bias audit of Section 4.8.3 can be exercised.

Synthetic generation has three consequences for the work. First, it removes privacy risk entirely, which allows the datasets, code and results to be shared and the experiments to be reproduced. Second, the expected department of every request is known by construction, which provides exact ground truth for routing evaluation. Third, it limits external validity: template-based texts are more regular than real correspondence, which is why Chapter 5 evaluates the engine on independently written challenge sets, and the performance figures derived from the historical cases describe the synthetic dataset rather than measured performance of the Comune. The connectors described below make the framework independent of this choice: when real data become available, the same pipeline reads them without changes to the downstream components.

Block 1 provides a connector for every source type in the methodology, each returning a data frame with a fixed schema so that downstream components are independent of the origin of the data. Structured sources (request, regulation and historical-case files, or any SQL database of historical logs) are read directly. E-mail is read from individual .eml files or .mbox archives using the Python standard library: sender, subject, date and plain-text body are extracted, and category and department are left empty for the engine to fill. Municipal documents in PDF, DOCX, TXT and Markdown are converted to text, with scanned PDFs and images passed to optical character recognition. Web portals and FAQ pages are read from saved HTML or live URLs, with scripts, styles and navigation removed. Every acquisition records its provenance (source, path and number of records) as required by Article 10 of the AI Act.

4.3 Preprocessing, Privacy Protection and Data-Quality Audit

4.3.1 Cleaning and Normalisation

Text fields are normalised to Unicode NFC, non-breaking spaces and typographic quotes are replaced, zero-width characters are removed and whitespace is collapsed. A second normalisation, *accent folding*, maps accented characters to their base letters (for example, è to e). It is applied only for matching, because the regulation corpus is written without accents (*difficolta*, *e gestita*) whereas citizens type them. Without folding, lexical matching between citizen language and regulatory language would fail systematically.

4.3.2 Language Detection

Language is detected from content rather than taken from declared metadata. For a text with tokens $t_1, \dots, t_n$, the score of language ℓ is the share of tokens that belong to its stop-word list S_ℓ , and the detected language is the one with the highest score:

$$\text{score}(\ell) = \frac{1}{n} \sum_{j=1}^n \mathbb{1}[t_j \in S_\ell], \quad \hat{\ell} = \arg \max_{\ell \in \{\text{it}, \text{en}\}} \text{score}(\ell) \quad (4.1)$$



The method needs no external model and is adequate for the Italian–English mix of the data. It proved necessary: all 55 requests declared as English are written in Italian (Section 4.3.6), so replies based on the declared language would have been in the wrong language.

4.3.3 PII Masking and Pseudonymisation

Direct identifiers are removed from the text before any analytical processing (FR3; GDPR Arts. 5 and 25). Table 4.2 lists the categories detected and their replacement tokens. The citizen's name is also masked where it appears in the free text, using both a self-identification pattern (*mi chiamo ...*, *my name is ...*) and the known name from the request record.

Table 4.2 – Personal data detected and masked in free text

Category	Detection method	Replacement
E-mail address	Regular expression on local-part@domain	[EMAIL]
Telephone number	Italian mobile and landline patterns with optional +39 prefix	[PHONE]
Codice fiscale	16-character Italian tax-code pattern	[CODICE _{FISCALE}]
IBAN	Italian IBAN pattern	[IBAN]
Payment card	13–16 digit sequences	[CARD]
Personal name	Self-identification phrases and known name of the requester	[NAME]

Requests from the same citizen must remain linkable, for instance to detect repeated submissions, without exposing identity. Each citizen is therefore assigned a keyed pseudonym computed with HMAC-SHA-256 over the normalised e-mail address:

$$\text{pid}(x) = \text{CIT} \parallel \text{trunc}_{12} \left(\text{HMAC}_{\text{SHA256}}(K, \text{lower}(x)) \right) \quad (4.2)$$

where K is a secret key held outside the analytical environment. Without K the pseudonym cannot be reversed or recomputed from a known e-mail address, which distinguishes keyed pseudonymisation from a plain hash (GDPR Art. 4(5)). Names, e-mail addresses and telephone numbers are then removed from the working data set and stored separately in an encrypted vault (Section 4.5.4).

4.3.4 Metadata Extraction

Structured metadata is extracted from every request and regulation article. Deadlines are detected with their qualifier (*entro*, *almeno*, *fino a*, *within*), value and unit, and converted to calendar days, with working days weighted by 1.4. Other extracted fields are references to legal acts (for example, *DPR 223/1989*), monetary amounts, numbers of people (relevant to the 200-attendee rule for events), ages stated by the citizen (relevant to transport concessions for those over 70), Genoese places (Bisagno, Polcevera, Sampierdarena, *centro storico*, ZTL and others) and agencies (ARPAL, AMT, SUAP, TAR, COC). These fields feed the rule checks of the Policy and Compliance agent and the knowledge graph.

4.3.5 Validation and Duplicate Handling

Twenty-three automated checks are applied across the three datasets. They verify identifiers, schema completeness, domain values (categories in the regulatory taxonomy, departments consistent with categories, valid channels and age groups), temporal validity (parseable dates, dates not after the reference date of the study) and statistical plausibility (response-time outliers above $Q3 + 3 \cdot \text{IQR}$). Exact duplicate submissions (same citizen, text and date) are removed. Requests with identical text from different citizens are kept, because



they are distinct administrative requests, but they are assigned a *template identifier* so that the evaluation can group them and avoid leakage between development and test material.

4.3.6 Data-Quality Findings

Table 4.3 lists the checks that found issues. None is blocking, but four have direct consequences for the design.

Table 4.3 – Data-quality findings (checks with at least one affected record)

Dataset	Check	Affected	Consequence for the design
Requests	Declared language matches detected language	55 / 319	Reply language taken from detected language, not metadata
Requests	Distinct message texts	Only 15 distinct texts in 319 requests	Independent bilingual and noisy test sets needed for a valid evaluation (Chapter 5)
Requests	Date not after the reference date (8 October 2026)	70 / 319	Temporal validation at intake; future-dated records flagged
Requests	Exact duplicate submissions	1 / 320	Removed
Regulations	Distinct article texts	14 / 45 duplicated	Articles 2 and 3 follow a shared template across regulations
Regulations	Substantive (non-boilerplate) articles	30 / 45 boilerplate	Re-ranker favours specific over procedural text
Regulations	Single effective date per regulation	15 / 15 inconsistent	Effective dates handled per article; validity checked per cited article

4.4 Semantic Knowledge Construction

4.4.1 Municipal Ontology

The service domain is formalised as an ontology of fifteen service categories, each with an Italian–English label and a bilingual vocabulary of 179 domain terms in total (for example, *tumulazione*, *loculo*, *burial* for cemetery services; *ZTL*, *contrassegno*, *parking* for parking permits). The vocabulary was written from the regulation texts and general municipal terminology, not from the citizen messages, so that it is not fitted to the evaluation data. The ontology supports intent recognition, query expansion and the knowledge graph.

4.4.2 Tokenisation and Chunking

All lexical components share one tokenizer: text is lower-cased and accent-folded, split into alphanumeric tokens, stripped of Italian and English stop words, and reduced by a light prefix stemmer that keeps the first six characters of longer words. This stemmer conflates Italian inflections such as *iscrizione*, *iscrizioni* and *iscrivere*.

Regulation articles are split into chunks of complete sentences with at most 60 words and an overlap of one sentence between consecutive chunks. Sentence splitting is aware of the abbreviation *Art.*, which would otherwise split every article heading into a meaningless fragment. Each chunk inherits the metadata of its article: regulation, article number, category, department, effective date, source and boilerplate flag.

In addition to the 45 regulation articles, fifteen **historical-statistics documents** are generated, one per category, from the historical cases. Each states, in Italian and English, the number of cases, the median and 90th-percentile response times and the outcome shares. These documents make empirical performance retrievable and citable: a reply can tell a citizen that 90% of comparable cases were closed within a given number of days, citing the source. The knowledge base therefore contains 60 chunks.



4.4.3 Embedding Generation

Dense representations are produced by concatenating two TF-IDF spaces: word unigrams and bigrams over the stemmed tokens, and character n-grams of length 3 to 5 within word boundaries, weighted by 0.6. Character n-grams make the representation robust to inflection, compounding and typing errors. The concatenated vector is projected with truncated singular value decomposition (latent semantic analysis; Deerwester et al., 1990) to 128 dimensions when the corpus exceeds 256 chunks. For the present corpus of 60 chunks the projection is skipped, because it would discard discriminative information. Vectors are L2-normalised, so that the inner product equals cosine similarity:

$$\mathbf{v}(x) = \frac{[\text{tfidf}_{\text{word}}(x); 0.6 \text{tfidf}_{\text{char}}(x)]}{\|[\text{tfidf}_{\text{word}}(x); 0.6 \text{tfidf}_{\text{char}}(x)]\|_2}, \quad \text{sim}(q, d) \\ = \mathbf{v}(q)^\top \mathbf{v}(d) \quad (4.3)$$

When the *sentence-transformers* library and a multilingual model (paraphrase-multilingual-MiniLM-L12-v2) are available, neural sentence embeddings (Reimers and Gurevych, 2019) replace the TF-IDF representation through the same interface, adding cross-lingual semantic matching.

4.4.4 Knowledge Graph

The ontology is instantiated as a directed multigraph built with NetworkX, whose node and edge types are listed in Table 4.4. The graph contains 133 nodes and 344 edges.

Table 4.4 – Knowledge-graph schema and size

Node type	Count	Edge type (relation)	Count
Municipality	1	has _{department} (Municipality → Department)	10
Department	10	handled _{by} (ServiceCategory → Department)	45
ServiceCategory	15	governed _{by} (ServiceCategory → Regulation)	45
Regulation	15	has _{article} (Regulation → Article)	45
Article	45	imposes (Article → Obligation)	24
Obligation (deadline)	24	involves (Article → Agency)	61
Agency	12	refers _{toplace} (Article → Place)	39
Place	6	historical _{outcome} (ServiceCategory → Outcome)	75
Outcome	5		

The graph serves three purposes. It is the authoritative source for routing, because the department is resolved through the *handled_by* relation rather than inferred from text. It supplies structured facts, namely obligations, agencies and historical outcome counts, to the prompts of LLM backends and to explanations. It is also exported for inspection in the dashboard and stored as subject–relation–object triples in the metadata store.

4.5 Municipal Knowledge Repository

4.5.1 Vector Database

Chunk vectors are stored in a FAISS (Johnson et al., 2021) flat inner-product index (*IndexFlatIP*), which performs exact cosine search. Exact search is appropriate at municipal scale: the scalability test of Chapter 5 shows interactive latency up to 3,000 chunks. An approximate index (IVF or HNSW) can replace it through the same interface for larger corpora. When metadata filters are applied, for example restricting retrieval to one



department, the search is performed exactly on the filtered subset. If FAISS is not installed, an equivalent NumPy implementation is used.

4.5.2 Metadata and Document Stores

A SQLite database stores one row per chunk (identifier, document, category, department, regulation, article, source, effective date, boilerplate flag, text and full metadata) with indexes on category and department, together with the knowledge-graph triples. A JSON document store keeps the original documents. The vector index, chunk list, metadata database and knowledge graph are persisted under a single store directory, so the repository can be rebuilt, versioned and audited as a unit.

4.5.3 Role-Based Access Control

Access to data and functions is governed by role-based access control (Table 4.5). Every permission check, granted or denied, is written to the audit log with the user, role, permission and resource. Denied attempts are therefore traceable, which supports both security monitoring and the accountability requirements of Article 32 of the GDPR.

Table 4.5 – Roles and permissions

Role	Permissions
Citizen-service operator	Read masked data; process requests; view dashboard
Department officer	Read masked data; process requests; review and decide on drafts; view dashboard
Strategic manager	Read masked data; view dashboard; run simulations; view reports
Data protection officer	Read masked data; read personal data (vault); view audit log; view reports
Auditor	Read masked data; view audit log; view reports
Administrator	All permissions, including index management

4.5.4 Encrypted PII Vault

Direct identifiers removed during preprocessing are stored in a vault encrypted with Fernet (AES-128 in CBC mode with HMAC-SHA-256 authentication), with a key derived from the deployment secret. Only roles holding the *read_pii* permission can decrypt it, and each read is logged. Personal data therefore exists in exactly one protected location and never in the analytical stores, indexes or prompts.

4.6 Hybrid Retrieval Pipeline

Retrieval supplies the evidence on which every downstream decision rests. The pipeline (Figure 4.2) has five stages: query understanding, query expansion, hybrid candidate generation, re-ranking and context compression.

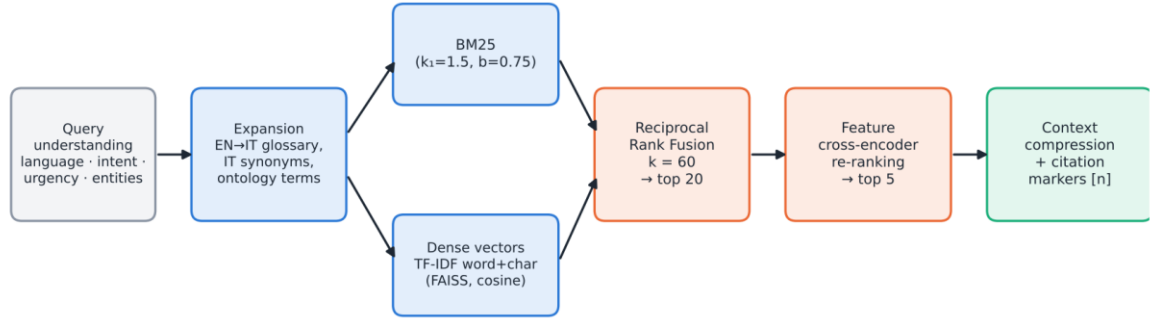


Figure 4.2: Hybrid retrieval pipeline

4.6.1 Query Understanding

The query-understanding stage extracts the language (Equation 4.1), urgency and safety signals, entities (Section 4.3.4) and an intent distribution over the fifteen categories. The intent score of category c counts the ontology terms found in the accent-folded query, with multi-word terms weighted 1.5 because they are more specific:

$$r_c(q) = \sum_{t \in T_c} (1 + 0.5 \cdot \mathbb{1}[t \text{ is multi-word}]) \mathbb{1}[t \in q], \quad I_c(q)$$

$$= \frac{r_c(q)}{\sum_{c'} r_{c'}(q)} \quad (4.4)$$

If no term matches, the distribution is uniform. Matching requires that a term is not preceded by a letter, which prevents false matches inside longer words. Urgency is detected from a list of 23 terms (*urgente, allerta, pericolosa, sta salendo, deceduto ieri, emergency* and others) and public-safety risk from a list of 10 terms.

4.6.2 Query Expansion

Citizens and regulations use different vocabularies, and the knowledge base is in Italian while some citizens write in English. Three expansion sources address this gap: a 43-entry English-to-Italian administrative glossary (*pothole* → *buca dissesto stradale*), 16 Italian synonym rules (*buca* → *dissesto stradale*) and, when the intent distribution is confident ($I_c \geq 0.5$), the first three ontology terms of the top category, a form of pseudo-relevance feedback. The expanded query repeats the original text twice to preserve its weight relative to the added terms.

4.6.3 Hybrid Candidate Generation

The expanded query is scored against every chunk by BM25 ($k_1 = 1.5, b = 0.75$), as defined in Section 1.4.2, and by cosine similarity in the embedding space (Equation 4.3). The two rankings are fused with reciprocal rank fusion with $k = 60$, and the 20 best-fused chunks become candidates for re-ranking. Rank fusion is preferred to score interpolation because BM25 and cosine scores are on incommensurable scales, and RRF requires no calibration between them.



4.6.4 Re-ranking

Candidates are re-ranked by a pairwise scorer that, like a cross-encoder, evaluates the query and the passage jointly. Six interaction features are computed for each query–chunk pair, and the re-ranking score is their weighted sum:

$$s_{rr}(q, d) = \mathbf{w}^T \mathbf{f}(q, d), \quad \mathbf{f} = \left[\cos, \frac{\text{BM25}}{\text{maxBM25}}, \text{cov}, I_{c(d)}(q), \text{spec}, \text{reg} \right] \quad (4.5)$$

Table 4.6 – Re-ranking features and weights

Feature	Definition	Weight
Dense similarity	Cosine similarity of query and chunk vectors	1.0
Normalised BM25	BM25 score divided by the best candidate's score	0.8
Query coverage	Share of original-query tokens present in the chunk	1.2
Intent agreement	Intent probability of the chunk's category	2.0
Specificity	1 if the chunk is not boilerplate (articles 2–3), else 0	0.6
Regulation source	1 if the chunk is a regulation article, 0 if a statistics document	0.2

The intent-agreement feature carries the largest weight because it injects the domain knowledge of the ontology into ranking; the ablation in Chapter 5 shows that removing the lexicon reduces routing accuracy markedly. The specificity feature addresses the boilerplate finding of Table 4.3. The weights can also be learned by logistic regression from relevance labels through the same interface, and a neural multilingual cross-encoder can be substituted when available. The five highest-scoring chunks are returned as evidence and labelled with citation markers [1] to [5].

4.6.5 Context Compression

Only the sentences of the evidence that are relevant to the request are passed to the generator, which reduces prompt length and the opportunity for the generator to use irrelevant material. Each sentence u is scored by its overlap with the expanded query, with a bonus for sentences that state a deadline or a quantity:

$$\text{rel}(u) = \frac{|\text{tok}(q^+) \cap \text{tok}(u)|}{|\text{tok}(u)|} + 0.3 \cdot \mathbb{1}[u \text{ states a deadline or quantity}] \quad (4.6)$$

Sentences with $\text{rel}(u) > 0.05$ are selected in order of relevance within each evidence chunk, up to a budget of 220 words. Each selected sentence keeps the citation marker of its chunk.

4.7 Multi-Agent Decision Engine

The decision engine is composed of five agents that run in sequence and communicate through a blackboard (Nii, 1986), a shared state to which each agent adds its outputs (Figure 4.3). An orchestrator executes the agents, times each step, attaches the explanation and writes the outcome to the audit log. When a request is not answered automatically, it also places it in the review queue.

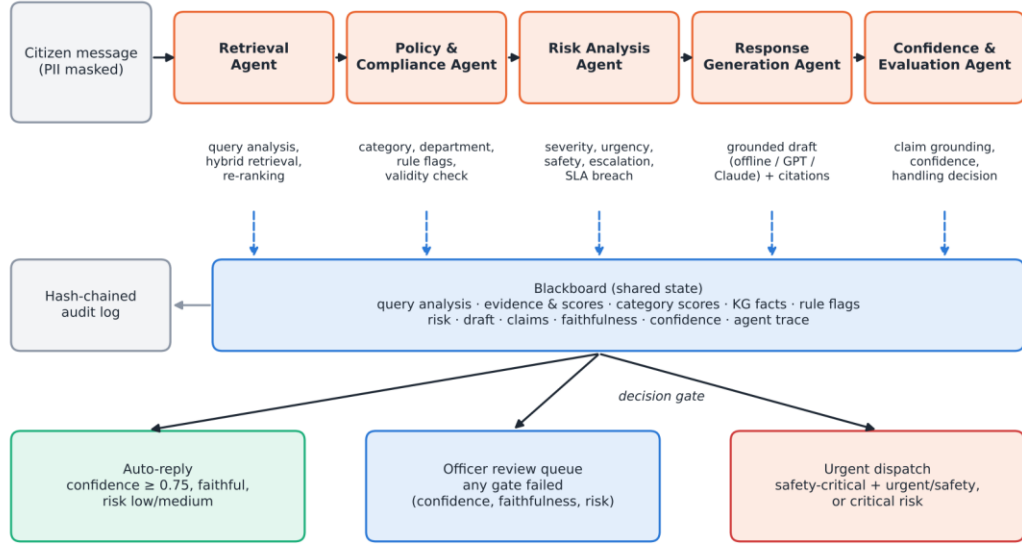


Figure 4.3: Multi-agent decision engine with blackboard, decision gate and audit log

4.7.1 Retrieval Agent

The Retrieval agent runs the pipeline of Section 4.6 on the masked request and writes the query analysis, the five evidence chunks with all their scores, the compressed context and the citation map. Its trace records the detected language, the top intent with its confidence, the expansion terms and the identifiers of the retrieved chunks.

4.7.2 Policy and Compliance Agent

This agent decides the service category and the competent department and identifies the rules that apply. Evidence chunks vote for their category with weights given by a softmax of their re-ranking scores at temperature $\tau = 0.5$. The final category score combines the vote with the intent distribution:

$$v_c = \sum_{k=1}^5 \frac{\exp(s_{rr,k}/\tau)}{\sum_j \exp(s_{rr,j}/\tau)} \mathbb{1}[c(d_k) = c], \quad S_c = 0.45 I_c(q) + 0.55 v_c \quad (4.7)$$

The predicted category is $\hat{c} = \operatorname{argmax}_c S_c$, and the department is resolved through the *handled_by* relation of the knowledge graph. The agent then applies rule checks derived from the regulations. It flags the need for a safety plan when an event is expected to exceed 200 attendees (REG-009), and the 30-day advance application. It flags the ISEE requirement for social benefits, waste-tax reductions and transport concessions; the over-70 concession category when the citizen states an age of 70 or more (REG-013); the five-working-day inspection for public-works reports with safety risk (REG-007); the 48-hour procedure for urgent burials (REG-012); activation of the COC under orange or red alert (REG-010); and the joint ARPAL and Local Police check for environmental reports (REG-008). Finally, it checks the **temporal validity** of every cited regulation article: an article whose effective date is later than the request date is flagged, because citing a rule not yet in force would breach the duty to state correct legal reasons.



4.7.3 Risk Analysis Agent

The Risk Analysis agent computes a risk score in $[0, 1]$ from additive factors (Table 4.7) and maps it to four levels.

$$R = \min(1, \sigma_c + 0.25 \mathbb{1}_{urg} + 0.20 \mathbb{1}_{safety} + 0.5 \text{esc}_c + 0.3 \text{breach}_c) \quad (4.8)$$

Table 4.7 – Risk factors and levels

Factor	Definition	Contribution
Category severity σ_c	Emergency 0.60; flood risk and environment 0.25; public works 0.20; cemetery and social benefit 0.15; noise 0.10; others 0.05	σ_c
Urgency	Urgency terms detected in the request	+0.25
Public-safety risk	Safety terms detected (danger, collapse, flooding)	+0.20
Historical escalation	Share of past cases of the category that were escalated	$0.5 \times \text{esc}_c$
Historical SLA breach	Share of past cases of the category closed after the regulatory deadline	$0.3 \times \text{breach}_c$
Levels	low $< 0.25 \leq$ medium $< 0.50 \leq$ high $< 0.75 \leq$ critical	—

The historical factors connect the risk of an individual request to the measured weaknesses of the current process. A cemetery request mentioning a death on the previous day is rated high risk not only because of its urgency, but because every comparable past case missed the 48-hour target. For high and critical levels the agent produces an alert naming the department and the main factors, which feeds the risk dashboard.

4.7.4 Response Generation Agent

The Response Generation agent drafts the reply in the detected language through one of three interchangeable backends. With an OpenAI or Anthropic API key configured, an LLM is prompted with a system instruction that constrains it to use only the supplied context and facts, cite every factual sentence with its marker, keep Italian legal terms unchanged, never promise approvals, amounts or dates not present in the context, never decide eligibility, and end with a fixed AI-disclosure sentence (AI Act Art. 50). The user message contains the masked request, the compressed context and the structured facts from the knowledge graph and rule flags.

Without an API key, an **offline grounded generator** composes the reply extractively. It writes the routing statement, an urgency notice when applicable, and the substantive sentences of the most specific regulation article with their citation. It then adds the historical timing from the statistics document, the procedural sentences on how to apply and how to appeal, and, for eligibility-related services, a statement that a municipal officer will take the final decision. In English replies, regulation sentences are quoted in the original Italian. If an API call fails, the agent falls back to the offline generator, so the system never returns an empty reply. Safety notices (call 112 in immediate danger), the AI disclosure and the human-decision statement are drawn from a registry of verified **standard notices** approved by the administration, not generated freely.

4.7.5 Confidence and Evaluation Agent

The last agent verifies the draft (Section 4.8.1), computes a confidence score and decides how the request is handled. Confidence combines five components, each in $[0, 1]$ (Table 4.8):

$$\text{Conf} = 0.30 m + 0.25 g + 0.15 \rho + 0.20 \phi + 0.10 I_c(q) \quad (4.9)$$



Table 4.8 – Components of the confidence score

Component	Definition	Weight
Routing margin m	$\min(1, (S_{(1)} - S_{(2)}) / 0.5)$: separation between the best and second-best category	0.30
Evidence agreement g	Share of the five evidence chunks belonging to the predicted category	0.25
Retrieval strength p	Logistic transform of the top re-ranking score: $1 / (1 + \exp(-(s_{rr} - 2.5)))$	0.15
Faithfulness ϕ	Share of claims in the draft supported by sources (Equation 4.11)	0.20
Intent confidence	Probability of the top intent from the ontology lexicon	0.10

The handling decision follows the rules in Table 4.9, applied in order. The rules encode the human-oversight principle of the AI Act and Law 132/2025. Automation is permitted only where confidence, grounding and risk are all favourable. Safety-critical situations go directly to the duty officer, and the case decision for eligibility-related services always remains with an officer, even when the informational reply is sent automatically.

Table 4.9 – Handling decision rules

Decision	Condition
Urgent dispatch	Category is safety-critical (emergency report, flood-risk inquiry, public-works complaint) and urgency or safety language is present; or risk level is critical
Officer review	Any of: confidence < 0.75; faithfulness < 0.80; risk high or critical; a cited article not yet in force on the request date
Auto-reply	None of the above conditions holds
Case decision	For eligibility-related services (social benefits, school enrolment, waste tax, transport passes, permits, licences, residency), always taken by an officer, whatever the reply path

4.8 Governance Layer

4.8.1 Hallucination Detection and Faithfulness

Every draft, whether produced by an LLM or by the offline generator, is decomposed into claims: sentences of at least four words, excluding salutations, signatures and headings. Each claim is checked against a set of **support units** comprising the sentences of the retrieved evidence, the standard notices and the system's own routing statements, which are decisions rather than claims about the world. For a claim a and a support unit u , support is the share of the claim's content tokens found in the unit, penalised when the claim contains numbers absent from the unit:

$$\text{sup}(a, u) = \frac{|\text{tok}(a) \cap \text{tok}(u)|}{|\text{tok}(a)|} \cdot \begin{cases} 0.3 & \text{if } \text{num}(a) \not\subseteq \text{num}(u) \\ 1 & \text{otherwise} \end{cases} \quad (4.10)$$

A claim is supported if $\max_u \text{sup}(a, u) \geq 0.5$. The faithfulness of a draft is the share of supported claims, and its hallucination rate is the complement:

$$\phi = \frac{1}{|A|} \sum_{a \in A} \mathbb{1} \left[\max_u \text{sup}(a, u) \geq 0.5 \right], \quad \text{HR} = 1 - \phi \quad (4.11)$$

The numeric penalty targets the most consequential form of hallucination in administrative correspondence, a wrong deadline, amount or threshold, which lexical overlap alone would not detect. For cited claims the detector also checks **citation correctness**, that is, whether the cited chunk is itself a supporting unit. Unsupported claims are listed in the explanation and shown to the reviewing officer. Chapter 5 validates the detector by fault injection.



4.8.2 Explainability

Each decision carries a structured explanation with four parts. A natural-language summary states why the category was chosen (with the score of the runner-up), how the department was resolved, the risk level and its factors, the confidence and the handling decision, and the reasons for review where applicable. The evidence table lists each cited chunk with its source, article, category, re-ranking score and the contribution of each feature of Table 4.6. The category ranking gives the top three categories with their scores. Finally, the explanation includes the confidence components and the rule flags. The agents' trace, with timings, completes the record. This design gives the officer local, decision-specific reasons rather than a global description of the model, which is what Article 13 of the AI Act and the Italian principle of *conoscibilità* require at the individual act.

4.8.3 Fairness Monitoring and Bias Audit

Two complementary analyses address non-discrimination. **Monitoring of the system's outputs** computes, for each protected or operational attribute (age group, detected language, channel), the auto-reply rate per group and derives demographic parity, disparate impact (Feldman et al., 2015) and equal opportunity (Hardt et al., 2016):

$$\begin{aligned} DP &= \max_g P(\hat{y} = 1 | g) - \min_g P(\hat{y} = 1 | g), \quad DI \\ &= \frac{\min_g P(\hat{y} = 1 | g)}{\max_g P(\hat{y} = 1 | g)} \end{aligned} \quad (4.12)$$

$$EO = \max_g P(\hat{y} = 1 | y = 1, g) - \min_g P(\hat{y} = 1 | y = 1, g) \quad (4.13)$$

Here $\hat{y} = 1$ denotes an automatic reply and $y = 1$ a request whose correct routing and faithful draft make automatic handling legitimate. Routing accuracy per group is also compared. The four-fifths rule ($DI \geq 0.8$; Feldman et al., 2015) is used as the alert threshold.

The **audit of historical decisions** tests whether past outcomes differ by age group, because historical data that is biased would contaminate any model or statistic derived from it. It applies a χ^2 test of independence between outcome and age group (with Cramér's V as effect size), a Kruskal–Wallis test on response times, and per-category χ^2 tests whose p-values are corrected for multiple testing with the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995). A category is flagged only when its disparate impact is below 0.8 *and* the corrected q-value is below 0.10. This avoids raising alarms on small-sample noise, which is a real risk with about 30 cases per category.

4.8.4 Regulatory Compliance Checks

At the end of each run, eleven compliance checks are evaluated automatically from measured facts and reported with a status of PASS, WARN or FAIL and the evidence used. They cover AI Act Articles 6 and Annex III, 9, 10, 11–12, 13, 14, 15 and 50, GDPR Articles 5, 22, 25 and 32, and the fairness requirement of Article 10(2)(f). For example, the check for Article 14 passes only if no eligibility decision was taken automatically. The Article 50 check passes only if every reply contains the AI disclosure, and the GDPR data-minimisation check passes only if no direct identifier is present in the working data set. The checklist and a generated system card, documenting intended use, components, measured performance and limitations, form the technical documentation required by Article 11.



4.8.5 Human-in-the-Loop Review

Requests that are not auto-replied enter a persistent review queue (SQLite) with a priority derived from the risk level (critical first), the predicted department and category, the confidence, the reasons for review and the draft. Officers filter the queue by department, read the reasons and the explanation, and then approve the draft, edit and approve it, or reject it with a comment. Each action is recorded in the queue and in the audit log with the officer's identity. Edited drafts are retained, which gives the administration a growing set of officer-corrected replies for future evaluation and improvement.

4.8.6 Tamper-Evident Audit Log

All significant events are written to an append-only audit log in JSON Lines format: system builds, index construction, access checks, request processing (with category, department, risk, confidence, decision, faithfulness, cited chunks and generator), review enqueueing, review decisions and simulation runs. Each record contains the hash of the previous record and its own hash, forming a hash chain (Figure 4.4):

$$h_i = \text{SHA256}(\text{canon}(ts_i, \text{event}_i, \text{actor}_i, \text{details}_i, h_{i-1})), \quad h_0 = 0^{256} \quad (4.14)$$

hash = SHA-256(record without hash). Editing or deleting any record breaks every later link, which verify() detects.

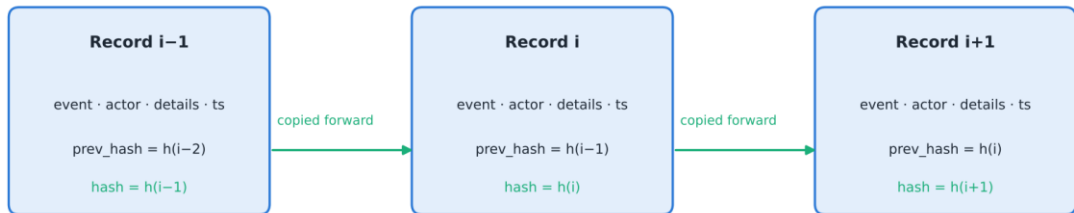


Figure 4.4: Hash-chained audit log

where canon denotes canonical JSON serialisation with sorted keys. Verification recomputes every hash and checks every link, following the forward-integrity principle of secure audit logs (Schneier and Kelsey, 1999); modifying, inserting or deleting any record breaks the chain at that point and is detected, with the position of the first inconsistency reported. Writes are serialised with a lock so that concurrent processes cannot interleave records. The log satisfies the record-keeping obligation of Article 12 of the AI Act and the traceability requirement of Law 132/2025, and it makes the audit trail itself auditable. For deployment, the head hash can additionally be anchored periodically to the municipal protocol system or a timestamping service, which protects against wholesale replacement of the log.

4.9 Implementation and Technology Stack

The framework is implemented in Python as a package of sixteen modules, one per methodology block plus shared configuration and utilities, with about 5,700 lines of code including the user interfaces, scripts and tests (Table 4.10). All parameters and assumptions are centralised in a configuration module, and every external dependency beyond the scientific Python stack is optional.



Table 4.10 – Software modules and technologies

Block	Module	Main technologies
1 Data acquisition	data _{acquisition} .py	pandas; email/mailbox; pypdf, python-docx, BeautifulSoup (optional)
2 Preprocessing	preprocessing.py	Regular expressions; HMAC-SHA-256; OCR via Tesseract (optional)
3 Knowledge construction	knowledge.py	scikit-learn TF-IDF / SVD; BM25; NetworkX; sentence-transformers (optional)
4 Repository	repository.py, audit.py	FAISS; SQLite; cryptography (Fernet)
5 Retrieval	retrieval.py	Hybrid BM25 + dense; RRF; feature or neural cross-encoder
6 Decision engine	agents.py, llm.py, strategy.py	Blackboard multi-agent pipeline; OpenAI / Anthropic SDKs (optional)
7 Governance	governance.py	SciPy statistical tests; SQLite review queue; system card generator
8 Simulation	simulation.py	SimPy; NumPy; Erlang-C; Poisson GLM
9 Evaluation	evaluation.py, challenge _{set} .py	Retrieval, generation, calibration and fairness metrics; fault injection
10 Dashboard	app/dashboard.py, app/web, app/api.py	Streamlit and Plotly; standalone HTML/JavaScript tool; FastAPI REST service
Orchestration	pipeline.py, run.py, scripts/	End-to-end pipeline; one-command launcher; pytest suite (18 tests)

The system is exposed through four interfaces that serve different users: a REST API for integration with the municipal protocol and web-portal systems, a Streamlit dashboard for officers and managers, a standalone HTML tool that runs entirely in the browser, and a command-line interface for single requests. A single launcher builds the knowledge base and results when needed and starts the HTML tool on a local server. The complete pipeline, from raw data to evaluation, simulation and compliance report, runs in under two minutes on a standard laptop, and the processing of an individual request takes a few milliseconds in the offline configuration. Chapter 5 evaluates the framework, and Chapter 6 describes its user interfaces and deployment.



CHAPTER 5 – EXPERIMENTAL EVALUATION AND RESULTS

This chapter evaluates the framework against the measures defined in Section 5.1. Sections 5.2–5.6 examine the AI decision-support engine, including retrieval performance, request routing, response quality, trustworthiness, explainability and fairness. Section 5.7 assesses the operational effects predicted by the Chapter 3 simulation model. Implementation efficiency and scalability are reported in Section 5.8, followed by the automated regulatory compliance assessment in Section 5.9. Finally, Section 5.10 considers what the results imply for the research questions in Section 1.6.1 and where the evidence remains limited.

5.1 Evaluation Design

5.1.1 Test Sets

The data-quality audit (Table 4.3) showed that the 319 validated citizen requests contain only fifteen distinct message texts, one per service category. Performance measured on these requests alone therefore reflects how well the system handles fifteen specific texts, not how well it generalises to the variety of real correspondence. To obtain a valid estimate, the evaluation uses three test sets:

- **Provided requests** (n = 319; 15 distinct texts). Used for all metrics, and for retrieval metrics computed once per distinct text so that duplicates do not inflate the averages.
- **Clean challenge set** (n = 60). Independently written requests, four per category (two in Italian, two in English), with wording, situations and vocabulary that differ from the provided texts, for example a bar opening in Sestri Ponente, a forest fire above houses or a burial at Staglieno. The set was written for evaluation only and was not used to build the ontology or to tune any parameter.
- **Noisy challenge set** (n = 120). Two perturbed variants of each challenge request, simulating real-world input defects. Each word longer than three characters has a probability of 0.08 of a typing error (swapped, deleted or duplicated character) and 0.08 of being dropped; accents are stripped with probability 0.5 and the text is lower-cased with probability 0.5.

Results on the provided requests are reported for completeness, but conclusions about generalisation are drawn from the two challenge sets. The noisy set is used as the reference for the conservative figures fed into the simulation (Section 3.4.1).

5.1.2 Configuration

All results were obtained in the offline configuration of the framework: TF-IDF word-and-character embeddings, a FAISS exact inner-product index, the feature-based re-ranker and the offline grounded generator, with random seed 42. This configuration requires no external service and is fully reproducible. It is also a conservative baseline for the components that optional neural models would replace (Section 5.10.3).

5.1.3 Metrics

The metric families follow the evaluation framework of the methodology (block 9). Retrieval relevance is graded per chunk: 3 for a substantive article of the request's category, 2 for the



category's historical-statistics document, 1 for a procedural (boilerplate) article of the category, and 0 for any chunk of another category. With rel_i the relevance of the chunk at rank i and R the set of relevant chunks, the retrieval metrics are

$$P@K = \frac{1}{K} \sum_{i=1}^K \mathbb{1}[rel_i > 0], \quad R@K = \frac{1}{|R|} \sum_{i=1}^K \mathbb{1}[rel_i > 0], \quad MRR$$

$$= \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{rank_q} \quad (5.1)$$

$$nDCG@K = \frac{DCG@K}{IDCG@K}, \quad DCG@K = \sum_{i=1}^K \frac{2^{rel_i} - 1}{\log_2(i + 1)} \quad (5.2)$$

where $rank_q$ is the rank of the first relevant chunk for query q , and $IDCG@K$ is the DCG of the ideal ordering (Järvelin and Kekäläinen, 2002). Calibration of the confidence score regarding routing correctness is measured by the expected calibration error (Guo et al., 2017) over $B = 10$ equal-width bins and by the Brier score (Brier, 1950):

$$ECE = \sum_{b=1}^B \frac{|B_b|}{N} |acc(B_b) - conf(B_b)|, \quad \text{Brier} = \frac{1}{N} \sum_{i=1}^N (conf_i - y_i)^2 \quad (5.3)$$

Response quality is measured by BLEU (Papineni et al., 2002), with add-one smoothing, ROUGE-L (Lin, 2004) and a BERTScore-style soft-matching F1 (Zhang et al., 2020). Because the datasets contain no human-written reference replies, the reference for each request is the text of articles 1 and 2 of the regulation of its true category. These metrics therefore measure how much of the governing regulation a reply conveys, not its stylistic quality. Faithfulness, hallucination rate and citation correctness are defined in Section 4.8.1, and the fairness metrics in Section 4.8.3.

5.2 Retrieval Performance

Table 5.1 reports the retrieval metrics of the four retrieval configurations on the three test sets, and Figure 5.1 shows the main metrics for the noisy set.

Table 5.1 – Retrieval performance by configuration and test set

Test set	Configuration	P@1	P@5	R@5	MRR	nDCG@5
Provided (15 texts)	BM25	0.933	0.600	0.750	0.967	0.868
	Dense	0.933	0.600	0.750	0.967	0.854
	Hybrid (RRF)	0.933	0.600	0.750	0.967	0.867
	Hybrid + re-rank	1.000	0.787	0.983	1.000	0.977
Challenge clean (60)	BM25	0.900	0.340	0.425	0.914	0.712
	Dense	0.867	0.360	0.450	0.910	0.712
	Hybrid (RRF)	0.900	0.350	0.438	0.926	0.717
	Hybrid + re-rank	0.883	0.677	0.846	0.914	0.881
Challenge noisy (120)	BM25	0.833	0.323	0.404	0.855	0.669
	Dense	0.833	0.348	0.435	0.875	0.685
	Hybrid (RRF)	0.850	0.343	0.429	0.884	0.684
	Hybrid + re-rank	0.842	0.648	0.810	0.869	0.844

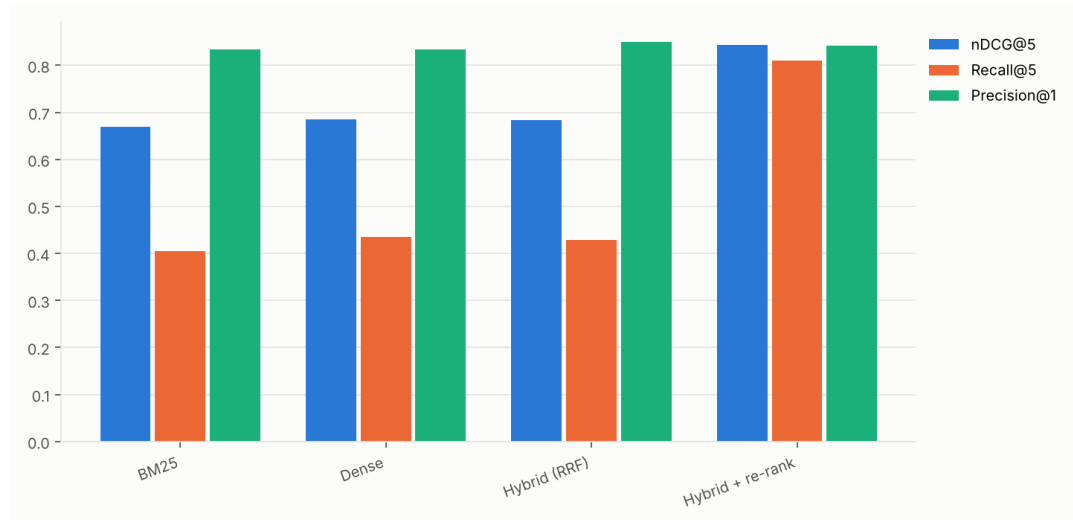


Figure 5.1: Retrieval quality by configuration on the noisy challenge set

Three observations stand out. First, all configurations find a relevant chunk at the first rank in most cases ($P@1$ between 0.83 and 0.93 on the challenge sets). The difference lies in the *completeness* of the evidence. Without re-ranking, only about 40% of the relevant chunks appear among the top five ($R@5 = 0.40$ – 0.45); with re-ranking the share rises to 81–85%, and $nDCG@5$ rises from 0.67–0.72 to 0.84–0.88. Completeness matters for the downstream agents. The reply needs the substantive article, the statistics document and the procedural articles of the *same* category, and the evidence-agreement component of the confidence score depends on how many of the five chunks belong to it.

Second, BM25 and dense retrieval perform similarly, and their fusion by RRF brings only a marginal gain. With a small, homogeneous corpus and a TF-IDF embedding, the two rankers share much of their signal, so fusion has little complementary information to exploit. The re-ranker contributes instead through features the first-stage rankers lack: intent agreement with the ontology, and the specificity feature that demotes the boilerplate articles shared by all regulations.

Third, performance degrades gracefully with noise. Between the clean and noisy challenge sets, $nDCG@5$ of the full pipeline falls by only 0.04 (0.881 to 0.844), which reflects the robustness of character n-grams and accent folding to typing errors and missing accents.

5.3 Routing and Decision-Support Quality

5.3.1 Routing Accuracy and Handling Decisions

Table 5.2 summarises the decision-support metrics. Category and department accuracy coincide on all sets, because the department is derived deterministically from the category through the knowledge graph.

Table 5.2 – Routing accuracy and handling decisions

Metric	Provided	Challenge clean	Challenge noisy
Department routing accuracy	1.000	0.883	0.842
Auto-reply share	0.787	0.700	0.692
Officer review share	0.091	0.250	0.267



Metric	Provided	Challenge clean	Challenge noisy
Urgent dispatch share	0.122	0.050	0.042
Misrouted requests	0 / 319	7 / 60	19 / 120
Misrouted requests sent to officer review	—	6 of 7	15 of 19
Misrouted requests auto-replied (unreviewed error)	0.0%	1.7%	3.3%
Mean confidence, correct / misrouted	0.870 / —	0.808 / 0.651	0.807 / 0.575
Policy-compliance rate	1.000	1.000	1.000

Accuracy of 0.883 on clean and 0.842 on noisy independent requests indicates a capability that is useful in practice but not sufficient for unsupervised routing. This is precisely why the framework does not rely on accuracy alone. Confidence is markedly lower for misrouted requests (0.575 against 0.807 on the noisy set), and the decision gate sends 15 of the 19 misrouted noisy requests and 6 of the 7 misrouted clean requests to an officer. As a result, only 3.3% of noisy requests (1.7% of clean requests) are both misrouted and auto-replied without review. This **unreviewed error rate** is the operationally relevant measure of risk, because it counts the errors that reach citizens without human control.

5.3.2 Error Analysis

Table 5.3 lists the misrouting patterns on the noisy set. Errors concentrate in two categories whose requests are intrinsically ambiguous.

Table 5.3 – Most frequent misrouting patterns (noisy challenge set)

True category	Predicted category	Cases	Interpretation
Emergency report	Public-works complaint	4	Landslide blocking a road, water entering garages: emergency and infrastructure overlap; both are safety-critical and dispatched with priority
Business licence	Event authorisation / noise complaint / transport pass	5	Bar, bakery or market-stall openings mention music, events or locations whose terms belong to other services
Environmental report	Waste tax	2	Abandoned hazardous waste shares the vocabulary of waste collection (<i>rifiuti</i>)
Residency change	Waste tax	2	Moving house triggers lexical overlap with the TARI registration of a new address
Other	Various	6	Single cases, mostly caused by dropped key words in the noisy variants

The emergency-to-public-works confusion is the least harmful: both categories are safety-critical, and urgent requests in either are dispatched to an officer immediately. The business-licence confusions point to a gap in the ontology vocabulary for commercial activities (*bar*, *panetteria*, *posteggio*), which is the simplest improvement available. Only lexicon terms need adding, and the change can be verified against the same test sets.

5.3.3 Ablation Study

Table 5.4 isolates the contribution of each component to routing accuracy on the 180 challenge requests (clean and noisy combined).



Table 5.4 – Ablation study of routing accuracy (challenge sets, n = 180)

Configuration	Retrieval	Category accuracy
Full system	BM25 only	0.861
Full system	Dense only	0.867
Full system	Hybrid (RRF)	0.844
Full system	Hybrid + re-rank	0.856
Without intent lexicon (evidence vote only)	Hybrid + re-rank	0.689
Without query expansion	Hybrid + re-rank	0.883
Lexicon only (no evidence vote)	Hybrid + re-rank	0.778

The ontology lexicon is the most important component: removing it lowers accuracy by 17 points, from 0.856 to 0.689. Combining the lexicon with the evidence vote adds 8 points over the lexicon alone (0.778), which confirms that the two sources of evidence are complementary. Routing accuracy is largely insensitive to the retrieval configuration (0.844–0.867), because routing depends mainly on the top-ranked chunks, where all configurations perform similarly (Table 5.1).

Query expansion slightly *reduces* routing accuracy, by 2.7 points. The pseudo-relevance step adds ontology terms of the top intent whenever the intent is confident, and when that intent is wrong the expansion reinforces the error. Expansion remains useful for retrieval completeness and for English requests, but the result suggests restricting pseudo-relevance feedback to cases where lexicon and retrieval agree. The finding is reported as observed; the system was not re-tuned on the test sets.

5.3.4 Urgency Detection

Against the *urgent* label of the provided requests, urgency detection (urgency terms or urgent dispatch) achieves recall 0.675 and precision 0.397 (F1 = 0.50, 40 positive cases). The label itself is unreliable outside emergency reports: identical routine texts, such as school-enrolment enquiries, are labelled urgent in some records and not in others, so no text-based detector can reproduce it. Every emergency report, the category in which urgency is genuine and unambiguous, is detected and dispatched urgently.

5.4 Response Quality

Table 5.5 – Response-quality metrics against the regulation reference

Metric	Provided	Challenge clean	Challenge noisy
BLEU	0.368	0.340	0.331
ROUGE-L (F1)	0.491	0.441	0.428
BERTScore (soft-match proxy, F1)	0.704	0.609	0.595

The values in Table 5.5 indicate that replies convey a substantial part of the governing regulation, and they decrease moderately as test conditions move away from the provided texts, in line with routing accuracy. Their absolute values should be interpreted with care: the reference is regulation text rather than a model reply, and the extractive generator quotes regulation sentences, which favours n-gram overlap. The proxy BERTScore uses character-trigram similarity in place of contextual embeddings and is automatically replaced by the original BERTScore when the relevant library is installed. GPT- or Claude-based quality judgement is implemented but requires an API backend, and was not run in the offline configuration. For human evaluation, a stratified sample of 30 replies (two per category) has



been exported with a five-point rating form for relevance, correctness, completeness, clarity and explanation satisfaction, ready for annotation by municipal officers.

5.5 Trustworthiness

5.5.1 Faithfulness and Hallucination Detection

All drafts on all three test sets are fully grounded: faithfulness is 1.0 and the hallucination rate 0.0. In the offline configuration this result holds by construction, since the extractive generator only states facts taken from evidence or from the registry of standard notices. The informative question is therefore whether the *detector* would catch hallucinations produced by an LLM backend. It was tested by fault injection on 379 drafts (the 319 provided requests and the 60 clean challenge requests). Two fabricated claims were added to each draft from a list of plausible but unsupported statements, such as *the maximum benefit is 1,500 euros per month* or *the application is approved automatically within 3 days*. Every claim containing a number was also copied with the number altered.

Table 5.6 – Hallucination-detector validation by fault injection

Measure	Value	Basis
Recall on fabricated claims	96.8%	758 injected claims
Recall on altered numbers	100.0%	774 corrupted claims
False-alarm rate on genuine claims	0.0%	3,008 original claims

The detector flags almost all fabricated statements and every altered number, without flagging any genuine claim. The few undetected fabrications share most of their content words with real regulation sentences (for example, statements about documents to attach), which is the expected limit of a lexical support criterion. A natural-language-inference model could close this gap in a future version, at a higher computational cost.

5.5.2 Calibration

Figure 5.2 shows the reliability diagram of the confidence score regarding routing correctness over the pooled 499 decisions of the three test sets. The pooled ECE is 0.120; per set, ECE is 0.131 (provided), 0.105 (clean) and 0.101 (noisy), with Brier scores of 0.018, 0.104 and 0.108.

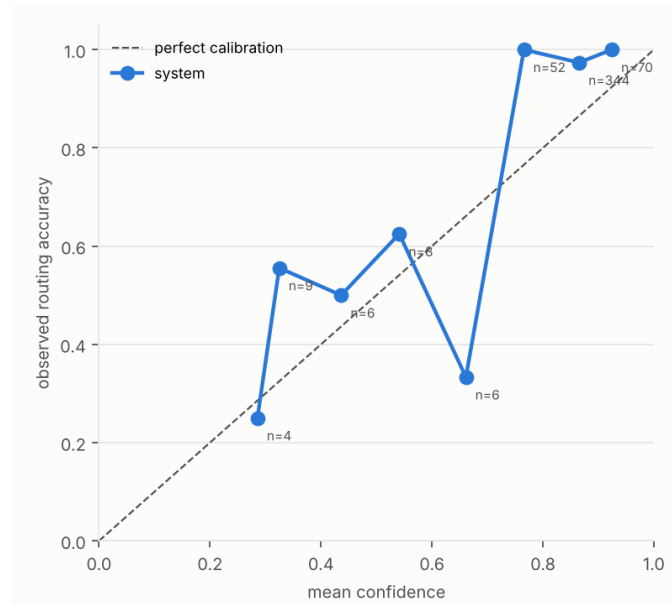


Figure 5.2: Reliability diagram of the confidence score (pooled test sets; n per bin shown)

The miscalibration is asymmetric. Above 0.7, where 466 of the 499 decisions fall, the score is *under-confident*: observed accuracy is 97–100% against mean confidence of 0.77–0.92. Below 0.7 the few decisions (33) are correct only 25–63% of the time. for the decision gate this behaviour is acceptable: the threshold of 0.75 separates a high-accuracy region from a low-accuracy one, and under-confidence errs on the side of human review. If the score were to be shown to officers as a probability, it should first be recalibrated, for example with isotonic regression fitted on officer-validated decisions.

5.6 Explainability and Fairness

5.6.1 Explainability

Every decision on every test set carries a complete explanation (transparency index 1.0): a natural-language summary, cited evidence with feature contributions, the confidence components, the AI disclosure, and a trace of all five agents. Citation accuracy, the share of cited chunks belonging to the true category of the request, is 1.000 on the provided requests, 0.883 on the clean and 0.842 on the noisy challenge set. Citations are therefore accurate whenever routing is accurate. When routing fails, the citations reveal the error to the reviewing officer, who sees immediately that the reply cites, for instance, the TARI regulation for a residency question. Explanation satisfaction requires a study with officers and is part of the exported evaluation form.

5.6.2 Fairness of Automated Handling

Table 5.7 reports fairness metrics for automatic replies across age groups and channels (provided requests) and across languages (challenge sets). Figure 5.3 shows the auto-reply rate by age group.



Table 5.7 – Fairness of automated handling

Attribute (test set)	Auto-reply rate by group	Parity difference	Disparate impact	Routing accuracy gap
Age group (provided)	18–30: 0.83; 31–50: 0.80; 51–70: 0.79; 70+: 0.74	0.089	0.892	0.000
Channel (provided)	PEC: 0.75; e-mail: 0.80; web portal: 0.82	0.070	0.914	0.000
Language (challenge clean)	English: 0.80; Italian: 0.60	0.200	0.750	0.033
Language (challenge noisy)	English: 0.75; Italian: 0.63	0.117	0.844	0.017

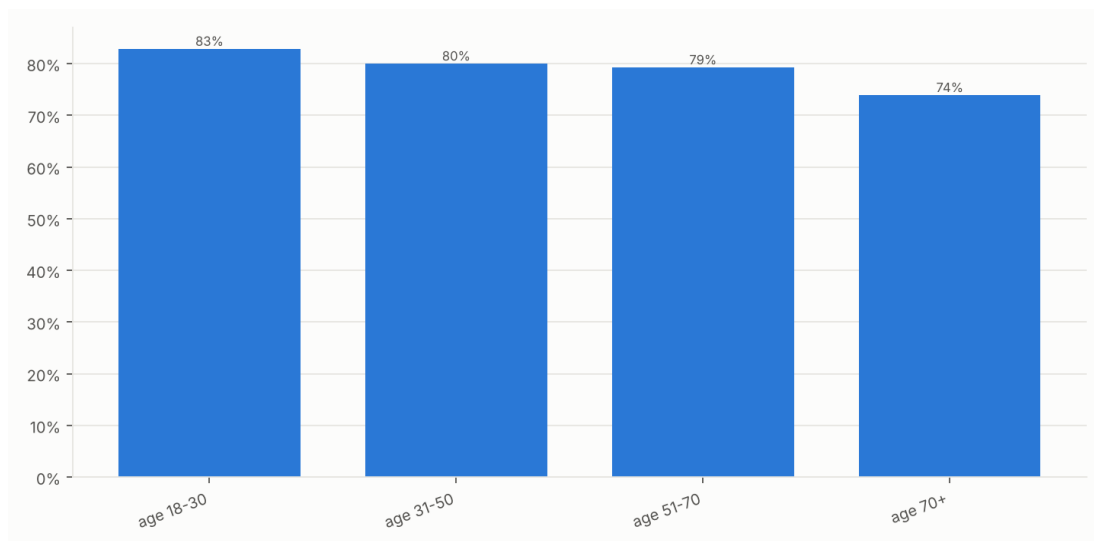


Figure 5.3: Auto-reply rate by age group (provided requests)

Routing accuracy is identical across age groups and channels, and nearly identical across languages, so the system does not *answer* any group less correctly. The differences lie in the *handling path*. Requests from citizens over 70 are auto-replied somewhat less often (0.74 against 0.83 for the 18–30 group). This reflects the category mix of the groups, not their age, since age is not an input of the engine. Citizens over 70 submit a larger share of safety-critical requests (emergency reports are 9.8% of their requests against 1.4% for the 18–30 group), and these are always dispatched urgently to an officer; 16.3% of their requests are dispatched urgently against 10.0% for the youngest group. The resulting disparate impact (0.892) satisfies the four-fifths rule. The direction of the difference also matters for interpretation: being routed to an officer is not a less favourable outcome for the citizen, since it adds human attention.

On the clean challenge set, the disparate impact between English and Italian requests is 0.750, below the 0.8 threshold. The difference arises from the gate rather than from routing: Italian challenge requests fell below the confidence threshold more often than English ones (7 against 3) and were slightly more often rated high or critical risk (5 against 4). A likely cause is that English requests pass through the English-to-Italian glossary, which inserts the exact administrative terms of the regulations and so raises evidence agreement and confidence, whereas colloquial Italian phrasing receives less expansion. The figure is based on 30 requests per language and rises to 0.844 on the noisy set, but it is reported as a monitoring alert, as the framework is designed to do. In deployment, language-level parity should be tracked on real correspondence.



5.6.3 Audit of Historical Decisions

The audit of the 500 historical cases finds no evidence that past decisions depended on the citizen's age group. The χ^2 test of independence between outcome and age group gives $\chi^2 = 2.89$ with 12 degrees of freedom ($p = 0.996$, Cramér's $V = 0.044$), and approval rates range only from 0.681 to 0.712 (disparate impact 0.956). Median response times range from 10 days (18–30) to 13 days (51–70); the Kruskal–Wallis test is not significant ($H = 5.65$, $p = 0.130$). At category level, 13 of the 15 categories show a disparate impact below 0.8. However, with about 30 cases per category and four age groups, none of these differences survives the Benjamini–Hochberg correction, so no category is flagged. The historical data can therefore be used to calibrate the risk factors and the simulation without evidence of transferring an age bias, while the small samples call for continued monitoring.

5.7 Simulation Results

The simulation model of Chapter 3 was run with the calibration of Section 3.4: AI routing accuracy 0.842 (noisy challenge set) and the auto-reply shares per category measured on the provided requests. Unless stated otherwise, results are means over six replications of 20 working days.

5.7.1 Policy Impact under Baseline Conditions

Table 5.8 – Operating policies under baseline conditions (mean $\pm$ 95% CI half-width)

Indicator	Manual office	AI routing	AI routing + drafting	AI + human oversight
SLA compliance	86.8% $\pm$ 0.9	86.7% $\pm$ 0.7	90.1% $\pm$ 0.5	96.4% $\pm$ 0.2
Average response time (working days)	1.19 $\pm$ 0.08	1.14 $\pm$ 0.05	0.86 $\pm$ 0.05	0.28 $\pm$ 0.01
90th-percentile response time (working days)	3.60	3.45	2.71	1.17
Urgent requests: average response (hours)	6.5	7.3	5.5	2.8
Task completion rate	89.2%	89.4%	92.7%	98.4%
Backlog after 20 days (requests)	303	299	206	46
Average department utilisation	0.73	0.74	0.53	0.18
Senior-desk utilisation	0.97	0.96	0.96	0.30
Misrouting rate	10.8%	14.5%	15.0%	15.0%
Cost per completed request	€79.4	€78.9	€75.7	€71.9

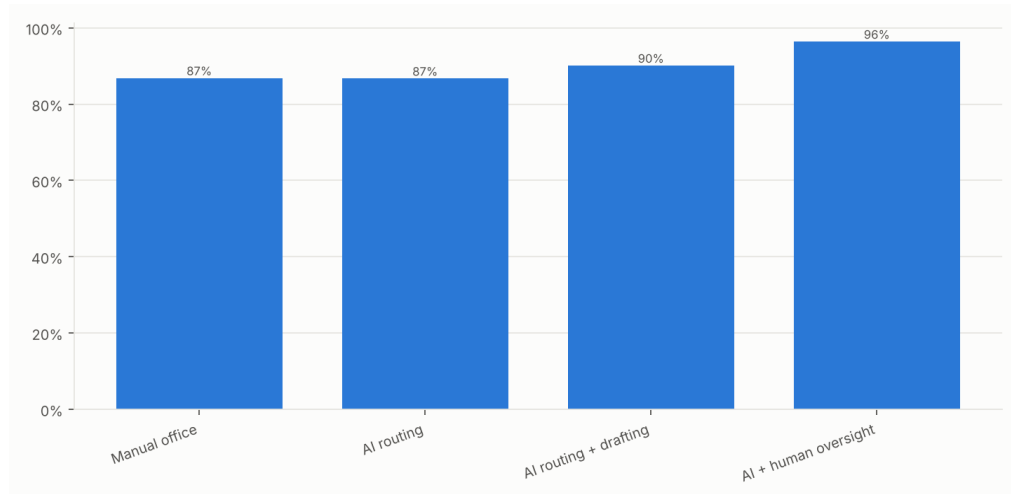


Figure 5.4: Simulated SLA compliance by operating policy (baseline scenario)

The results establish a clear ordering of the policies, and their decomposition is instructive. **AI routing alone brings no measurable benefit.** With the conservative noisy-set accuracy of 0.842, AI misrouting (14.5%) is slightly higher than the assumed manual rate (12%), and the triage time saved at the front office does not relieve the departments where queues form. **AI drafting** reduces handling effort and raises SLA compliance to 90.1%. The largest gain comes from **automatic replies under human oversight** combined with priority for urgent requests: SLA compliance rises to 96.4%, the average response time falls by 77%, and the backlog falls from 303 to 46 requests. The policy also relieves the senior review desk, which is the hidden bottleneck of the manual process at 97% utilisation, because auto-replied informational requests do not generate escalations.

Two qualifications apply. First, the cost per completed request falls by only 9.5%, because staff costs are fixed in the short term. The capacity released by AI support appears as lower utilisation (0.73 to 0.18), not as lower expenditure, and is available for reallocation (Section 5.7.4) or for work outside request handling. Second, as discussed in Section 3.4.2, the magnitude of the gain of the full policy depends on the assumed oversight effort for auto-replies. The ordering of the policies is robust; the size of the gap is not, and should be measured in a pilot.

5.7.2 Scenario Analysis

Table 5.9 – Manual office vs. AI + human oversight across scenarios

Scenario	SLA manual	SLA AI + HITL	Avg days manual	Avg days AI	Urgent hours manual	Urgent hours AI	Backlog manual	Backlog AI
Baseline	86.8%	96.4%	1.19	0.28	6.5	2.8	303	46
High e-mail volume	27.8%	96.4%	6.50	0.32	46.3	3.1	5,101	140
Natural disaster	18.2%	41.8%	6.44	2.00	58.1	39.3	5,129	3,404
Tourism season	62.2%	97.3%	2.69	0.25	16.1	2.6	1,269	56
Infrastructure incident	61.2%	78.9%	2.24	0.61	24.0	15.2	1,181	642
Staff shortage (–30%)	72.8%	96.3%	1.95	0.30	14.4	3.2	634	51
Policy change	85.2%	95.7%	1.38	0.36	7.7	3.7	366	59
Budget constraints (–15%)	81.2%	96.5%	1.44	0.28	8.1	3.0	456	44

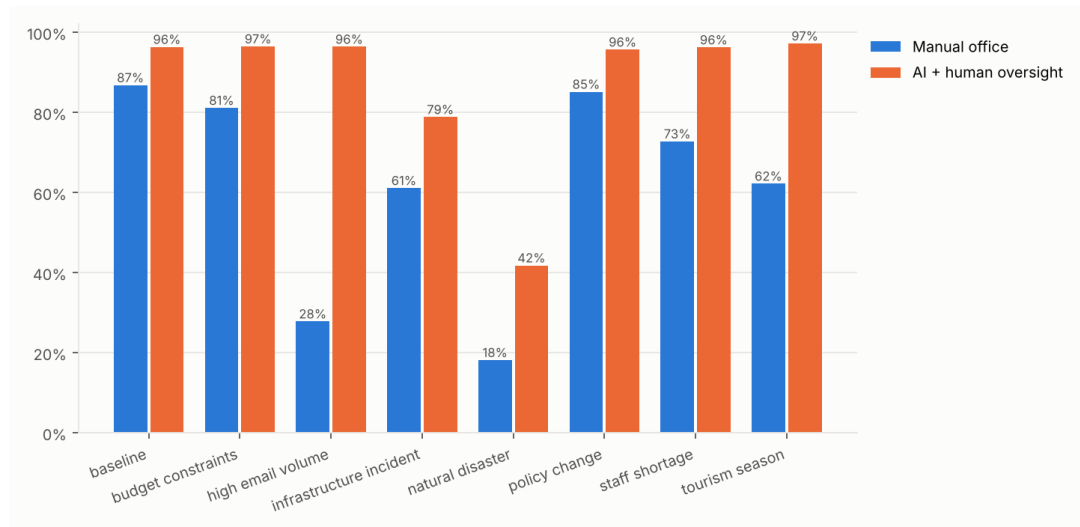


Figure 5.5: SLA compliance by scenario, manual office vs. AI with human oversight

Two groups of scenarios emerge (Table 5.9, Figure 5.5). In the **volume and capacity scenarios** (e-mail surge, tourism season, staff shortage, budget constraints, policy change), the manual office degrades sharply, down to 27.8% SLA compliance in an e-mail surge, while AI support keeps compliance at about 96%. These scenarios increase demand for informational services or reduce staff across all departments, which is exactly the load that automatic replies and drafting absorb. In the policy-change scenario, AI support remains effective even with an eight-point accuracy loss until the knowledge base is re-indexed.

In the **concentrated-shock scenarios** (natural disaster, infrastructure incident), AI support improves performance (from 18.2% to 41.8% and from 61.2% to 78.9%) but cannot restore the service. These shocks concentrate safety-critical demand on one or two departments, *Protezione Civile* and *Urbanistica*, whose requests are by design never auto-replied, and whose officer capacity is exceeded several times over. This is the correct behaviour of a trustworthy system: AI does not substitute human judgement in emergencies. It follows that resilience to these scenarios must come from contingency staffing, not from more automation.

5.7.3 What-If Analysis and Risk Forecast

Figures 5.6 and 5.7 show SLA compliance over the grid of volume and staffing multipliers for the two extreme policies.

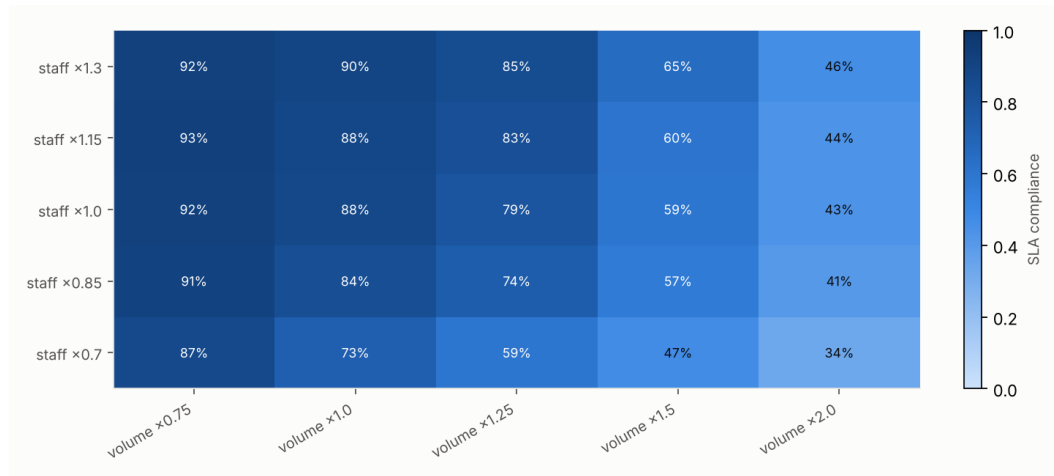


Figure 5.6: What-if surface of SLA compliance – manual office

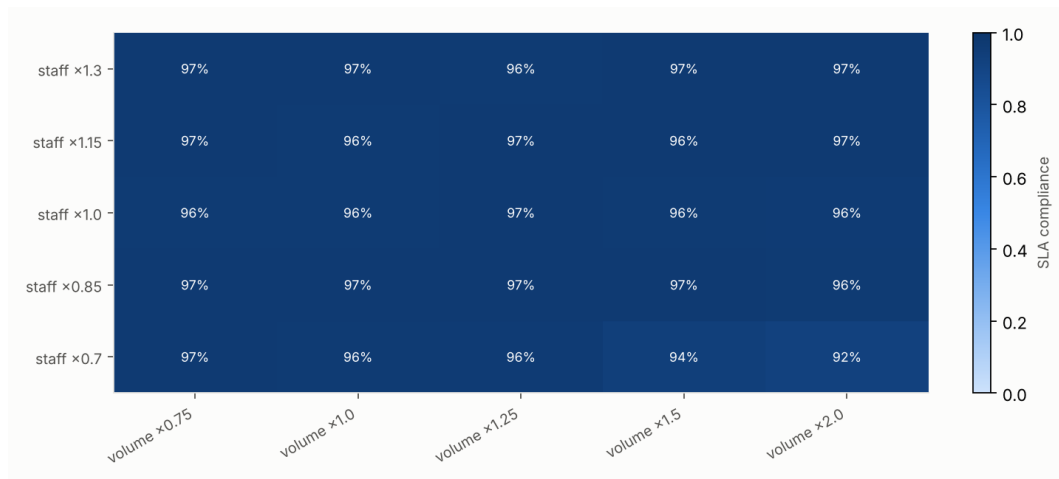


Figure 5.7: What-if surface of SLA compliance – AI with human oversight

Under the manual policy, compliance falls steeply with volume. At current staffing it drops from 92% at 0.75× volume to 43% at 2× volume, and even 30% more staff only lifts it to 46% at 2× volume. Under AI with human oversight the surface is almost flat between 92% and 97%, and it falls below 95% only when a 30% staff cut coincides with volumes of 1.5× or more. AI support thus changes the *shape* of the capacity problem: it turns a system that is sensitive to every fluctuation in demand into one that tolerates large fluctuations, and confines staff sensitivity to the departments it cannot automate.

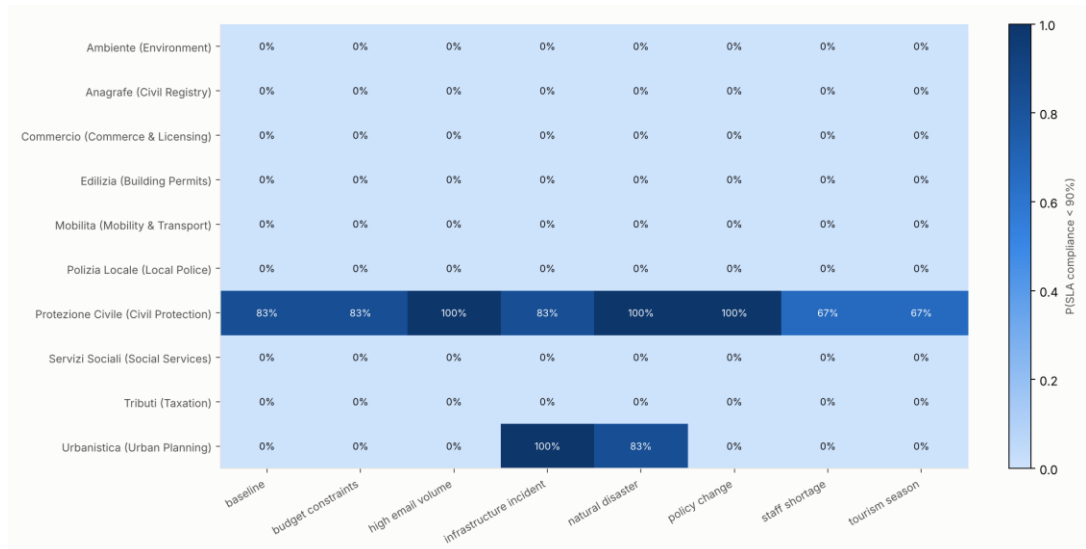


Figure 5.8: Risk forecast – probability that a department's SLA compliance falls below 90% (AI with human oversight)

The risk forecast (Figure 5.8) locates the residual risk precisely. Under AI with human oversight, eight of the ten departments have zero probability of falling below 90% SLA compliance in any scenario. *Protezione Civile* has a probability between 67% and 100% in every scenario, including the baseline, because its one-day target for emergency reports leaves no slack for queueing. *Urbanistica* reaches 83–100% in the infrastructure-incident and natural-disaster scenarios. These two departments are the strategic priority for staffing and contingency planning.

5.7.4 Staffing Optimisation

Table 5.10 reports the Erlang-C staffing recommendation for the baseline scenario under AI with human oversight, keeping the total headcount of 34 officers unchanged (target wait 4 hours, utilisation cap 0.85).

Table 5.10 – Erlang-C staffing redistribution, baseline, AI + HITL, fixed headcount

Department	Load (Erlangs)	Current officers	Recommended	Change	Utilisation now → after
Urbanistica (Urban Planning)	0.83	2	5	+3	42% → 17%
Protezione Civile (Civil Protection)	0.96	3	5	+2	32% → 19%
Anagrafe (Civil Registry)	1.27	5	6	+1	25% → 21%
Commercio (Commerce & Licensing)	0.23	2	3	+1	11% → 8%
Servizi Sociali (Social Services)	0.41	6	3	−3	7% → 14%
Edilizia (Building Permits)	0.44	4	3	−1	11% → 15%
Polizia Locale (Local Police)	0.34	4	3	−1	8% → 11%
Mobilità (Mobility & Transport)	0.13	3	2	−1	4% → 7%
Tributi (Taxation)	0.17	3	2	−1	6% → 8%
Ambiente (Environment)	0.19	2	2	0	10% → 10%

The redistribution moves request-handling capacity from departments whose demand is largely automated (social services, taxation, mobility) to the two departments whose demand is safety-critical and never automated. The utilisation figures are low because only request-



handling work is modelled; the released officers are available for casework outside the model. For the natural-disaster scenario the optimiser requires 17 officers in *Protezione Civile* (offered load 13.9 Erlangs) and 3 in *Urbanistica*. This quantifies the surge capacity that a civil-protection contingency plan should be able to mobilise, for example through cross-trained staff from the departments whose load AI support reduces.

5.7.5 Demand Forecast and Strategic Impact

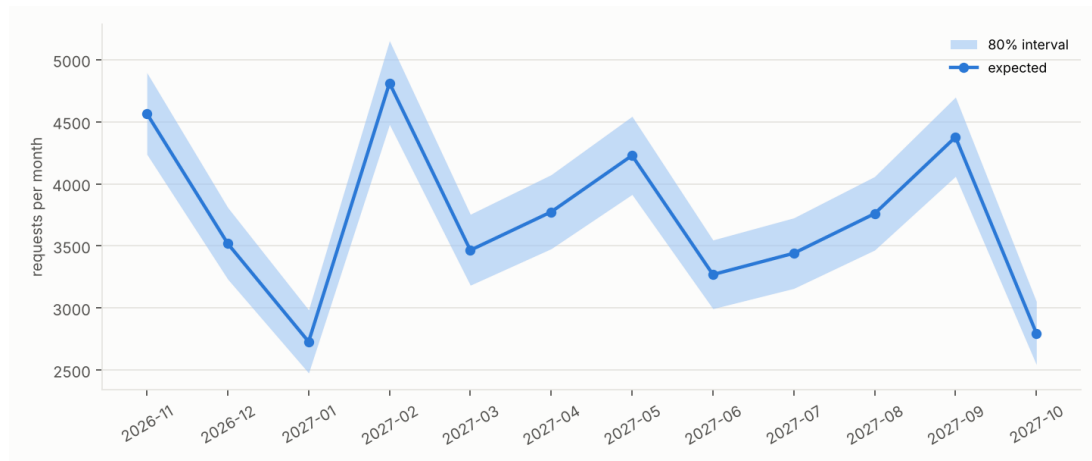


Figure 5.9: Forecast of total monthly requests, November 2026 – October 2027, with 80% prediction interval

The demand forecast (Figure 5.9) projects between about 2,700 and 4,800 requests per month over the next twelve months at the baseline volume, with peaks in February and November and troughs in January and October. Among categories, the trend estimates indicate the fastest growth for environmental reports (+38% per year), cemetery services (+21%), noise complaints (+13%), building permits (+12%) and school enrolment (+11%), and decline for social benefits (−10%) and transport passes (−8%). These trends are estimated from five years of historical counts and are indicative only.

The strategic impact score summarises the gain of AI with human oversight over the manual office as the mean of four capped relative improvements. These are the share of the SLA gap closed (0.726), the reduction in average response time (0.767), the reduction in cost per case (0.095) and the share of the completion gap closed (0.849), giving a score of **0.61** on a scale from −1 to 1. Cost is the weakest component, consistent with the observation that the benefit materialises as capacity rather than as immediate savings.

5.8 System Efficiency and Scalability

On a standard laptop the engine processes a request in 7.2 ms on average (median 7.1 ms, 95th percentile 8.2 ms) on the provided requests and in about 6 ms on the challenge sets, corresponding to a throughput of 138–172 requests per second on a single core. To assess scalability, the knowledge base was replicated up to 50 times (3,000 chunks) and retrieval latency measured over 30 queries (Figure 5.10). Mean latency grows from 4.8 ms at 60 chunks to 9.9 ms at 3,000 chunks (95th percentile 12.9 ms), and the index is rebuilt in 1.35 s. Exact search is therefore adequate for the full regulatory corpus of a large municipality, and approximate indexes would only be needed at a much larger scale. In the offline configuration the compute cost is negligible (about €0.0001 per thousand requests at €0.05 per CPU-hour). With an LLM backend, cost would be dominated by tokens, which the



context compression of Section 4.6.5 limits. The complete pipeline, including evaluation and all simulations, runs in under two minutes.

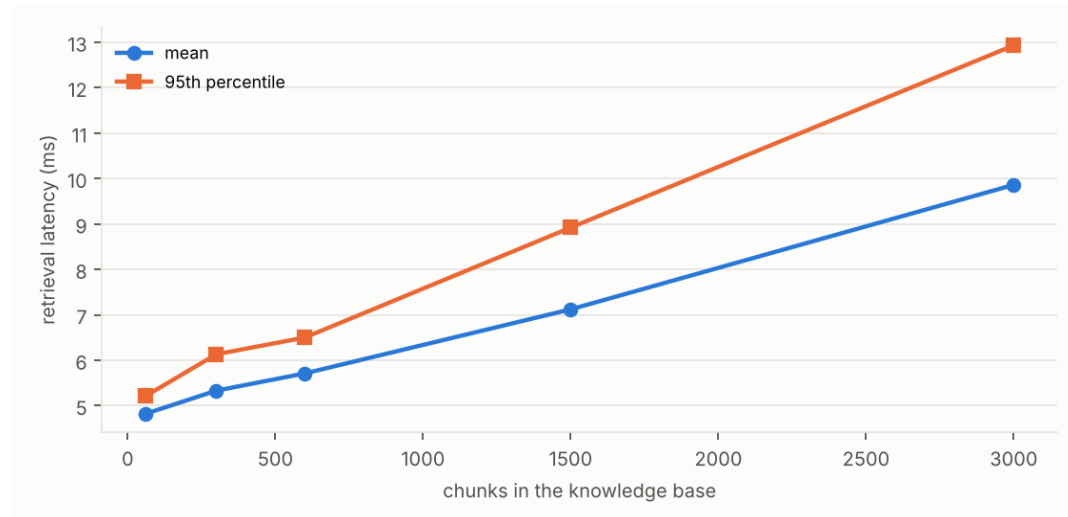


Figure 5.10: Retrieval latency as a function of knowledge-base size

5.9 Regulatory Compliance Assessment

Table 5.11 reproduces the automated compliance assessment generated at the end of the evaluation run.

Table 5.11 – Automated EU AI Act and GDPR compliance assessment

Reference	Requirement	Status	Evidence from the run
AI Act Art. 6, Annex III(5)(a)	Risk classification documented	PASS	Classified high-risk where outputs could affect access to benefits; decision support only
AI Act Art. 9	Risk scoring and alerts per request	PASS	100% of requests risk-scored
AI Act Art. 10	Data governance and bias examination	PASS	23 quality checks, 0 blocking errors; historical bias audit $p = 0.996$
AI Act Arts. 11–12	Documentation and tamper-evident logging	PASS	System card generated; audit chain valid over 390 records
AI Act Art. 13	Explanations with sources for every decision	PASS	100% of decisions explained and cited
AI Act Art. 14	Human oversight; no automated eligibility decisions	PASS	21% routed to officers; 0 automated eligibility decisions
AI Act Art. 15	Accuracy and robustness above threshold (0.90)	WARN	Routing accuracy 0.842 on the noisy set; faithfulness 1.000
AI Act Art. 50	Disclosure of AI interaction	PASS	100% of replies contain the disclosure
GDPR Arts. 5(1)(c), 25, 32	Minimisation, pseudonymisation, protected identifiers	PASS	0 direct identifiers in the working data set
GDPR Art. 22	No solely automated decision with legal effect	PASS	Eligibility outcomes always decided by an officer
AI Act Art. 10(2)(f)	Disparate impact of automated handling ≥ 0.8	PASS	Minimum disparate impact across age groups 0.892

Ten of the eleven checks pass. The Article 15 check returns a warning because routing accuracy on the noisy test set (0.842) is below the 0.90 target set for this study. The warning is mitigated, not resolved, by the decision gate, which keeps the unreviewed error rate at 3.3%. Two further points deserve attention. First, the fairness check is computed over age



groups on the provided requests; the language disparity found on the clean challenge set (Section 5.6.2) is not covered by it, and the check should be extended to language once real bilingual correspondence is available. Second, the checks verify technical properties of the system; they do not replace the organisational obligations of a deployer, such as the fundamental-rights impact assessment required of public authorities by Article 27 of the AI Act.



CHAPTER 6 – STRATEGIC DECISION-SUPPORT DASHBOARD AND DEPLOYMENT

A decision-support system is useful only when its findings are accessible to the people responsible for acting on them. This chapter describes the framework's user-facing component (methodology block 10) and the proposed route to deployment at the Comune di Genova. Section 6.1 identifies municipal users and their information needs. Section 6.2 presents the two interfaces the standalone Genova Civic Desk tool and the Streamlit dashboard and their principal views. Human review and approval are addressed in Section 6.3, while Section 6.4 describes the strategic outputs available to municipal managers. Section 6.5 covers deployment architecture, integration interfaces and the phased roll-out roadmap.

6.1 Requirements of Municipal Decision Makers

Four groups of users interact with the framework, each with different decisions to make and different time horizons (Table 6.1). Front-office operators and department officers make *operational* decisions on individual requests within minutes or hours. Heads of department and strategic managers make *tactical* and *strategic* decisions on staffing, contingency planning and adoption policy over weeks to years. The data protection officer and auditors make *assurance* decisions on whether the system operates lawfully. A single interface cannot serve all four groups well, so the dashboard is organised in views matched to these needs, and access to each view is governed by the role-based access control of Section 4.5.3.

Table 6.1 – Users, decisions and information needs

User group (role)	Typical decisions	Information needed	Main views
Citizen-service operator	Confirm routing; respond to simple requests	Proposed department, draft reply, risk level	Request desk / console
Department officer	Approve, edit or reject drafts; handle escalations	Reasons for review, explanation, cited evidence, grounding check, rule flags	Review queue; request console
Head of department, strategic manager	Staffing, contingency plans, adoption policy, priorities	KPIs, scenario results, risk forecast, staffing recommendations, action plans	Overview; scenarios; risk; staffing; strategic decisions
Data protection officer, auditor	Verify lawfulness, investigate incidents	Compliance checklist, audit log and its integrity, access records, PII vault	Governance; audit and transparency

Three design requirements follow from Chapter 1, from the users' needs and from the principles of human-centred AI (Shneiderman, 2020) and of dashboard design (Few, 2006). First, every result shown to a decision maker must be **traceable to its evidence**: a recommendation states the figures it is based on, and a reply shows the regulation articles it cites. Second, **state must be visible at a glance**: risk levels, compliance status and review priority are encoded as coloured status markers with text labels, so that what needs attention stands out without reading every row. Third, the interface must be **operable without technical expertise and without network dependencies**, so that it can be used on standard municipal workstations and in emergency conditions.



6.2 Dashboards and Views

Two complementary dashboards implement the views. The **Genova Civic Desk** is a standalone HTML and JavaScript tool that runs entirely in the browser: it can be opened as a file or served on the local network, needs no server-side Python, and embeds the knowledge base and the latest results. The **Streamlit dashboard** is connected to the live Python system: it processes requests with the full engine, runs new simulations on demand, writes review decisions to the shared queue and reads the audit log. Table 6.2 maps the eight components of the strategic decision-support dashboard foreseen by the methodology to the views that implement them.

Table 6.2 – Methodology dashboard components and their implementation

Methodology component	Genova Civic Desk (HTML)	Streamlit dashboard
AI recommendations & automated insights	Overview	AI Recommendations & Insights
KPI monitoring & analytics	Overview; Governance	KPI Monitoring & Analytics
Scenario comparison & what-if analysis	Scenarios	Scenario Comparison & What-if (with live simulation)
Risk heatmaps & alerts	Scenarios; Overview (alerts)	Risk Heatmaps & Alerts
Resource planning & optimisation	Staffing planner (live Erlang-C)	Resource Planning & Optimisation
Policy insights & compliance	Governance	Policy Insights & Compliance
Human review & approval	Review queue (per-browser)	Human Review & Approval (shared queue, audited)
Strategic decisions & action plans	Overview (recommendations)	Strategic Decisions & Action Plans; Audit & Transparency

6.2.1 Request Desk

The request desk (Figure 6.1) serves as the operational entry point. An operator can paste a citizen's message in Italian or English or choose an example; the five agents then process the request in a few milliseconds. The result panel presents five decision fields: service, competent department, risk level and score, confidence, and handling path. It also displays the proposed reply with highlighted citation markers, the reasons for the handling decision, and a grounding-check table that links each claim to its source and support score. Retrieved evidence appears with relevance bars, alongside the agent trace, per-agent timings and regulation-specific rule flags. Personal information is masked before processing, and the interface indicates when masking has occurred.



GENOVA CIVIC DESK
Strategic decision support & trustworthy AI - Comune di Genova
Run 2026-10-08 17:58
319 requests · 180 test messages

Request desk Overview Scenarios Staffing planner Governance Review queue

Request desk
Paste a citizen message in Italian or English. Five agents retrieve the governing regulation, route it, score the risk, draft a cited reply, and check every claim against the sources before deciding who handles it.

Citizen message (PII is masked before processing)

URGENTE: c'e una forte allerta meteo, l'acqua del torrente sta salendo vicino a casa mia in Val Bisagno. Cosa devo fare?

EXAMPLES
Flood alert Street festival Burial (EN) Bus pass Pothole
Renovation (EN) TARI + PII

Process request processed in 8.5 ms

AGENT TRACE

Agent	Task	Time
1 Retrieval Agent	language it, intent Emergency reports (0.67), 5 passages	6.6 ms
2 Policy & Compliance Agent	routed to Protezione Civile (score 0.83); 1 rule flags	0.1 ms
3 Risk Analysis Agent	risk critical (1.00)	0.0 ms
4 Response Generation Agent	drafted 180-word reply in it	0.2 ms
5 Confidence & Evaluation Agent		1.4 ms

SERVICE	DEPARTMENT	RISK	CONFIDENCE	HANDLING
Emergency reports	Protezione Civile - Civil Protection	critical 1.00	0.82	Urgent dispatch

Draft reply Copy reply

Gentile cittadino/a,
la sua richiesta riguarda: Emergenze. E di competenza dell'ufficio Protezione Civile (Civil Protection).
La richiesta è stata classificata come URGENTE e inoltrata con priorità.
In caso di pericolo immediato per persone chiami subito il numero unico di emergenza 112.
Secondo la normativa comunale applicabile:
- Definisce le procedure di allerta e risposta per emergenze idrogeologiche, incendi e rischio sismico. [1]
- In caso di allerta arancione o rossa, attiva il Centro Operativo Comunale (COC) e le procedure di evacuazione delle zone a rischio esondazione del torrente Bisagno e del Polcevera. [1]
Tempi indicativi:
- In base allo storico, il tempo mediano di risposta è di 13 giorni e il 90% delle pratiche è chiuso entro 20 giorni. [2]
Come procedere:
- Il cittadino può presentare istanza tramite sportello digitale, PEC o sportello fisico, allegando la documentazione richiesta. [3]
Questa risposta è stata predisposta con il supporto di un sistema di intelligenza artificiale del

Figure 6.1: Request desk of the Genova Civic Desk processing a flood-alert message (urgent dispatch to Protezione Civile)

The browser edition of the engine is a JavaScript port of the Python pipeline, using BM25 with the feature re-ranker instead of the dense index. It reproduces the routing accuracy of the Python engine on the clean challenge set (88.3%). The request desk can therefore be used offline, for training and demonstration as well as for daily triage.

6.2.2 Overview and Strategic Insights

The overview (Figure 6.2) is the entry point for managers. A row of key indicators gives the number of requests processed, the routing accuracy on the noisy test set, reply faithfulness, the simulated SLA compliance of the recommended policy with its gain over the manual office, the simulated average response time and the number of live requests at high or critical risk. Below these, charts compare the simulated SLA compliance of the four operating policies and show how the engine handled the processed requests. The overview closes with the policy recommendations, each with its priority and the evidence on which it rests, and the list of risk alerts by severity.

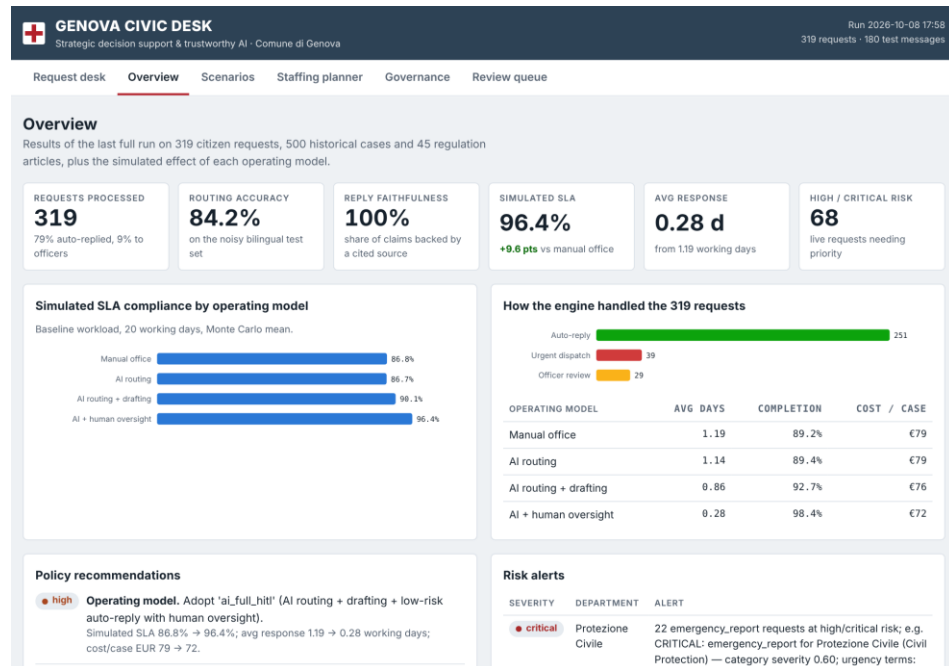


Figure 6.2: Overview with key indicators, policy comparison, recommendations and risk alerts

6.2.3 Scenario Comparison, What-If and Risk

The scenarios view (Figure 6.3) lets managers select any of the eight scenarios of Chapter 3 and compare the manual office with AI support under human oversight on four indicators: SLA compliance, average response time, response time to urgent requests and backlog after twenty days. A grouped chart shows SLA compliance for both policies across all scenarios. The view also contains the what-if grid of SLA compliance over volume and staffing multipliers for either policy, the risk heatmap of the probability that each department falls below 90% SLA compliance under each scenario, and the twelve-month demand forecast with its prediction interval. In the Streamlit dashboard, managers can also run a new simulation with their own volume and staffing multipliers and see the resulting queue lengths over time for every department.

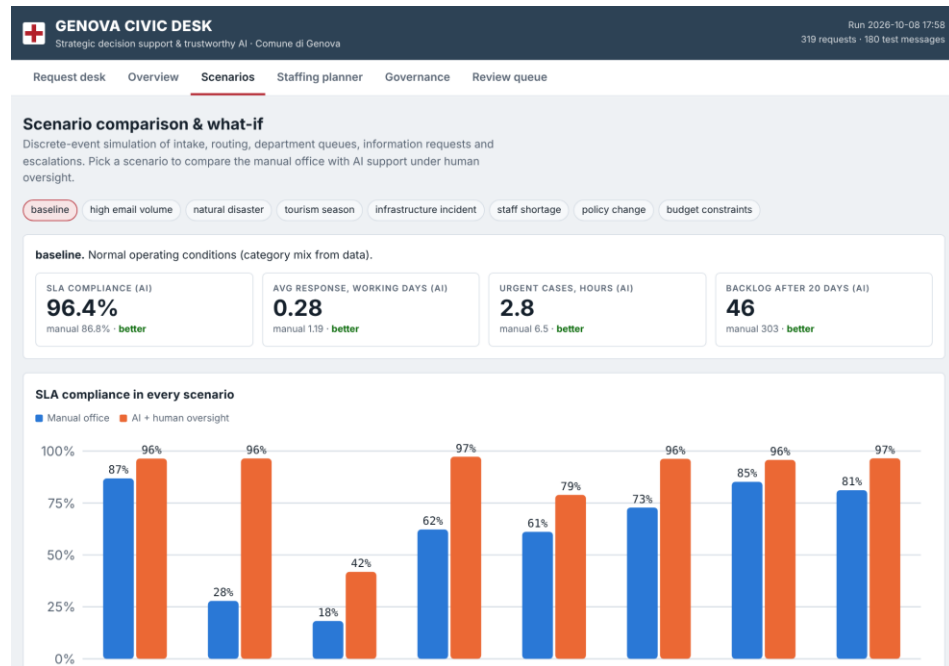


Figure 6.3: Scenario comparison view (baseline scenario selected)

6.2.4 Staffing Planner

The staffing planner (Figure 6.4) recomputes the Erlang-C optimisation of Section 3.7 interactively in the browser. The manager chooses a scenario and an operating policy, and sets the target waiting time and the utilisation cap with sliders. Optionally, the planner keeps today's total headcount and only redistributes it. The view shows current and recommended officers per department, the load before and after, and the expected waiting times, and it flags departments whose current load exceeds their capacity. In the natural-disaster scenario shown in the figure, the planner moves officers from departments whose demand is absorbed by automation to *Protezione Civile*, which would need 19 of the 34 officers to keep waiting times under four hours. This quantifies the scale of the surge capacity a civil-protection contingency plan must be able to mobilise.

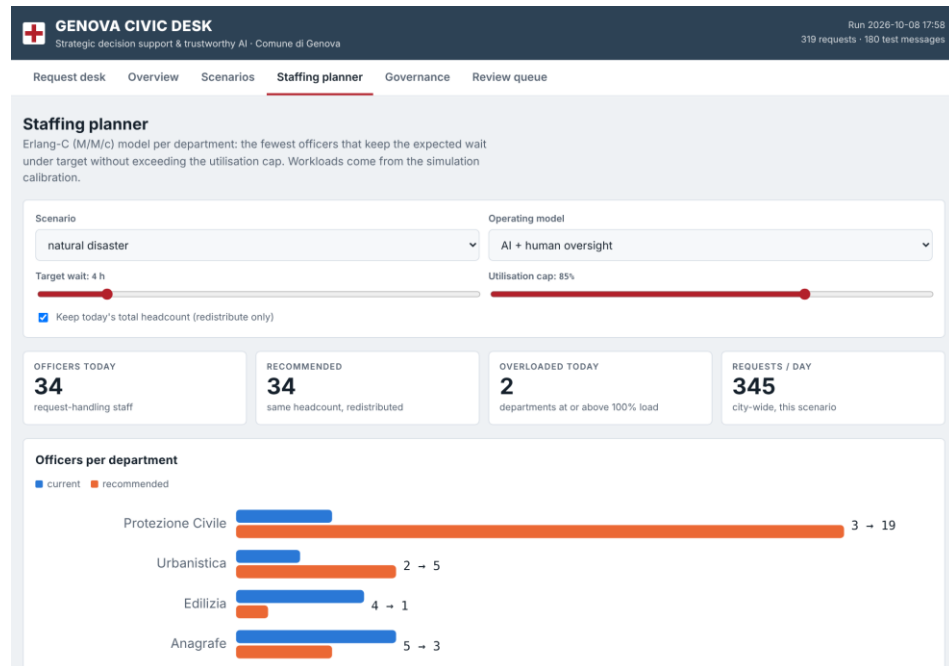


Figure 6.4: Staffing planner in the natural-disaster scenario with fixed total headcount

6.2.5 Governance, Compliance and Audit

The governance view presents the automated EU AI Act and GDPR checklist with its status markers, the evaluation results on the three test sets, the retrieval comparison, the results of the hallucination-detector validation, the fairness analysis, the calibration diagram, the routing ablation and the data-quality findings. In the Streamlit dashboard, the *Audit & Transparency* page verifies the integrity of the hash-chained audit log on demand and lists its records, filterable by event type. It also displays the system card and, only for the data protection officer, gives access to the encrypted PII vault, an access that is itself logged. The *Policy Insights & Compliance* page adds a regulation browser and the knowledge-graph neighbourhood of each service. Figure 6.5 shows the entry page of the Streamlit dashboard as seen by a strategic manager; the role selector in the sidebar demonstrates how the views available change with the user's role.

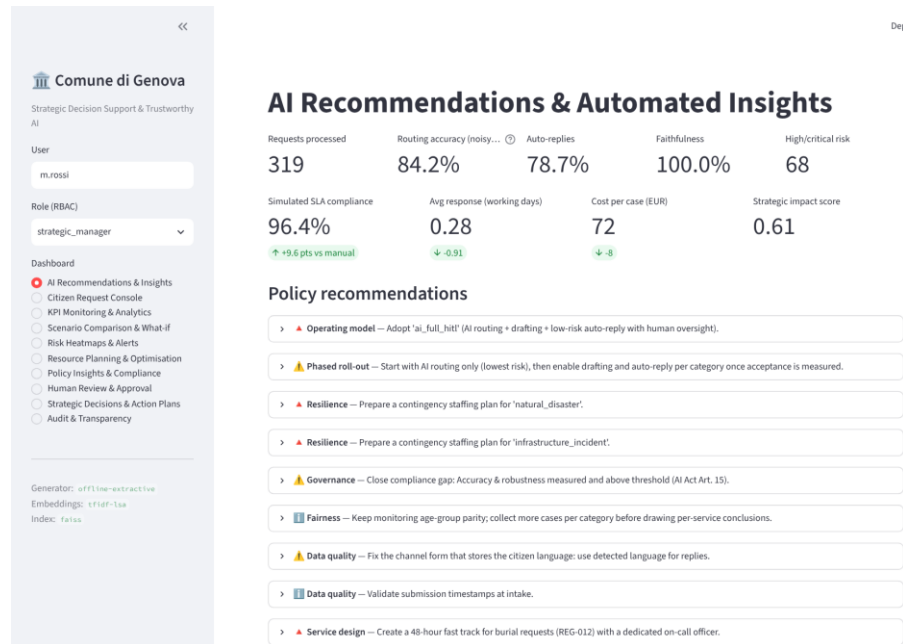


Figure 6.5: Streamlit dashboard – AI Recommendations & Insights page for the strategic-manager role

6.3 Human Review and Approval Workflow

The review workflow is where the human-oversight obligations of Article 14 of the AI Act and Article 14 of Law 132/2025 become an operational routine. Its states and transitions are shown in Figure 6.6.

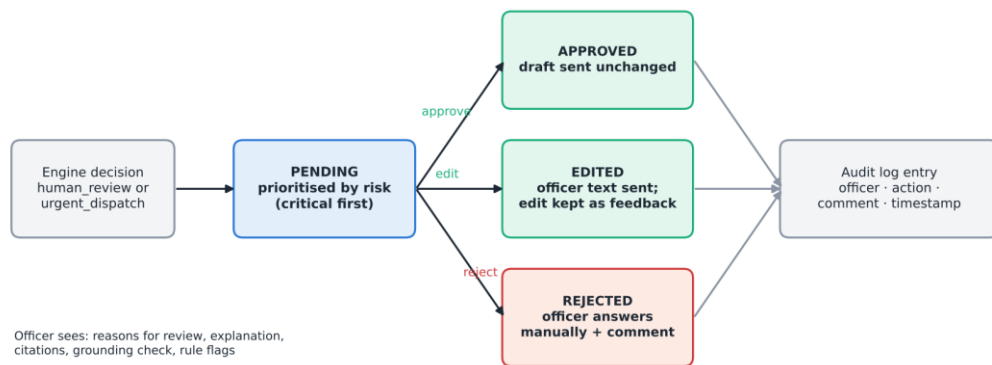


Figure 6.6: States of a request in the human review and approval workflow

Every request that the engine does not auto-reply enters the queue as *pending*, with a priority derived from its risk level: critical requests appear first, then high, medium and low. Within the same priority, older requests come first. Officers can filter the queue by department. Opening a request shows a summary of the request, risk, confidence and engine decision, followed by the reasons the engine held it back (for example, *risk level high* or *confidence* $0.65 < 0.75$) and the draft in an editable text box (Figure 6.7). The officer then takes one of three actions:



- **Approve** – the draft is sent unchanged. If the officer has modified the text before approving, the action is recorded as *edited* rather than approved, so that silent modifications cannot occur.
- **Edit** – the officer's version is sent, and the original draft and the edited version are both retained. The pairs of drafts and officer corrections form a growing data set for evaluating and improving the engine.
- **Reject** – the draft is discarded; the officer answers the citizen through the normal procedure and records the reason in a comment.

GENOVA CIVIC DESK
Strategic decision support & trustworthy AI - Comune di Genova

Run 2026-10-08 17:58
319 requests - 180 test messages

Request desk Overview Scenarios Staffing planner Governance **Review queue**

Human review & approval

Drafts the engine would not send on its own: high-risk, urgent or below the confidence gate. Approve, edit or reject; the reason the engine held each one is shown above the draft.

9 pending 0 decided

Department: All departments

- REQ-0016 Emergency reports - Protezione Civile **critical**
- REQ-0019 Public works - Urbanistica **critical**
- REQ-0020 Public works - Urbanistica **critical**
- REQ-0025 Emergency reports - Protezione Civile **critical**
- REQ-0052 Emergency reports - Protezione Civile **critical**
- REQ-0057 Public works - Urbanistica **critical**
- REQ-0006 Cemetery services - Anagrafe **high****
- REQ-0009 **high**

REQUEST	RISK	CONFIDENCE	ENGINE DECISION
REQ-0006	high	0.88	Officer review

Held because: risk level high

Draft reply (edit before approving if needed)

Gentile cittadino/a,
la sua richiesta riguarda: Servizi cimiteriali. E di competenza dell'ufficio Anagrafe (Civil Registry).
La richiesta è stata classificata come URGENTE e inoltrata con priorità.
Secondo la normativa comunale applicabile:
- Regola concessioni, tumulazioni ed estumulazioni nei cimiteri comunali. [1]
- Le concessioni loculo hanno durata di 30 o 99 anni; le pratiche urgenti di tumulazione sono evase entro 48 ore. [1]
Tempi indicativi:
- In base allo storico, il tempo mediano di risposta è di 14 giorni e il 90% delle pratiche è chiuso entro 20 giorni. [2]
Come procedere:
- Il cittadino può presentare istanza tramite sportello digitale, PEC o sportello fisico, allegando la documentazione richiesta. [3]

Questa risposta è stata predisposta con il supporto di un sistema di intelligenza artificiale del Comune di Genova.

Officer comment
e.g. checked with the duty officer

Approve Save as edited Reject

Figure 6.7: Review queue with risk-prioritised pending requests and the draft of a high-risk cemetery request

Each action records the officer's identity, the decision, the comment and the time, both in the review queue and in the hash-chained audit log. Accountability therefore remains with an identifiable official, as Law 241/1990 requires of the *responsabile del procedimento*. The acceptance rate shown at the top of the queue, the share of reviewed drafts that were approved or edited rather than rejected, is the real-world counterpart of the simulated acceptance rate of Chapter 5. It is the first indicator to monitor in a pilot. Only the officer and administrator roles may decide on drafts; an operator who opens the queue is shown an access-denied message, and the attempt is logged.

Two properties of the workflow deserve emphasis. First, review is *targeted*: officers do not review every request, only those that fall below the confidence threshold, carry high risk or contain an unsupported claim. In the evaluation this was 9% of the provided requests and 25–27% of the independent challenge requests, together with the urgent dispatches, which go directly to the duty officer. Second, review is *informed*: the officer receives the evidence, the explanation and the specific reason for the review, so that the review is a focused verification rather than a repetition of the work from scratch. This is what makes human oversight sustainable at scale, rather than a formality that would be bypassed under workload pressure.



6.4 Strategic Outputs

The strategic layer translates the per-request decisions, the simulation results and the governance findings into four outputs for municipal leaders, each stated with the figures it rests on.

6.4.1 Policy Recommendations

Table 6.3 lists the recommendations generated from the evaluation run, ordered by priority.

Table 6.3 – Policy recommendations generated by the framework

Priority	Area	Recommendation	Evidence
High	Operating model	Adopt AI routing and drafting with automatic replies under human oversight for low-risk informational requests	Simulated SLA 86.8% → 96.4%; average response 1.19 → 0.28 working days; cost per case €79 → €72
Medium	Phased roll-out	Start with AI routing, then enable drafting and automatic replies per category once officer acceptance is measured	Routing alone keeps SLA at 86.7% with no AI-written text reaching citizens
High	Resilience	Prepare contingency staffing plans for natural disasters and infrastructure incidents	Even with AI support, SLA falls to 41.8% and 78.9%, with backlogs of 3,404 and 642 requests
High	Service design	Create a 48-hour fast track for burial requests (REG-012) with an on-call officer	All historical cemetery cases exceeded the 48-hour target
Medium	Governance	Close the accuracy gap identified by the AI Act Art. 15 check before extending automation	Routing accuracy 0.842 on noisy requests against a 0.90 target
Medium	Data quality	Correct the intake form that records the citizen's language; use the detected language for replies	55 of 319 requests declare a language that does not match their text
Low	Fairness	Continue monitoring parity across age groups and languages; collect more cases per category	No significant age effect in historical decisions ($p = 0.996$); language alert on the challenge set
Low	Data quality	Validate submission timestamps at intake	70 requests dated after the reference date

6.4.2 Risk Alerts

Risk alerts are generated from three sources: live requests at high or critical risk, grouped by department and service; historical SLA performance, for services where more than a quarter of past cases exceeded the regulatory deadline; and the simulation risk forecast, for department–scenario combinations with a probability of at least 50% of falling below 90% SLA compliance. The evaluation run produced 17 alerts, 2 critical, 13 high and 2 medium. The critical alerts concern 22 emergency reports for *Protezione Civile* and 17 public-works reports with safety implications for *Urbanistica*. The high alerts include the 29 urgent cemetery requests for the *Anagrafe* and the historical breach rates of cemetery (100%), emergency (90%), public-works (77%) and environmental (52%) cases, together with the forecast risk for *Protezione Civile* in every scenario.

6.4.3 Resource Suggestions and Action Plans

Resource suggestions are taken from the Erlang-C optimisation under the recommended policy and the current headcount (Table 5.10): add officers to *Urbanistica* (+3), *Protezione Civile* (+2), *Anagrafe* (+1) and *Commercio* (+1), and release request-handling capacity in *Servizi Sociali* (–3) and in four other departments (–1 each). Each suggestion states the expected change in load and waiting time.



For each department, the framework then assembles an **action plan** that combines its staffing suggestion, its main risk alerts and its expected review workload, and assigns it to the head of department with a horizon of thirty days, if a high or critical alert is present, or of one quarter otherwise. For example, the plan for the *Anagrafe* combines the addition of one officer, the alert on urgent cemetery requests with a 100% historical breach of the 48-hour target, and the indication that 63% of its AI drafts will require review or urgent dispatch. Action plans are the bridge between analysis and management: they turn the results of the framework into assignments with an owner and a deadline.

6.5 Deployment

6.5.1 Deployment Architecture

Figure 6.8 illustrates the proposed deployment architecture. The framework operates on a municipal server or in an EU-hosted cloud environment, positioned between citizen-facing channels and the existing protocol and case-management system. Requests registered in the protocol are submitted through a REST API. The engine returns routing information, risk, a draft response, an explanation and a handling decision, which are recorded in the case file. Officers, managers and the data protection officer access the relevant dashboards according to their roles. The vector index, metadata, review queue, audit log and encrypted PII vault all remain within the municipal environment. If an external LLM backend is used, only the masked request and compressed context are transferred, subject to a data-processing agreement and processing within the European Union.

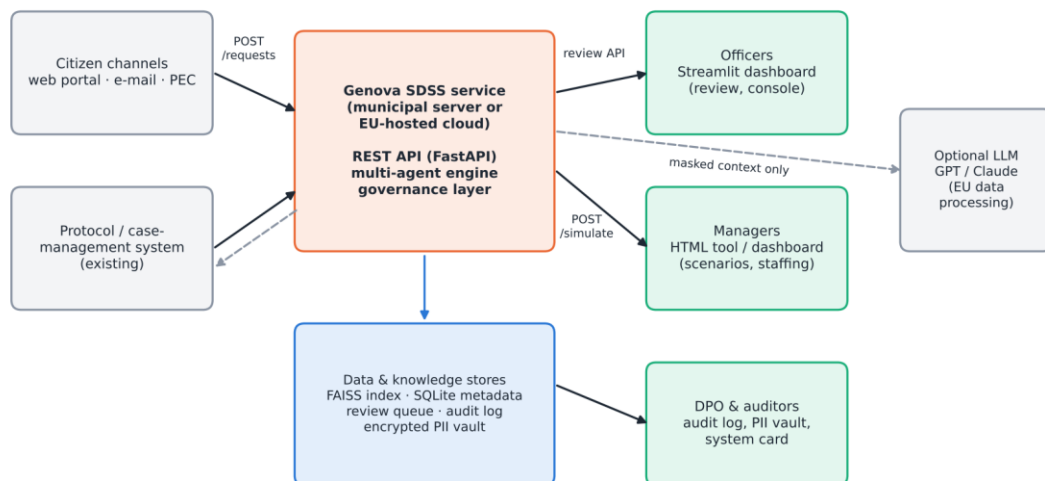


Figure 6.8: Target deployment architecture at the Comune di Genova

6.5.2 Integration Interfaces

The REST API, implemented with FastAPI, exposes the functions needed for integration (Table 6.4). Every endpoint checks the caller's role before acting. Personal data is masked before a submitted message reaches the engine, even if the calling system has not removed it.



Table 6.4 – REST API endpoints

Endpoint	Permission required	Function
GET /health	—	Service status, active generator and audit-chain integrity
POST /requests	process request	Process a citizen message: returns category, department, decision, confidence, risk, draft, faithfulness, rule flags, explanation and citations
GET /review	review decision	Pending review items, optionally filtered by department
POST /review/{id}	review decision	Record an officer's approval, edit or rejection with comment
POST /simulate	run simulation	Run a scenario and policy simulation and return the KPIs

Other interfaces support different modes of use. A command-line tool processes single messages for testing. A one-command launcher builds the knowledge base and results when needed, exports the HTML tool and starts it on a local server, and Windows, Linux and VS Code configurations are provided for it. The complete evaluation pipeline can be re-run with a single script whenever regulations or data change, regenerating every table, figure, compliance check and the system card.

Listing 6.1 – Starting the system

```
# build results if needed, serve the HTML tool and open the browser
python run.py
# re-run the full pipeline (modules 1-9) on new data
python run.py --fresh
# officer and manager dashboard connected to the live engine
streamlit run app/dashboard.py
# REST API for integration with the protocol system
uvicorn app.api:app --port 8000
```

6.5.3 Configuration of the Generator

The framework runs offline by default with the extractive grounded generator. A GPT or Claude backend is enabled by setting an environment variable for the provider and an API key, with no change to the code; replies produced by an LLM pass through the same grounding check, confidence gate and review queue. Before enabling an LLM in production, the administration should repeat the evaluation of Chapter 5 with that backend. This means measuring faithfulness and hallucination rate on the challenge sets, comparing them with the offline baseline, and adjusting the review threshold if necessary.

6.5.4 Roll-Out Roadmap

Consistent with the simulation finding that routing alone is low-risk but brings little benefit, while drafting and automatic replies bring most of the benefit, a four-phase roll-out is proposed (Table 6.5). Each phase is entered only when the exit criteria of the previous phase are met.

Table 6.5 – Proposed roll-out roadmap

Phase	Duration	Scope	Exit criteria
0 – Preparation	2 months	Fundamental-rights impact assessment (AI Act Art. 27) and DPIA (GDPR Art. 35); full regulation corpus indexed; integration with protocol system; officer training	Assessments approved; index covering all services; officers trained
1 – Shadow mode	3 months	Engine runs on real requests without acting; officers work as today and the system's proposals are compared with their decisions	Measured routing accuracy and draft acceptance on real data; workload calibration of the simulation



Phase	Duration	Scope	Exit criteria
2 – Assisted routing and drafting	6 months	AI-proposed routing confirmed by operators; cited drafts offered to officers for every request	Routing accuracy ≥ 0.90 (closing the Art. 15 warning); draft acceptance $\geq 80\%$; no fairness alert
3 – Supervised automatic replies	From month 12	Automatic replies for low-risk informational categories, enabled per category, with sampled oversight; quarterly re-evaluation	Unreviewed error rate below an agreed limit; citizen complaints monitored; annual compliance review

Shadow mode is the critical phase for the scientific validity of the framework. It provides the workload data that the simulation lacks (effort per file, actual staffing, real misrouting rates), the real-world routing accuracy and draft acceptance that replace the challenge-set estimates, and a fairness analysis on genuine bilingual correspondence. All of these are collected without any AI output reaching a citizen.

6.5.5 Operational Governance

Deployment also requires organisational measures beyond the software itself. An owner of the AI system should be designated, responsible for the system card, re-evaluation and change control. The data protection officer reviews the audit log and access records periodically. Heads of department own the action plans for their services. A change procedure ensures that every modification of regulations, ontology, thresholds or generator is followed by a re-run of the evaluation pipeline, and that the resulting compliance checklist is approved before the change goes live. The regulation corpus must be re-indexed whenever a municipal regulation is amended; the policy-change scenario of Chapter 3 quantifies the cost of failing to do so promptly. Together with the technical controls described in Chapter 4, these measures make the framework not only compliant at the time of its evaluation but governable over its lifetime.

The conclusions that follow summarise what the framework achieves, the lessons learned during its development and the developments needed to take it from a validated prototype to an operational service of the Comune di Genova.

CONCLUSIONS

Italian municipalities are expected to answer a growing volume of citizen requests within statutory deadlines, with stable or shrinking staff, while any use of artificial intelligence in that process must satisfy the EU AI Act, the GDPR, Law 132/2025 and the principles of administrative procedure set by Law 241/1990. This thesis addressed that tension for the Comune di Genova. Its aim was to show that AI can be introduced into municipal request handling as a *decision-support* capability that is effective, verifiably compliant and quantitatively justified for strategic decision makers, rather than as an opaque automation layer whose benefits and risks remain unmeasured.



TECHNICAL & SCIENTIFIC REFERENCES

- Agenzia per l'Italia Digitale. (2025, February 18). *Intelligenza artificiale: In consultazione le linee guida per la PA* [News release]. AgID.
- Alon-Barkat, S., & Busuioc, M. (2023). Human–AI interactions in public sector decision making: “Automation bias” and “selective adherence” to algorithmic advice. *Journal of Public Administration Research and Theory*, 33(1), 153–169. <https://doi.org/10.1093/jopart/muac007>
- Androutsopoulou, A., Karacapilidis, N., Loukis, E., & Charalabidis, Y. (2019). Transforming the communication between citizens and government through AI-guided chatbots. *Government Information Quarterly*, 36(2), 358–367. <https://doi.org/10.1016/j.giq.2018.10.001>
- Balci, O. (1997). Verification, validation and accreditation of simulation models. In *Proceedings of the 29th Conference on Winter Simulation* (pp. 135–141). IEEE Computer Society. <https://doi.org/10.1145/268437.268462>
- Banks, J., Carson, J. S., II, Nelson, B. L., & Nicol, D. M. (2010). *Discrete-event system simulation* (5th ed.). Pearson.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Bovens, M., & Zouridis, S. (2002). From street-level to system-level bureaucracies: How information and communication technology is transforming administrative discretion and constitutional control. *Public Administration Review*, 62(2), 174–184. <https://doi.org/10.1111/0033-3352.00168>
- Brier, G. W. (1950). Verification of forecasts expressed for probability. *Monthly Weather Review*, 78(1), 1–3. [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2)
- Bruzzone, A. G. (2018, September). *MS2G as pillar for developing strategic engineering as a new discipline for complex problem solving* [Keynote presentation]. I3M Multiconference, Budapest, Hungary.
- Bruzzone, A. G., Agresta, M., & Sinelshchikov, K. (2017a). Simulation as decision support system for disaster prevention. In *Proceedings of the International Workshop on Simulation for Energy, Sustainable Development & Environment (SESDE 2017)*. Barcelona, Spain.
- Bruzzone, A. G., Massei, M., Agresta, M., Poggi, S., Camponeschi, F., & Camponeschi, M. (2014). Addressing strategic challenges on mega cities through MS2G. In *Proceedings of the International Workshop on Modeling & Applied Simulation (MAS 2014)*. Bordeaux, France.



- Bruzzzone, A. G., Massei, M., Longo, F., Maglione, G. L., Di Matteo, R., Di Bella, P., & Milano, V. (2017b). Verification and validation applied to an interoperable simulation for strategic decision making involving human factors. In *Proceedings of the International Workshop on Applied Modeling & Simulation (WAMS 2017)*. Florence, Italy.
- Consiglio di Stato, Sezione VI. (2019a). *Sentenza 8 aprile 2019, n. 2270*.
- Consiglio di Stato, Sezione VI. (2019b). *Sentenza 13 dicembre 2019, n. 8472*.
- Cormack, G. V., Clarke, C. L. A., & Büttcher, S. (2009). Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 758–759). Association for Computing Machinery. <https://doi.org/10.1145/1571941.1572114>
- Decreto Legislativo 7 marzo 2005, n. 82. (2005). *Codice dell'amministrazione digitale*. Gazzetta Ufficiale della Repubblica Italiana, 16 May 2005.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391–407. [https://doi.org/10.1002/\(SICI\)1097-4571\(199009\)41:6<391::AID-ASI1>3.0.CO;2-9](https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9)
- Dunleavy, P., Margetts, H., Bastow, S., & Tinkler, J. (2006). *Digital era governance: IT corporations, the state, and e-government*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199296194.001.0001>
- Dyer, M. E., & Proll, L. G. (1977). Note On the validity of marginal analysis for allocating servers in M/M/c queues. *Management Science*, 23(9), 1019–1022. <https://doi.org/10.1287/mnsc.23.9.1019>
- Erlang, A. K. (1917). Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Elektroteknikerer*, 13, 5–13.
- Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 259–268). Association for Computing Machinery. <https://doi.org/10.1145/2783258.2783311>
- Few, S. (2006). *Information dashboard design: The effective visual communication of data*. O'Reilly Media.
- Gans, N., Koole, G., & Mandelbaum, A. (2003). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service*



- Operations Management*, 5(2), 79–141.
<https://doi.org/10.1287/msom.5.2.79.16071>
- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 1321–1330). PMLR.
<https://doi.org/10.48550/arXiv.1706.04599>
- Harding, J. (2018, August 21). CCTV shows Genoa bridge collapse. *BBC News*.
- Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems* (Vol. 29, pp. 3315–3323). Curran Associates.
<https://doi.org/10.48550/arXiv.1610.02413>
- High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI*. Publications Office of the European Union.
<https://doi.org/10.2759/346720>
- Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4), 422–446. <https://doi.org/10.1145/582415.582418>
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1–38.
<https://doi.org/10.1145/3571730>
- Johnson, J., Douze, M., & Jégou, H. (2021). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3), 535–547.
<https://doi.org/10.1109/TBDATA.2019.2921572>
- Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 6769–6781). Association for Computational Linguistics.
<https://doi.org/10.18653/v1/2020.emnlp-main.550>
- Kelly, A. (2021). A tale of two algorithms: The appeal and repeal of calculated grades systems in England and Ireland in 2020. *British Educational Research Journal*, 47(3), 725–741. <https://doi.org/10.1002/berj.3705>
- Kleinrock, L. (1975). *Queueing systems: Vol. 1. Theory*. Wiley.
- Law, A. M. (2015). *Simulation modeling and analysis* (5th ed.). McGraw-Hill Education.
- Legge 23 settembre 2025, n. 132. (2025). *Disposizioni e deleghe al Governo in materia di intelligenza artificiale*. Gazzetta Ufficiale della Repubblica Italiana, 25 September 2025.



- Legge 7 agosto 1990, n. 241. (1990). *Nuove norme in materia di procedimento amministrativo e di diritto di accesso ai documenti amministrativi*. Gazzetta Ufficiale della Repubblica Italiana, 18 August 1990.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 9459–9474). Curran Associates. <https://doi.org/10.48550/arXiv.2005.11401>
- Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out* (pp. 74–81). Association for Computational Linguistics. <https://aclanthology.org/W04-1013/>
- McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Chapman & Hall. <https://doi.org/10.1007/978-1-4899-3242-6>
- Mergel, I., Edelmann, N., & Haug, N. (2019). Defining digital transformation: Results from expert interviews. *Government Information Quarterly*, 36(4), Article 101385. <https://doi.org/10.1016/j.giq.2019.06.002>
- Nii, H. P. (1986). Blackboard systems: The blackboard model of problem solving and the evolution of blackboard architectures. *AI Magazine*, 7(2), 38–53. <https://doi.org/10.1609/aimag.v7i2.537>
- Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Association for Computational Linguistics. <https://doi.org/10.3115/1073083.1073135>
- Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons regarding the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). (2016). *Official Journal of the European Union*, L 119, 1–88. <http://data.europa.eu/eli/reg/2016/679/oj>
- Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). (2024). *Official Journal of the European Union*, L, 2024/1689. <http://data.europa.eu/eli/reg/2024/1689/oj>
- Regulation (EU) 2026/1744 of the European Parliament and of the Council of 8 July 2026 amending Regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI). (2026). *Official Journal of the European Union*, L, 2026/1744. <http://data.europa.eu/eli/reg/2026/1744/oj>



- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982–3992). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1410>
- Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4), 333–389. <https://doi.org/10.1561/15000000019>
- Robinson, S. (2014). *Simulation: The practice of model development and use* (2nd ed.). Palgrave Macmillan.
- Sargent, R. G. (2013). Verification and validation of simulation models. *Journal of Simulation*, 7(1), 12–24. <https://doi.org/10.1057/jos.2012.20>
- Schneier, B., & Kelsey, J. (1999). Secure audit logs to support computer forensics. *ACM Transactions on Information and System Security*, 2(2), 159–176. <https://doi.org/10.1145/317087.317089>
- Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6), 495–504. <https://doi.org/10.1080/10447318.2020.1741118>
- Team SimPy. (2024). *SimPy: Discrete event simulation for Python* [Computer software]. <https://simpy.readthedocs.io>
- Van Bekkum, M., & Zuiderveen Borgesius, F. (2021). Digital welfare fraud detection and the Dutch SyRI judgment. *European Journal of Social Security*, 23(4), 323–340. <https://doi.org/10.1177/13882627211031257>
- Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law*, 7(2), 76–99. <https://doi.org/10.1093/idpl/ix005>
- Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2019). Artificial intelligence and the public sector Applications and challenges. *International Journal of Public Administration*, 42(7), 596–615. <https://doi.org/10.1080/01900692.2018.1498103>
- Wooldridge, M. (2009). *An introduction to multiagent systems* (2nd ed.). Wiley.
- Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR 2023)*. <https://doi.org/10.48550/arXiv.2210.03629>
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. In *Proceedings of*



the Eighth International Conference on Learning Representations
(ICLR 2020). <https://doi.org/10.48550/arXiv.1904.09675>