



INTERNATIONAL
HELLENIC
UNIVERSITY

Dissertation Title Goes Here

Student Name

SID: 12345678

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in Information and Communication Systems

OCTOBER 2012

THESSALONIKI – GREECE



INTERNATIONAL
HELLENIC
UNIVERSITY

(title page – delete this line)

Dissertation Title Goes Here

Student Name

SID: 12345678

Supervisor:

Prof. Name Surname

Supervising Committee Members: Assoc. Prof. Name Surname

Assist. Prof. Name Surname

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in Information and Communication Systems

OCTOBER 2012

THESSALONIKI – GREECE

Abstract

This dissertation was written as a part of the MSc in ICT Systems at the International Hellenic University. Here goes a summary of the dissertation (1-2 paragraphs). Add one last paragraph acknowledging the supervisor and the people who contributed (collaborators, other scientists, etc.). Abstract length should not exceed one page.

Summarize the problem of manual eligibility verification, the proposed Neuro-Symbolic pipeline solution, the experimental comparisons and the critical findings.

Student Name Date

Contents

Abstract	iii
Contents	v
1 Introduction	1
1.1 Background and Motivation	1
1.2 Problem Statement	1
1.3 Objectives	2
1.4 Dissertation Structure	2
2 Systematic Literature Review	3
2.1 Introduction	3
2.2 Methodology	3
2.2.1 Research Questions	3
2.2.2 Search Strategy	4
2.2.3 Inclusion/Exclusion Criteria	4
2.3 Results	5
2.3.1 PRISMA Flow Diagram	6
2.3.2 Data Extraction	6
2.4 Thematic Analysis	8
2.4.1 Neuro-Symbolic Pipelines in Public Administration	8
2.4.2 State-of-the-art in Logic Synthesis	10
2.4.3 Methodologies for Logic Validation and Trust	12
2.5 Discussion and Research Gap	14
2.5.1 The Prescriptive Synthesis Gap: From Facts to Rules	14
2.5.2 The Stability Paradox: Correctness vs. Resilience	14
2.5.3 The Validation Vacuum: The Need for Mutation Testing	15
2.5.4 Contribution of this Study	15
3 Pilot Study	17
3.1 Overview	17
3.2 Methodology and System Architecture	17
3.2.1 Setup Environment	17
3.2.2 Semantic Data Modeling	18

3.2.3	The Extraction and Generation Pipeline	19
3.2.4	The Validation Engine	24
3.3	Experimental Design	25
3.3.1	The Mutation Testing Framework	26
3.3.2	Experimental Configurations	27
3.3.3	Evaluation Metrics	30
3.4	Conclusion	31
4	Results	33
4.1	Experimental Dataset	33
4.2	Syntactic Validity	34
4.2.1	Success Rate	34
4.2.2	Failure Mode Analysis	35
4.3	Logic Validity	36
4.3.1	Success Rate	36
4.3.2	The Syntax-Logic Gap	37
4.4	Overall Pipeline Reliability	37
4.4.1	Pipeline Feasibility and Attrition	37
4.4.2	In-context (Recommender System) reliability	38
5	Discussion	41
5.1	Cognitive Dissonance in Code Generation	41
5.1.1	The Illusion of Fluency	41
5.1.2	Structural Logic Failures	41
5.2	Syntax Hallucination and Language Bleed	42
5.2.1	SQL Contamination	42
5.2.2	Token-Level Hallucinations	43
5.2.3	Namespace Invention	43
5.3	The Trade-off of Abstraction vs. Fidelity	43
5.3.1	The "Smart Model" Trap	43
5.3.2	The Deterministic Superiority	44
5.4	The Complexity Ceiling	44
5.5	The Contribution of Prompt Engineering	45
5.5.1	The "Copy-Paste" Bias	45
5.5.2	The Failure of Self-Correction	45
5.5.3	The Engineering Ceiling	45
5.6	Feasibility and Sovereignty	45
5.6.1	The Semantic Drift of "Eligibility"	46
5.6.2	Replication Crisis	46
5.7	Risk Asymmetry in Public Administration	46

6	Limitations & Future Work	49
6.1	Infrastructure Dependencies & Digital Sovereignty	49
6.1.1	Model and Resource Constraints	49
6.1.2	Operational Fragility and the Replication Crisis	49
6.1.3	Model Specialization	50
6.1.4	A Roadmap Toward Sovereign AI	50
6.2	Scalability	50
6.2.1	Document Corpus and Prompting Strategies	50
6.2.2	From Pilot to Interoperability	51
6.3	Methodological Refinement	51
6.3.1	Root Cause Analysis	51
6.3.2	Artifact Evaluation	52
6.4	Trust-Centric Optimization	52
	Bibliography	55
A	Appendix placeholder	61

List of Figures

2.1	PRISMA Flow Diagram of the selection process	6
3.1	Flow Chart of the core pipeline	20
3.2	A citizen-service graph according to the Meta-Model	22
4.1	Syntactic Validity Rates by Configuration	34
4.2	Distribution of Syntax Error Types by Model and Document	35
4.3	Logic Validity: Percentage of Flawless Runs (All Scenarios Correct) . . .	36
4.4	Distribution of Final Pipeline Outcomes.	38
4.5	Confusion Matrix of Eligibility Recommendations	39

List of Tables

2.1	Inclusion and Exclusion Criteria	5
2.2	Final Synthesis Matrix of Included Studies ($n = 25$)	7
4.1	Distribution of Experimental Runs per Configuration	33
4.2	Definition of Classification Outcomes in the Recommender Context . . .	39

Listings

3.1	JSON Information Model snippet	21
3.2	A SHACL Shape generated by the pipeline	23
3.3	YAML description of a scenario	27
3.4	Excerpt of the reflexion prompt	29

1 Introduction

An idea I wrote, not sure if it fits here: Consider this analogy: Building a neuro-symbolic pipeline for public administration is akin to building a high-speed railway between two distinct cultures. The LLM acts as the versatile translator and engineer who interprets the local landscape (unstructured text), while the intermediate YAML or JSON models serve as the blueprints and tracks. The Knowledge Graph is the rigid, reliable station where everyone must follow specific rules (logic). Currently, we have built successful short-distance test tracks (prototypes), but the challenge remains in connecting the entire national network while ensuring the trains (data) always run on time and stay on the correct side of the tracks (normative compliance).

Another nice Analogy: If Descriptive QA is like asking a librarian to find a specific book (retrieval), then Prescriptive synthesis is like asking a lawyer to write a set of building codes (rule-making). In the descriptive case, the model translates a search request into a precise map of the library. In the prescriptive case, the model must understand the underlying intent of a policy and translate it into a rigid framework of "if-then" laws that govern future actions.

And another Analogy: Validating a neuro-symbolic system is like inspecting a bridge built by a highly creative but occasionally distracted architect. Probabilistic methods are like counting how many cars successfully cross the bridge (F1 scores), while deterministic methods are like a safety inspector using a physical measuring tape to ensure the steel beams meet code (SHACL/OBQC). Currently, we are very good at checking if the bridge stands up once (static validation), but we lack a way to simulate an earthquake or a flood (mutation testing) to ensure the architect's logic doesn't fail under pressure.

1.1 Background and Motivation

The burden of manual bureaucracy in public administration and the potential of AI to automate legislative interpretation.

Why is this work important?

1.2 Problem Statement

the challenge of bridging unstructured text with deterministic validation logic while mitigating LLM hallucinations.

1.3 Objectives

the specific goals of building a Text-to-SHACL pipeline and evaluating its semantic accuracy and operational feasibility.

1.4 Dissertation Structure

Outline of the organization of the subsequent chapters and the logical flow of the research.

2 Systematic Literature Review

This chapter details a Systematic Literature Review (SLR) with the goal to establish the necessary theoretical foundations of Neuro-Symbolic AI.

2.1 Introduction

We approach the current research in Neuro-Symbolic AI specifically focusing on how Large Language Models (LLMs) and Knowledge Graphs (KGs) are combined. We aim to find existing approaches for extracting rules from text and generating formal logic (with a particular focus on SPARQL and SHACL), as well as methods of evaluating the results of such a process.

2.2 Methodology

To ensure scientific strictness and reproducibility, the review adheres to the PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) guidelines. The process was structured into four phases. First, we defined specific Research Questions (RQs). Second, we executed an automated search strategy on the *Scopus* database. Third, we applied a two-stage screening process: an initial "practical" screening of only titles and abstracts, and then a more thorough "quality assessment" screening of the full texts. This phase utilized specific inclusion/exclusion criteria and quality assessment criteria. Finally, data was extracted from the selected primary studies into a standardized matrix to synthesize key themes and composed into a thematic analysis.

2.2.1 Research Questions

To achieve our objective, we defined three Research Questions that guided the data extraction and synthesis process:

- **RQ1:** How are Large Language Models (LLMs) currently utilized to extract structured knowledge and conditional rules from unstructured text?
- **RQ2:** What are the state-of-the-art approaches for translating natural language requirements into executable constraint languages (specifically SHACL and SPARQL)?
- **RQ3:** What methodologies exist for evaluating the functional correctness and operational stability of LLM-generated logic?

RQ1 explores the initial phase of the proposed pipeline (Text-to-Graph), while **RQ2** focuses

on the core challenge of logic generation. **RQ3** allows us to examine how existing studies critically evaluate trust and correctness.

2.2.2 Search Strategy

To identify relevant records, we conducted an automated search on the *Scopus* database. Scopus was selected as the source due to its extensive coverage of academic literature. The search was executed on the 1st of December 2025. The search query was formulated to find the intersection of Generative AI and Semantic Web technologies. We employed Boolean logic to combine three conceptual blocks:

1. **Generative AI Terms:** ("Large Language Model" OR "LLM")
2. **Target Logic:** ("SHACL" OR "SPARQL")
3. **Symbolic Terms:** ("Semantic Web" OR "Knowledge Graph")

These blocks were combined using the AND operator. Thus, the final search string applied to the Title, Abstract, and Keywords fields was:

```
( "Large Language Model" OR "LLM" ) AND  
( "SHACL" OR "SPARQL" ) AND  
( "Semantic Web" OR "Knowledge Graph" )
```

We also applied some necessary metadata filters during this phase:

- **Language:** Only papers written in English were considered.
- **Document Type:** We restricted the search to Articles and Conference Papers, excluding trade journals and errata.

Interestingly, despite the Date Range not being restricted, all results fell in the range of years 2023–2026. This could be an indication that the application of Large Language Models to formal constraint languages like SHACL is a newly emerging field, that appeared primarily after the widespread adoption of "GPT-4 class" models.

The described search strategy yielded an initial set of 125 candidates which, after removing 3 duplicates, were then subjected to the screening process described next.

2.2.3 Inclusion/Exclusion Criteria

We established a set of inclusion and exclusion criteria that reflect the focus of this review. These were applied to Titles and Abstracts during the initial "Practical Screening" phase of the 122 records. Table 2.1 summarizes the criteria used. This screening process excluded 61 papers from the review. The rest were sought for retrieval from official channels. Of the 61 papers sought, 8 could not be retrieved due to access restrictions (paywall). The remaining 53 were downloaded and assessed for eligibility by reading the full text. In this "Quality Screening" phase, we applied a second set of quality exclusion criteria (QE), with the goal to further focus our scope and increase relevance to this study.

Table 2.1: Inclusion and Exclusion Criteria

Category	Inclusion Criteria	Exclusion Criteria
Task Focus	Text-to-Graph extraction, Text-to-SPARQL, SHACL shapes generation, GraphRAG architectures.	Pure NLP (summarization), low-level graph mechanics (Entity Alignment, Link Prediction, Subgraph Extraction), Dataset creation.
Methodology	Neuro-Symbolic architectures, Prompt Engineering for logic generation, Fine-tuning, Evaluation Frameworks for Semantic Accuracy.	Traditional Machine Learning (non-generative), Reinforcement Learning without LLMs.
Data Flow	<i>Forward:</i> Transforming unstructured text into formal logic or structured data (Text \rightarrow Logic).	<i>Reverse:</i> Transforming structured data into natural language (Verbalization/Explanation).
Mode	Textual inputs with or without pre-processing.	Multimodal studies (Speech/Image), Computer Vision, Temporal Data.
Type	Peer-reviewed Articles and Conference Papers.	Conference Proceedings (Meta-entries), Posters, Editorials, Preliminary Results.

- **QE1 (Domain & Logic Mismatch):** From articles situated in descriptive scientific domains (e.g., bioinformatics, chemistry), exclude those where the knowledge structure is purely factual or relational rather than normative or rule-based, offering low transferability to eligibility logic.
- **QE2 (Complexity & Task Focus):** From studies focusing on simple factoid Question Answering (KGQA), exclude those that do not analyze the extraction or generation of complex conditional constraints (if-then-else logic) required for compliance or eligibility.
- **QE3 (Methodological Maturity):** From studies limited to model-vs-model benchmarking or evaluation of existing datasets, exclude those that do not propose novel neuro-symbolic pipeline architectures or logic-validation frameworks.

Following this quality assessment, 28 papers were excluded from the review (13 due to QE1, 6 due to QE2 and 9 due to QE3), leaving 25 papers to be included.

2.3 Results

From an initial set of 125 records, 25 studies were identified as meeting all eligibility criteria.

2.3.1 PRISMA Flow Diagram

The search and screening process can be summarized in the PRISMA flow diagram (Figure 2.1).

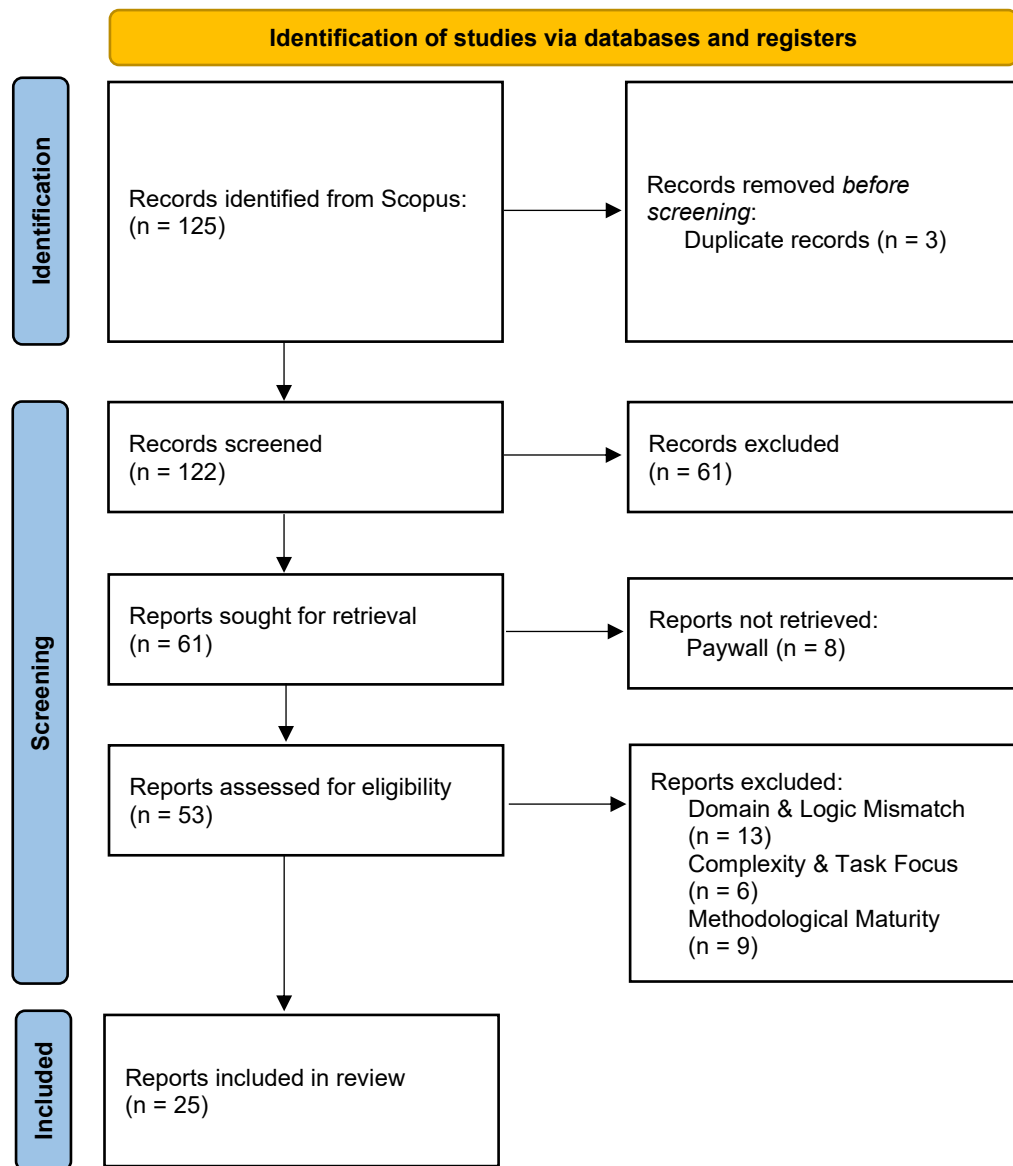


Figure 2.1: PRISMA Flow Diagram of the selection process

2.3.2 Data Extraction

Table 2.2 presents the data extraction summary for the 25 included studies. The studies are categorized thematically, to reflect the research trajectory: from domain-specific applications in public administration and normative compliance, through the intricate technical mechanisms of logic synthesis, to the frameworks of validation and architectural trust.

Table 2.2: Final Synthesis Matrix of Included Studies ($n = 25$)

Study	Core Task / Do-main	Logic	Neuro-Symbolic Inte-gration	Validation Method
<i>Category 1: Public Administration & Normative Compliance</i>				
Konstantinidis (2025) [1]	Recommendation (Public Services)	RDF, SHACL	LLM extraction, SHACL validation	Human Expert As-sessment
Oranekwu (2026) [2]	Cybersecurity Compliance (IoT)	OWL, SPARQL	Ontology-driven RAG	Similarity over ground truth
Spyropoulos (2025) [3]	Entity Mining (Po-lice Reports)	RDF, OWL	LLM Entity Extraction & linking	Visual and SPARQL Verification
Hanuragav (2025) [4]	CSR Validation (Medical)	SHACL, SPARQL	RTF to JSON to RDF with YAML mapper	SHACL (structure), SPARQL (content)
<i>Category 2: Automated Logic Synthesis & Semantic Parsing</i>				
Agarwal (2024) [5]	Complex QA (KGQA)	KoPL	SymKGQA: Symbolic Program Generation	Hits@1 and F1 on Benchmarks
Avila (2025) [6]	Scientific QA (KGQA)	SPARQL	Text-to-SPARQL with RAG + Few-shot ICL	F1 on Benchmarks
Jiang (2025) [7]	Multi-KG Gener-alization (KGQA)	SPARQL	Semantic sketch + KG population	Hits@1 and F1 on Benchmarks
Shah (2024) [8]	Multi-hop QA (KGQA)	Cypher, SPARQL	Text-to-Logic with Few-shot + CoT	Match Accuracy on Benchmarks
Walter (2026) [9]	Reasoning / QA (Multi-domain)	SPARQL	Zero-shot Iterative Agent KG exploration	F1 on Benchmarks
Soularidis (2024) [10]	Rule Generation (Search & Rescue)	SWRL	Ontology-driven Text-to-SWRL	F1 on Human Expert
Lehmann (2023) [11]	Semantic Parsing (Wikidata)	CNL, SPARQL	Controlled Natural Language to Logic	Hits@1 on Bench-marks
Kovriguina (2023) [12]	SPARQL Genera-tion (Fantasy)	SPARQL	Augmenting prompts with RDF subgraphs	F1-macro on Bench-marks
Mountantonakis [13] (2025)	Cultural Heritage (Art)	SPARQL	Path Pattern prediction + query generation	Accuracy on Bench-mark
Ongris (2024) [14]	Wikidata QA (KGQA)	SPARQL	Sequential LLM Chain-ing + GraphRAG	Jaccard Similarity on Ground Truth
Vieira da Silva (2024) [15]	Capability Model-ing (IoT)	OWL	TBox-contextualized prompting	Pellet (OWL reason-ing) + SHACL
Emonet (2025) [16]	Federated QA (Bioinformatics)	SPARQL	ShEx/VoID-driven RAG query generation	Execution Success Rate and F1
Mashhaditafreshi (2025) [17]	Digital Twins (IoT)	RDF, SHACL	JSON to Aspect Mod-els via bootstrapping	Human evaluation, Jena (RDF syntax)
Continued on next page...				

Table 2.2 – continued from previous page

Study	Core Task / Do-main	Logic	Neuro-Symbolic Inte-gration	Validation Method
<i>Category 3: Evaluation, Stability & Trustworthiness</i>				
Sequeda (2025) [18]	SQL Databases (Enterprise)	SPARQL	LLM query correction	Comparison with SQL ground truth
Allemang (2024) [19]	SQL Databases (Enterprise)	SPARQL	Ontology-based Error Detection + Repair	Execution Accuracy on Benchmark
Gashkov (2025) [20]	Query Filtering (Multilingual)	SPARQL	LLM-as-a-Judge binary classifier	Answer Trustworthiness on Benchmark
Adam (2025) [21]	Statement Veri-fication (Bio-sci)	RDF	RAG using External Snippets	Precision / Recall on fixed dataset
Meyer (2025) [22]	KGE Benchmark-ing (Web)	RDF, SPARQL	LLM-KG-Bench 3.0 Framework	Parseable Syntax and F1
Kosten (2024) [23]	Complex QA (KGQA)	SPARQL	Ontology-based prompt engineering	Execution Accuracy on Benchmark
Schmidt (2026) [24]	Systematicity Test-ing (Wiki)	SPARQL	CompoST: Composi-tional Testing	Compositionality F1 on ground truth
Tufek (2025) [25]	Artifact Validation (Industrial)	SPARQL	Zero-shot Instruction Prompting	Domain-specific Pre-cision, Recall, F1

2.4 Thematic Analysis

Following the described systematic selection process, the included studies were synthesized into a thematic analysis. This analysis aims to move beyond providing paper summaries, but instead constructing a coherent narrative that explores how Large Language Models and formal logic systems interact within the current research literature. This narrative is structured around three primary themes. First, an exploration of how high-stakes legal and regulatory texts are currently being formalized. Second, a technical deep-dive into the mechanics of translating natural language into structured symbolic logic. Finally, a critical assessment of how the correctness and stability of these systems are verified.

2.4.1 Neuro-Symbolic Pipelines in Public Administration

The integration of Large Language Models (LLMs) and Knowledge Graphs (KGs), frequently characterized as Neuro-Symbolic AI, seeks to combine the flexibility of neural networks with the logical rigor of formal ontologies [19]. One of the many goals of this intersection is to address the complexities of public sector data management [1]. In this domain, authors are increasingly moving away from unstructured legislative texts and narrative reports and towards formal logic, to enable proactive and data-centric governance [1][3].

Knowledge Extraction and Formalization

The transition from unstructured text to formal logic typically follows a multi-stage pipeline designed to reduce LLM hallucinations and preserve the semantic intricacies of legal rules [1][2].

Konstantinidis et al. [1] utilize LLMs to interpret complex legislative documents (in raw PDF format) describing public services, extracting preconditions for eligibility and translating them into SHACL (Shapes Constraint Language) rules. Their approach uses Retrieval-Augmented Generation (RAG) and prompt chaining to transform raw text into RDF-based evidence models, while ensuring that the extracted rules are grounded in established EU standards like *CPSV-AP* and *CCCEV*.

Similarly, Oranekwu et al. [2] employ a RAG pipeline to ingest regulatory texts and manufacturer privacy policies, using LLMs to extract subject-predicate-object triples that are then mapped into a compliance knowledge graph.

Spyropoulos and Tsiantos [3] focus on law-enforcement archives, using instruction-tuned models like OpenAI o3 to parse narrative police reports to extract entities and their interrelationships, subsequently converting this knowledge into OWL-compliant triples for ingestion into a triplestore.

The Role of Intermediate Models

Intermediate representations serve as critical "blueprints" or "mappers" that bridge the gap between unstructured narratives and executable logic [4][3].

Hanuragav and Gopinath [4] demonstrate the utility of intermediate representations through a multi-stage pipeline designed for regulatory validation. In their framework, the transition from unstructured rich-text documents into formal RDF is facilitated by a JSON-to-YAML mapper. By using LLMs to draft YAML mapper files rather than direct triples, their architecture decouples the semantic extraction of data from the technical generation of the knowledge graph (thus essentially decoupling logic and reasoning from syntax and formality). This reliance on non-executable intermediate schemas suggests a shift toward modularity in public sector pipelines, where the LLM's role is confined to architectural drafting, a failsafe to ensure that the resulting logic is structurally "anchored" before final conversion.

Konstantinidis et al. [1] also utilize intermediate steps to formulate natural language rules into a template format before final SHACL generation, allowing for the hierarchical structuring of evidence data.

Spyropoulos and Tsiantos [3] employ intermediate tabular forms to organize recognized entities before they are formally mapped to the OWL ontology, making human-in-the-loop validation much easier.

Current Status and Limitations

Despite promising results, these systems are currently characterized as conceptual or pilot-scale prototypes [1][2].

Konstantinidis et al. [1] emphasize that their pipeline is not yet end-to-end operational and faces significant hurdles regarding data fragmentation across administrative silos.

A primary critique of current methods is the lack of automated testing at scale. For instance, Oranekwu et al. [2] note that their ground truth dataset remains limited in statistical generalizability and has not yet undergone testing with end-users, in real-world conditions.

Furthermore, Spyropoulos and Tsiantos [3] admit to the use of simulated reports rather than authentic documents due to confidentiality, which may raise concerns about not fully capturing the complexity of real-world law-enforcement data.

Proposing future work in this sector, authors focus on overcoming legal and policy complexities, as continuous updates are required to accommodate rapidly evolving regulations [1][2]. Authors also suggest that federated Knowledge Graphs and decentralized technologies (like blockchain) may be necessary to address issues of data ownership and privacy compliance (such as GDPR requirements) [1]. Additionally, there is an identified need for more robust benchmarking methods to validate AI-driven interpretations against human-expert evaluations in high-stakes public environments [2][1].

2.4.2 State-of-the-art in Logic Synthesis

The field of logic synthesis has evolved from monolithic sequence-to-sequence translation toward modular, Neuro-Symbolic pipelines that compartmentalize the semantic parsing of natural language into manageable logical components [5][7]. These state-of-the-art approaches take advantage of the linguistic fluency of Large Language Models while enforcing the structural constraints of Knowledge Graphs, through various technical mechanisms and intermediate representations [16][12][9].

Technical Mechanisms for Translation

Notable Neuro-Symbolic techniques for bridging natural language and structured logic include Few-shot In-Context Learning (ICL) [6], Retrieval-Augmented Generation (RAG) [16] and Iterative Agentic Exploration [9].

Frameworks such as *SymKGQA* [5] combine few-shot ICL with function definitions, to generate symbolic programs in *KoPL* (Knowledge Oriented Programming Language), allowing step-by-step reasoning that is independent of the model’s pre-trained knowledge of language grammars. Shah et al. [8] further enhance this via what they refer to as "Planned Query Guidance", where few-shot examples demonstrate a code-style reasoning process that handles multi-hop transitions line-by-line.

To ground logic in specific KG schemas, authors utilize RAG variations that inject

minimal subgraphs [12], *VoID* (Vocabulary of Interlinked Datasets) descriptions or *ShEx* (Shape Expression) schemas into the prompt [16]. For example, *SPARQLGEN* [12] enriches prompts with a minimal RDF subgraph, sufficient to answer the query, reducing the need for models to memorize the entire graph. Emonet et al. [16] utilize *ShEx* to define available predicates for specific classes, which proved to significantly improve the model’s ability to generate valid federated queries.

The use of RAG is further extended beyond simple fact retrieval to the generation of *ABox* (Assertional Box, storing factual statements) instances for complex domain models. Vieira da Silva et al. [15] demonstrate that providing the full *TBox* (Terminological Box, the schema-level knowledge of an ontology), within the prompt context is essential for reducing hallucinations when modeling industrial capabilities. By explicitly injecting the *TBox*, the LLM is forced to conform with predefined class hierarchies and relationship constraints, such as domain–range axioms. This contextualization ensures that the generated instances are both linguistically plausible and logically consistent with the underlying ontology, successfully reducing the occurrence of model contradictions or invented properties.

Recent research introduces iterative frameworks like *GRASP* [9], which treat the LLM as an agent, tasked with exploring a graph through sequential function calls (search, list, execute). This methodology allows the model to iteratively refine its understanding of the graph’s topology, without being constrained by a fixed context window. Similarly, *SAMM Copilot* [17] employs iterative prompting and feedback loops to generate semantic Aspect Models from JSON data.

Beyond the choice of underlying model, the selection of the formal target language itself is a point of contention. A significant shift in logic synthesis came with the proposal by Lehmann et al. [11] to use Controlled Natural Languages (CNLs), such as *SQUALL* or *Sparklis*, as the target logical form instead of formal languages like SPARQL. The authors argue that because CNLs are linguistically closer to both the input question and the vast amounts of natural language text used in LLM pre-training data, they require significantly less fine-tuning to achieve high accuracy. Their findings indicate that despite the LLMs struggling with the strict syntax of SPARQL, they show a "deeper understanding" of CNLs, allowing them to generate valid syntactic variations that are semantically equivalent to ground truth. For especially complex queries (comparatives, ordinals, differences), switching the parsing target from SPARQL to a natural and compact language like *SQUALL* can effectively double the semantic parsing accuracy.

Managing Structural Complexity

Handling complex schema mapping, particularly in event-based or multi-hop scenarios, requires more advanced strategies than simple triple-matching [13][7].

Jiang et al. [7] propose *OntoSCPrompt*, a two-stage architecture that separates Query Structure Prediction (dubbed Stage-S) from KG Content Population (dubbed Stage-C).

In the first stage, the model predicts a sketch or "skeleton" of the SPARQL query, using special placeholders (e.g., [ent], [rel], [cct]), which is then populated with graph-specific identifiers in the second stage.

For event-based models like *CIDOC-CRM*, where answering a single question often requires traversing long complex paths, Mountantonakis and Tzitzikas [13] introduce a two-stage process using *Ontology Path Patterns*. Their method first predicts the required properties and classes to identify relevant paths of a specific radius before synthesizing the final query, effectively reducing the search space for the LLM.

Descriptive vs. Prescriptive Logic Synthesis

At this point of the analysis, it is important to make a distinction. While many (the majority, in fact) technical advancements focus on Descriptive Question Answering (QA), which is retrieving factual data such as birthplaces or award winners [5][14][9], a distinct and smaller subset of research addresses the synthesis of Prescriptive or Normative Rules [10][1].

Systems like *GraphRAG* [14] and *UniKGQA* [7] primarily focus on factual retrieval, done by translating natural language into executable queries (SPARQL, Cypher) to fetch stored static values [5][9].

In contrast, Soularidis et al. [10] and Konstantinidis et al. [1] attempt to synthesize formal logic that encodes rules for behavior or eligibility. Soularidis et al. utilize template-driven prompts and RAG to translate natural language rules from the Search and Rescue (SAR) domain into *SWRL* (Semantic Web Rule Language). Similarly, Konstantinidis et al. use LLMs to extract regulatory preconditions from legislative texts and formalize them as SHACL shapes, which serve as machine-readable rules for public service eligibility checks and recommendations. Oranekwu et al. [2] bridge these two domains by extracting triples from manufacturer privacy policies to verify compliance against structured *NIST* standards.

2.4.3 Methodologies for Logic Validation and Trust

If neuro-symbolic systems are to transition from conceptual pilots to operational environments, the methodologies used to validate their logical outputs have to become a primary focus of research; this idea is already shifting the emphasis from simple performance metrics to the establishment of trust and traceability [11]. Knowledge Graphs serve as the fundamental "source of trust" in these architectures, providing a formal framework to evaluate the validity of queries generated by Large Language Models and acting as a foundation for explaining results [18].

Knowledge Graphs as a Source of Grounding and Repair

The integration of KGs into the validation pipeline provides the necessary grounding to remedy the risks that stem from the probabilistic nature of LLMs [11].

Allemang and Sequeda [19] introduce *Ontology-based Query Check (OBQC)*, a deterministic approach that utilizes the semantic constraints of an ontology (such as domain-range rules) to identify errors in generated SPARQL queries without relying on an LLM. When an error is detected, an LLM-Repair mechanism utilizes the previously output deterministic error explanations, to iteratively prompt the model for correction, a cycle that continues until the query passes the ontological rules or reaches a predefined iteration limit.

This focus on traceability is further advanced by Adam and Kliegr [21], who propose an inherently traceable approach to verification by instructing LLMs to avoid internal factual knowledge and instead find justification for RDF statements within retrieved external text snippets, while directly generating references for every claim they make in the process and the end result.

The Probabilistic vs. Deterministic Divide

Current literature reveals a clear methodological divide between probabilistic and deterministic validation techniques.

Probabilistic approaches often rely on benchmarking and "LLM-as-a-Judge" frameworks. Gashkov et al. [20] employ instruction-tuned LLMs as binary classifiers to act as post-processing filters, removing incorrect SPARQL query candidates to enhance their *Answer Trustworthiness Score (ATS)*. Similarly, Kosten et al. [23] utilize the *Spider4SPARQL* benchmark [26] to evaluate model performance across varying levels of query hardness, finding that even state-of-the-art models struggle to surpass 51% accuracy on complex multi-hop tasks. Meyer et al. [22] provide an extensible framework, *LLM-KG-Bench 3.0*, which automates the evaluation of LLM answers using metrics such as normalized triple F1 and string similarity.

In contrast, high-stakes domains prioritize deterministic approaches and rigid constraints. Tufek et al. [25] focus on validating semantic artifacts against industrial standards (e.g., *OPC UA*), where natural language requirements are translated into SPARQL queries to ensure compliance. The requirement for absolute precision in high-stakes domains has led to the adoption of the "*draft only, never execute*" paradigm; as proposed by Hanuragav and Gopinath [4], this methodology explicitly prohibits the Large Language Model from direct interaction with the triple store. Instead, the model's output is restricted to intermediate mappers that are subsequently processed by a deterministic Python translator. This creates a "symbolic circuit breaker", ensuring that the final RDF output is idempotent, auditable and strictly compliant with regulatory requirements. This is a level of reliability that probabilistic filtering methods struggle to guarantee. Furthermore, SHACL shapes are frequently used as "logical gates" to enforce structural integrity and detect hallucinations in digital twins and public administration service models [4][1][15].

The Stability and Systematicity Gap

A critical shortcoming in current validation research is the *stability gap*, where models fail to maintain logical consistency across novel tasks. Schmidt et al. [24] investigate *systematicity*, the ability of an agent to understand complex expressions built from known components, through the *CompoST* benchmark. Their findings indicate that LLMs struggle to systematically interpret questions when the complexity of components deviates from their training samples, with performance scores rarely exceeding 0.57 even under optimal self-contained conditions. This highlights that most current validation is static, that is, comparing an LLM's output against a single "golden standard" answer in a single-pass execution [11][18][8].

2.5 Discussion and Research Gap

The systematic review of the current literature paints a certain picture regarding the trajectory of Neuro-Symbolic research, which is moving from monolithic sequence-to-sequence translation to modular pipelines, that distinguish semantic parsing from logical execution. However, the synthesis of the included studies identifies three fundamental gaps that remain unaddressed, as explained next.

2.5.1 The Prescriptive Synthesis Gap: From Facts to Rules

While the broader field of logic synthesis has achieved high accuracy in descriptive tasks, there is a marked lack of research into the synthesis of prescriptive governance logic. As established in the thematic analysis, dominant frameworks such as *UniKGQA* [7] and *GRASP* [9] focus almost exclusively on Knowledge Graph Question Answering (KGQA), that is, translating natural language into queries to retrieve stored facts. In contrast, the requirements of digital governance necessitate the synthesis of *Rules* rather than just *Questions*.

Even within Category 1 (Public Administration), where studies like Konstantinidis et al. [1] attempt to extract service preconditions, the methodology remains largely conceptual. There is a notable absence of a formal "Neuro-symbolic Bridge" that can serve as a structural blueprint to ensure that synthesized prescriptive logic (SHACL shapes) can survive the edge cases of diverse citizen profiles.

2.5.2 The Stability Paradox: Correctness vs. Resilience

A significant paradox emerges in how current literature defines "trust" and "correctness." Current state-of-the-art mechanisms, such as *Ontology-based Query Check (OBQC)* and *LLM-Repair* [19][18], are designed as one-off error-correction loops. These systems ensure that a specific generated query is semantically valid according to an ontology schema for a

single execution pass.

However, in the context of public service eligibility, a synthesized rule is not a one-off query. It is a permanent logic structure intended for high-frequency application across potentially heterogeneous datasets. The literature focuses on making the LLM a more accurate "translator" in the moment, but fails to ensure that the resulting logic is *stable* across a spectrum of inputs. There is no evidence of research into whether a synthesized SHACL shape, once generated, remains logically consistent when tested attributes are systematically varied, a requirement for the proactive "No-Stop Government" vision [1].

2.5.3 The Validation Vacuum: The Need for Mutation Testing

The most critical shortcoming identified is the reliance on static, single-pass validation methodologies. Benchmarks like *CompoST* [24] and *LLM-KG-Bench 3.0* [22] measure accuracy by comparing model outputs against a "golden standard" answer. While these metrics are useful for measuring retrieval precision, they are insufficient for verifying the functional correctness of eligibility rules.

As identified in Table 2.2, current validation is predominantly probabilistic. Even deterministic approaches, such as those by Hanuragav and Gopinath [4] or Tufek et al. [25], focus on structural syntax or version-control compliance. To date, no study has implemented a comprehensive *Mutation Testing Framework* to strictly test the structural and logical stability of LLM-generated SHACL shapes. For high-stakes digital governance, an 88% precision rate (as reported in descriptive tasks [21]) is not a useful nor representative result. Future frameworks must address the variability of non-deterministic systems by establishing testing methodologies that guarantee logical stability across constantly evolving regulatory landscapes.

While Sequeda et al. [18] acknowledge the importance of regression testing to ensure that accuracy does not decrease as ontologies are extended, there is no evidence of widespread use to ensure logic remains resilient under variable conditions. For high-stakes governance and public administration, measuring accuracy on a single pass is insufficient.

2.5.4 Contribution of this Study

Based on the gaps identified above, this dissertation proposes a Neuro-Symbolic framework that utilizes a *JSON Information Model* as an intermediate architectural bridge to generate SHACL-based eligibility rules. SHACL is chosen specifically for its ability to act as a "logical gate", providing the unified graph structure, explainability and automation required for public administration.

To address the validation vacuum, this study introduces a *Deterministic Mutation Testing* methodology. By systematically mutating citizen attributes to evaluate the "rejection rate" of synthesized SPARQL logic, the framework moves beyond probabilistic retrieval to provide a functional guarantee of stability. This approach, detailed in the following Pilot

Study, represents a transition from checking if a model is "correct once" to verifying that the synthesized law is "functionally infallible" across variable scenarios.

3 Pilot Study

This chapter details the design, implementation and experimental testing of a novel Neuro-Symbolic pipeline for modeling automating public service eligibility checks.

3.1 Overview

The proposed architecture addresses the limitations of "black-box" Large Language Models [27] by enforcing a strict separation between neural interpretation (extracting meaning from text) and symbolic execution (validating logic against data).

The methodology is structured around a "Text-to-Graph-to-Logic" workflow. The system transforms unstructured administrative documents into formal Knowledge Graphs and executable SHACL shapes through a chain of intermediate structured representations. This design prioritizes explainability and determinism, ensuring that the final eligibility decision is derived from explicit, auditable rules rather than probabilistic approaches.

This chapter is organized as follows: Section 3.2.2 defines the semantic schemas that ground the system. Section 3.2.3 details the four-stage extraction and generation pipeline. Section 3.2.4 describes the validation engine, and Section 3.3 outlines the experimental framework used to stress-test the system's capabilities through automated mutation testing.

3.2 Methodology and System Architecture

3.2.1 Setup Environment

The pipeline was implemented using Python 3.12.9, utilizing a modular architecture to separate core processing logic from experimental orchestration. The system relies on local processing for semantic graph operations and cloud-based APIs for interfacing with Large Language Models.

System Architecture

The codebase follows a functional separation of concerns, organized into three distinct layers:

1. **The Core Logic Layer:** A modular custom Python library encapsulating the functional logic of the system. It contains the reusable logic, such as API communication, graph operations, parsing and testing utilities. It also contains the end-to-end extraction-generation workflow, dubbed the *pipeline core*.

2. **The Orchestration Layer:** An interactive Jupyter Notebook serves as the control interface. This layer manages the experimental loop, injects configuration variables into the core modules and handles exceptions without interrupting the iterative execution of experiments.
3. **The Persistence Layer:** To ensure auditability and reproducibility, the system employs a meticulous "Artifact Preservation" strategy. Every experimental run generates a dedicated directory locally, containing all intermediate outputs of the core pipeline. Furthermore, during testing, metrics and metadata are saved in a Master CSV file for post-hoc analysis.

Technologies and Libraries

The system combines standard Semantic Web technologies with modern Data Science tools:

- **RDFLib:** Used for parsing, manipulating and serializing RDF graphs (in Turtle format), as well as executing local SPARQL queries.
- **PySHACL:** The standard Python implementation of the SHACL validation engine, used to validate the LLM-generated shapes against the citizen data.
- **Pydantic:** Employed for schema enforcement of the intermediate JSON representations. It ensures that the unstructured extractions from the LLM adhere to strict data types and structural constraints, acting as a structural "gatekeeper".
- **Pandas:** Used for the post-hoc aggregation and analysis of the testing logs.

3.2.2 Semantic Data Modeling

This pipeline was designed specifically with public service documents in mind. To bridge the gap between unstructured administrative text and deterministic validation logic, two distinct semantic layers were defined. These RDFS schemas serve as the symbolic "grounding" for the Large Language Model.

The Public Service Meta-Model

To ensure semantic interoperability and standardization, the modeling of the public service itself adheres to European formal vocabularies, specifically the Core Public Service Vocabulary Application Profile (CPSV-AP) and the Core Criterion and Evidence Vocabulary (CCCEV). The schema follows the following hierarchical structure:

- **cpsv:PublicService:** The root node representing the public service itself.
- **cccev:Constraint:** Connected to the root node via `cpsv:holdsRequirement`, these nodes represent individual preconditions extracted from the text.
- **cccev:InformationConcept:** These nodes are connected to Constraint nodes via `cccev:constrains` and represent the abstract information required to evaluate a constraint.

The adoption of established EU standards is a deliberate architectural choice, made to ensure cross-border interoperability and extensibility [28]. By anchoring the pipeline's output in the CPSV-AP and CCCEV ecosystems, the generated graphs are compatible with the broader European e-Government infrastructure (such as the Single Digital Gateway). Furthermore, this modular design allows for future expansion where the pipeline could automatically ingest the full breadth of these ontologies (complex Evidence mappings, Agent definitions, Output representations), without requiring a fundamental restructuring of the core logic.

Citizen Schema

While the Public Service Meta-Model describes the *rules*, the Citizen Schema describes the *applicant*. This work utilizes a domain-specific RDFS schema tailored to the requirements of each document and generated in a separate workflow (not presented here) by the same LLM used in the implementation of the rest of the pipeline. The model is instructed to use granular instead of aggregate data as nodes (e.g. prefer "Date of Birth" rather than "Age") and is encouraged to use abstract and reusable classes.

It has been demonstrated that the generation of such schemas can be automated as part of the pipeline [1]. However, for the scope of this pilot study, the Citizen Schema is treated as fixed input context. This methodological choice serves two purposes:

1. **Experimental Control:** By fixing the target schema, we isolate the performance of the LLM in *logic generation* (SHACL/SPARQL) and *extraction*, without the confounding variable of schema generation errors.
2. **Prerequisite for Testing:** The automated testing framework that will be presented relies on injecting specific faults into the citizen graph (e.g., modifying property values to trigger violations). This requires a schema structure that is known in advance. Had the schema been generated dynamically during each run, it would be impossible to define a library of test scenarios targeting specific graph nodes.

3.2.3 The Extraction and Generation Pipeline

The core contribution of this work is the following multi-stage, Neuro-Symbolic pipeline. The process follows a sequential data flow, depicted in Figure 3.1, consisting of four primary stages.

Stage 1: Document Summarization and Preconditions Extraction

The pipeline begins with the ingestion of the raw public service document (PDF). Using a Large Language Model, the unstructured text is processed to extract a summary of eligibility preconditions. The prompt is designed to filter out administrative noise and standardize the format of the rules. Since this is the first and most vital piece of information to be passed downstream, summarization is necessary as it reduces the cognitive load required for the

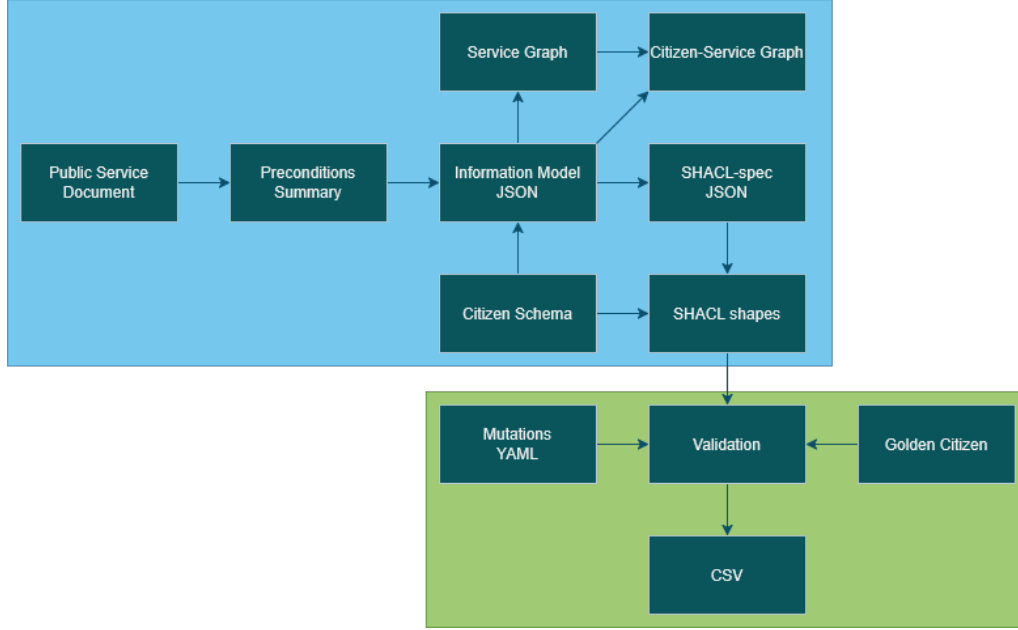


Figure 3.1: Flow Chart of the core pipeline

subsequent logic generation steps.

Stage 2: Information Model Generation

In this critical neuro-symbolic step, the extracted preconditions are transformed into a structured JSON "Information Model". The Information Model organizes the unstructured list of rules into a strict hierarchy that mirrors the Meta-Model structure:

- **Constraints:** Each eligibility rule is encapsulated as a Constraint object, containing the natural language description of the rule.
- **Information Concepts:** Nested within each Constraint are the abstract Information Concepts, representing the specific pieces of evidence or data required to evaluate that rule.

Inferring these concepts from the list of preconditions is the main reasoning task of the LLM at this stage. However, a second task it is prompted with is to act as a semantic mapper. The LLM is provided with the Citizen Schema (defined in Section 3.2.2) as a strict vocabulary constraint to prevent the hallucination of non-existent properties. With it, it is instructed to connect each Information Concept with a number of Citizen nodes, by constructing specific traversal paths through the ontology (e.g., mapping the concept of "Applicant Age" to the path `:Applicant → :birthDate`). This method has been shown to potentially increase the consistency and reliability of the produced results [29].

The output is strictly enforced using a Pydantic schema definition, ensuring valid JSON structure. The resulting artifact effectively creates a "blueprint" for downstream tasks. It contains all the necessary semantic links to be deterministically serialized into valid RDF triples in the subsequent stage, while ensuring that all data references are grounded in the

controlled vocabulary of the Citizen Schema. Listing 3.1 provides an example snippet of the JSON Information Model.

Listing 3.1: JSON Information Model snippet

```
1 {
2   "name": "family_income_condition",
3   "desc": "Annual family income must not exceed 30,000,
4         increased by 3,000 for each dependent child after the
5         first.",
6   "constrains": [
7     {
8       "name": "total_family_income",
9       "related_paths": [
10        {
11          "path": ["hasIncome", "amount"],
12          "datatype": "xsd:decimal"
13        },
14        {
15          "path": ["hasParent", "hasIncome", "amount"],
16          "datatype": "xsd:decimal"
17        }
18      ]
19    },
20    {
21      "name": "is_dependent_child_of_parents",
22      "related_paths": [
23        {
24          "path": ["hasParent", "hasChild", "isDependent"],
25          "datatype": "xsd:boolean"
26        }
27      ]
28    }
29  ]
30 }
```

Stage 3: Semantic Graph Construction

Once the Information Model is established, the system deterministically (via Python code) constructs two RDF artifacts without further LLM inference:

1. **The Service Graph:** A formal representation of the public service using the CPSV-AP and CCCEV vocabularies and following the Meta-Model schema defined in Section 3.2.2.

2. **The Citizen-Service Graph:** By loading an "Example Citizen" (a valid applicant instance), the system uses the Information Model to link the abstract Information Concepts from the Service Graph directly to the actual data nodes in the Citizen Graph via `ex:mapsTo` edges. This unified graph serves as a visual "audit trail," allowing human inspectors or automated agents to trace exactly which specific data points are being used to evaluate a specific constraint.

Both Graphs are serialized using Turtle syntax and saved to file as artifacts. Interactive visualizations of them are generated using the *PyVis* library and also saved to file as `html` files.

Figure 3.2 shows a simplified view of a citizen-service graph as it was automatically produced by the pipeline. The yellow node is the `cpsv:PublicService` node, the red nodes are `cccev:Constraint` nodes, the blue nodes are `cccev:InformationConcept` nodes. These connect to the Citizen nodes (Green, the root, and grey, the properties) completing the Citizen-Service Graph.

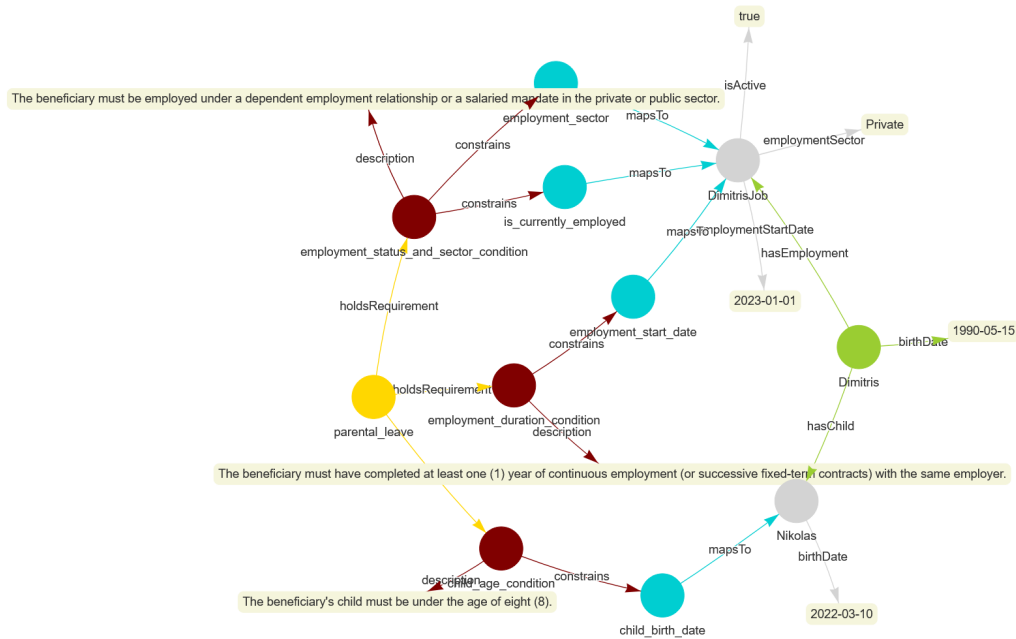


Figure 3.2: A citizen-service graph according to the Meta-Model

Stage 4: SHACL Shapes Generation

The final stage of the pipeline is responsible for synthesizing the executable validation logic. This stage transforms the abstract requirements from the Information Model into a strictly valid Shapes Constraint Language (SHACL) document. It should be noted that this is theoretically the most demanding task the LLM performs throughout this pipeline, from a semantic understanding perspective. The generation happens in two steps.

First, the system deterministically distills the rich Information Model into a simplified, noise-free JSON structure termed the "SHACL-Spec". This intermediate representation

reorganizes the structure and retains only the logical primitives required for validation (e.g. rules, target paths and data types). This step acts as a "context cleaner", helping the LLM focus exclusively on code synthesis.

Second, the LLM is invoked to translate this specification into RDF triples (Turtle format). The prompt enforces a Dual-Strategy Protocol for logic synthesis. For atomic constraints involving single-hop properties and literal comparisons (e.g., `Citizenship = 'GR'`), the model generates standard `sh:property` shapes. SHACL-Core provides a rich set of structural and logical components (e.g., `sh:and`, `sh:or`, `sh:not`) and can encode fairly complex conditional logic and exception patterns [30].

For requirements involving arithmetic, aggregations, date comparisons or cross-referenced data (e.g., `now() - birthDate > 18`), the model encapsulates the logic within a `SPARQLConstraint` as per the standard and recommended approach in the literature [31][32]. This allows for the expression of complex conditional logic that exceeds the expressivity of the SHACL Core vocabulary [33]. The model is once again restricted to using the fixed Citizen Schema, which is once again given as context to act as a failsafe, in case earlier path generation failed to include crucial nodes.

As a last addition, the LLM generates an error message for every shape, which is intended to be displayed as part of the Validation Engine report in case of a violation (e.g., "Income exceeds threshold").

The output is a fully serialized `ttl` file containing the `sh:NodeShape` definitions. This file serves as the executable input for the Validation Engine, the mechanics of which are detailed in the next section.

Listing 3.2 shows a SHACL Shape and its embedded SPARQL query generated by this process. Specifically, it is the shape that corresponds to the constraint we showed in Listing 3.1 in its Information Model form.

Listing 3.2: A SHACL Shape generated by the pipeline

```

1 :family_income_shape
2   a sh:NodeShape ;
3   sh:targetClass :Applicant ;
4   sh:sparql [
5     a sh:SPARQLConstraint ;
6     sh:message "Annual family income exceeds the limit." ;
7     sh:select """
8       SELECT ?this
9       WHERE {
10         {
11           SELECT ?this (SUM(?val) AS ?totalIncome)
12             WHERE {
13               { ?this :hasIncome ?inc .

```

```

13         ?inc :amount ?val . }
14     UNION
15     { ?this :hasParent ?p .
16         ?p :hasIncome ?p_inc .
17         ?p_inc :amount ?val . }
18     } GROUP BY ?this
19 }
20 OPTIONAL {
21     {
22         SELECT ?this (COUNT(DISTINCT ?child)
23             AS ?depChildCountResult) WHERE {
24             ?this :hasParent ?parent .
25             ?parent :hasChild ?child .
26             ?child :isDependent true .
27         } GROUP BY ?this
28     }
29     BIND(COALESCE(?depChildCountResult, 0) AS ?
30         depChildCount)
31     BIND(30000 + (3000 * (IF(?depChildCount > 1, ?
32         depChildCount - 1, 0))) AS ?limit)
33     FILTER(?totalIncome > ?limit)
34 }

```

3.2.4 The Validation Engine

The final component of the architecture is the Validation Engine, which functions as the execution core of the system's symbolic layer. While previous stages focus on structuring and grounding the data, this engine is responsible for applying the generated constraints against specific citizen data to render a final, deterministic eligibility decision.

The engine operates on two distinct RDF graphs:

- **The Shapes Graph:** The .ttl file generated by Stage 4 of the pipeline, containing the `sh:NodeShape` definitions and SPARQL constraints within.
- **The Data Graph (Citizen Instance):** An RDF graph representing a specific applicant (a concrete instantiation of the Citizen Schema). It contains the factual assertions about an individual, structured strictly according to the domain ontology.

For the Execution and Reasoning step, the system utilizes PySHACL, a Python-based implementation of the W3C SHACL standard, to perform the validation. The execution follows a standard protocol:

- **Targeting:** The engine identifies the "Focus Node" in the Data Graph (defined during Stage 4 as the sole instance of the class `:Applicant`).
- **Constraint Evaluation:** For every Shape mapped to the Applicant, the engine evaluates the corresponding logic. Simple property shapes are validated via graph traversal, while complex conditions trigger the execution of the embedded SPARQL queries against the Data Graph.
- **Entailment:** The engine operates under the RDFS entailment regime, allowing it to infer class hierarchies (e.g., understanding that a `:Child` is also a `:Person`) during validation.

The output of the engine is a formal *Validation Report Graph* adhering to the SHACL standard. This report provides as output:

1. **Boolean Conformance:** A global `sh:conforms` value (True/False), which serves as the system's final decision on eligibility.
2. **Violation Details:** A set of `sh:ValidationResult` nodes in cases of non-conformance. Each one links to the specific Shape that failed and includes the generated error message, providing explanation for the rejection.

3.3 Experimental Design

To evaluate the reliability, functional correctness and operational stability of the proposed architecture, an experimental framework was developed. The design moves beyond simple anecdotal testing, implementing a means to quantify the performance of the Neuro-Symbolic pipeline under varying conditions.

The core unit of the experiment is defined as a "run". A run represents a single end-to-end execution of the pipeline, characterized and configured by a specific combination of variables, dubbed the *Configuration Tuple*:

(Document, Model, Prompting Strategy)

Even when generating executable code, Large Language Models are inherently non-deterministic when operating at non-zero temperature settings and can generate vastly different results given the exact same configuration (inputs, prompt, context) [34]. For some model architectures, even setting temperature to *zero* reduces but *does not eliminate* non-determinism [35]. Consequently, a single generation is insufficient to prove or disprove their reliability.

To address this, the framework executes a loop of multiple iterations for each unique configuration. In fact, empirical recommendations emphasize multiple generations to draw reliable conclusions [36]. This repetition allows for "drowning out" stochasticity and for the results metrics to converge to values that describe the actual stability of the pipeline with more fidelity.

The execution of these runs is done in The Orchestration Layer (see Section 3.2.1), which oversees the following lifecycle for every iteration:

1. **Context Initialization:** At the start of a run, a dictionary is initialized. This volatile data structure acts as a "flight recorder", accumulating outputs and metadata.
2. **Pipeline Execution:** The extraction and generation pipeline is triggered. If the pipeline encounters a critical failure, the failure mode is logged and the run is marked as incomplete before moving to the next.
3. **Scenario Validation:** Upon successful generation of a valid SHACL graph, the system proceeds to the Mutation Testing phase (detailed in the following subsection), where the generated logic is stress-tested against a battery of specific, pre-made scenarios.
4. **Persistence:** Finally, the accumulated metrics are "flushed" to a CSV file. Results are persisted immediately to prevent data loss during long-running batch experiments.

3.3.1 The Mutation Testing Framework

To evaluate the functional correctness of the generated SHACL shapes, the system implements a Mutation Testing Framework. Mutation testing has matured and gained popularity in evaluating test suites and supporting experimentation [37]. Unlike traditional unit tests that might check for static string matches, this framework dynamically generates RDF graph instances to test whether the generated logic correctly distinguishes between eligible and ineligible applicants. The framework operates on a "Baseline and Perturbation" model, consisting of the components analyzed below.

The "Golden Citizen" Baseline

For each public service document, a single, syntactically perfect RDF graph termed the *Golden Citizen* is manually constructed. This data instance represents an applicant who satisfies *all* eligibility preconditions. This baseline graph is constructed to adhere strictly to the Citizen Schema. The data values are calibrated to demonstrate marginal eligibility (e.g., if an income upper limit is €12,000, the Golden Citizen might have €11,999). This ensures that the testing framework evaluates the precision of the logic, not just its general functionality.

Scenarios

The test cases (Scenarios) are defined in a declarative YAML configuration file. Each entry in this file represents a distinct Scenario, designed to isolate and test a specific logical constraint found in the document. A Scenario definition includes:

1. **Expected Violation Count:** The ground truth for the test. A compliant scenario expects 0 violations, a failure scenario typically expects 1.
2. **Mutation Actions:** A set of instructions to alter ("mutate") the Golden Citizen.

Listing 3.3 provides an example of one mutation scenario.

Listing 3.3: YAML description of a scenario

```
1 - id: SCN_03
2   description: "Base income is too high."
3   expected_violation_count: 1
4   actions:
5     - type: patch_node
6       turtle: ex:DadIncome :amount 26001.0 .
```

Crucially, mutations are designed to be atomic. Each scenario targets a single "fact" in the graph (e.g., changing a Literal value or a URI reference) to nudge the applicant from an "Eligible" state to a "Non-Eligible" state. This isolation allows the Validation Engine to pinpoint exactly which specific rule the LLM failed to generate correctly, if any.

For this work, all mutation scenarios were hand-crafted to reinforce confidence in the results. However, recent work has used LLMs to generate mutants, showing that mutation scores align with detection of real errors, extending classic mutation assumptions into LLM settings [38].

The Mutation Engine

For every iteration ("run"):

1. The system loads the Golden Citizen graph into memory.
2. It creates a deep copy of the graph to ensure test isolation.
3. Once per scenario, it applies the Patch Logic. The engine parses the Turtle snippets defined in the YAML actions (e.g., `ex:Income :amount 12,000.1`) and updates the graph triples accordingly. This allows for complex graph transformations, such as replacing nodes or updating relationships, without manual RDF manipulation.

The resulting Mutated Citizen Graph and the Generated Shapes Graph (from Stage 4) are then passed to the aforementioned Validation Engine (section 3.2.4). The boolean outcome (conforms) and the number of violations with their associated messages are captured and logged to later be compared against the Expected Violation Count defined in the scenario.

3.3.2 Experimental Configurations

Recall the configuration tuple around which the experiment was designed:

(Document, Model, Prompting Strategy)

For this work we chose 2 documents, 2 models and 3 prompting strategies, for a total of 12 different experimental configurations. This combinatorial approach allows for the isolation of specific failure modes, distinguishing between errors caused by document complexity, model reasoning capacity, or prompting sufficiency. Below we analyze each component of the tuple and the configurations explored in the scope of this work.

Document Corpora (Use Cases)

This selection tests the pipeline’s ability to generalize across different domains and logical structures. This discrepancy was deemed necessary, as research indicates that LLMs often experience a non-linear degradation in performance, as reasoning depth and contextual density increase [39].

Two public service documents were selected to represent two different levels of bureaucratic complexity.

Student Housing Allowance (High Complexity)

Selected as the "Stress Test" for the system. This document is characterized by:

- **Deep Graph Traversal:** Verification requires traversing multiple hops (Applicant → Parents → Properties → Location).
- **Recursive Arithmetic:** It involves dynamic income thresholds, calculated based on the count of dependent children (e.g., $Limit = Base + (N \times Bonus)$).
- **Referential Integrity Constraints:** Verification requires comparing the identity of URI nodes rather than literal values (e.g., validating that the :UniversityCity node is distinct from the :FamilyResidenceCity node).

Special Parental Leave Allowance (Intermediate Complexity)

Selected to evaluate standard administrative processing. This document focuses on:

- **Categorical Classification:** Eligibility relies on specific enumerated values (e.g., Employment Sector must be "Private" or "Public").
- **Temporal Logic:** Involves duration calculations (e.g., "1 year of continuous employment") rather than complex arithmetic aggregations.

Large Language Models

The experiment uses two of the models in the Google Gemini 2.5 family, to evaluate the trade-off between reasoning capability and computational efficiency.

- **Gemini 2.5 Pro:** The high-parameter "reasoning" model. It is hypothesized to excel at complex SPARQL generation and abstracting vague requirements into formal logic, potentially at the cost of higher latency.
- **Gemini 2.5 Flash:** The lightweight, low-latency model. It serves to test the feasibility of a "high-throughput" pipeline. A key research question is whether this smaller model can adhere to the strict SPARQL syntax requirements without the deep reasoning capabilities of the Pro variant.

Prompting Strategies

Three distinct prompting strategies were implemented to evaluate the impact of "In-Context Learning" and "Self-Correction" on code quality.

Default Strategy (Few-Shot with Guardrails)

This strategy represents the baseline optimized approach. The system prompt instructs the model to act as an "Expert", a *role-playing* strategy that can potentially improve answer quality for some models and tasks [40]. The prompt also provides:

- **Proposed Strategy:** Explicit instructions to choose between, a simplification of the *Plan-and-Solve* strategy [41].
- **Syntactic Guardrails:** A set of negative constraints, derived from pilot testing frequent errors, as an attempt to improve alignment with syntactic expectations [42].
- **Few-Shot Examples:** Concrete examples demonstrating correct and desired outputs, a method that has shown to consistently improve results with harder extraction tasks, especially when examples clarify desired structure [43].

Zero-Shot Strategy (Ablation Study)

To quantify the value of the engineering effort put into the Default prompt, the Zero-Shot strategy removes all Few-Shot Examples: the model is given the instructions but no reference implementations. This tests the model's innate reasoning prowess and knowledge of syntax versus its reliance on pattern matching from examples. This was also an attempt to see if diminishing returns in simple tasks and small context, observed in other studies [44], will affect our use cases.

Reflexion Strategy (Iterative Self-Correction)

This strategy implements a *Prompt Chaining* loop that has the following steps:

1. The model generates a draft response using the Default strategy.
2. The output is passed back to the model with a new "persona": "*Senior Data Quality Assurance Auditor*." This agent is instructed to critique the quality of the draft with regards to criteria such as completeness, logical contradictions and syntactic validity.
3. If errors are found, the model rewrites the response based on its own critique.

Listing 3.4 shows an excerpt of the reflexion prompt.

Listing 3.4: Excerpt of the reflexion prompt

```
1 You are a Senior Data Quality Assurance Auditor .
2 Your task is to review the DRAFT RESPONSE provided below ,
   which was generated based on the ORIGINAL INPUT DATA .
3 1. Analyze the Draft Response for :
4   - Completeness: Did it miss any details from the input ?
5   - Logic: Are there contradictions or hallucinations ?
6   - Syntax: If the output is code (JSON/SPARQL/Turtle), is it
   syntactically valid ?
7 ...
```

This configuration evaluates the efficacy of self-correction mechanisms in code generation, specifically testing whether the computational overhead of iterative refinement yields a statistically significant reduction in errors [45].

3.3.3 Evaluation Metrics

To move beyond qualitative observation, the experimental framework was designed in such a way to capture a granular dataset for every execution cycle. This data collection strategy was designed to decouple structural failures (code that does not compile) from logical failures (code that compiles but yields incorrect decisions), enabling a multi-dimensional analysis of pipeline performance.

Data Collection

For every experimental run, the system persists a dataset that captures the complete state of the pipeline at the moment of execution, categorized into five distinct dimensions:

- **Configuration Metadata:** Contextual fields regarding a unique Run ID, timestamp, the specific document input, the LLM employed and the active prompting strategy.
- **Artifact Fingerprinting:** To track the stability and uniqueness of the LLM's output, the system computes and logs the cryptographic hashes (MD5) of the generated graphs. This allows for the detection of potentially identical artifacts generated across different runs.
- **Syntactic Integrity Verification:** Before execution, the system first verifies if the generated text is a valid RDF/Turtle graph (parsable by RDFLib), and secondly, it performs a "deep compile" check on every embedded SPARQL constraint to ensure the query syntax adheres to the SPARQL standard. Both errors, if they occur, are flagged differently to be distinguishable.
- **Validation Outcome Metrics:** The raw output of the validation engine is captured in detail. This includes the Actual Violation Count, the Expected Violation Count (derived from the scenario definition) and a serialized list of the specific Violated Shapes and their associated error messages. These fields enable the calculation of granular error metrics beyond simple binary accuracy.
- **Operational Diagnostics:** To monitor system health, metrics such as end-to-end Execution Time (latency) and Runtime Error Messages (e.g., Python exceptions) are logged. These fields are critical for quantifying the operational stability of components such as the external API.

Performance Indicators

The analysis of this dataset focuses on two primary dimensions of success.

Syntactic Validity

The first hurdle for any code-generating system is the production of executable syntax [46]. This metric quantifies the percentage of runs where the LLM produced a `.ttl` file that could be successfully parsed by the RDFLib graph library *and* whose embedded SPARQL queries could be compiled without error. A run that fails this check is distinguished from runs that simply produce incorrect logic.

Functional Logic Accuracy

For runs that pass the syntax check, the focus shifts to logical fidelity. This is measured by comparing the system's eligibility decision against the known ground truth of the mutation scenarios. By treating the validation outcome as a binary classification task, where a "Conformance" is the Positive class and "Violation" is the Negative class, standard machine learning metrics can be calculated.

3.4 Conclusion

This chapter has detailed the architectural and experimental foundations of the Neuro-Symbolic pipeline. By combining a schema-grounded generation process with a deterministic mutation testing framework, the system is designed to provide a quantifiable evaluation of LLM capabilities in the context of this task. The following chapter presents the results of these experiments, analyzing the pipeline's performance across the aforementioned dimensions.

4 Results

This chapter presents the quantitative findings obtained from the experimental evaluation of the neuro-symbolic pipeline. The analysis strictly follows the performance metrics defined in the methodology, assessing the system across three cascading thresholds of success: syntactic validity (code generation), functional logic accuracy (reasoning fidelity), and overall operational reliability (end-to-end feasibility). Broader interpretation of these patterns and their implications for public administration systems are discussed in Chapter 5.

4.1 Experimental Dataset

The experimental campaign consisted of a total of 170 end-to-end pipeline executions ("Runs"). Each run consisted of the steps that were analyzed on the previous chapter. The distribution of these runs across the varying configurations is detailed in Table 4.1. Due to the operational constraints discussed in Section 6.1, the dataset is unbalanced, with the "Flash" model variant accounting for a larger proportion of the total runs.

Table 4.1: Distribution of Experimental Runs per Configuration

Document	Model	Prompt Strategy	Runs (<i>N</i>)
Parental Leave	gemini-2.5-flash	Default	20
		Reflexion	20
		ZeroShot	20
	gemini-2.5-pro	Default	10
		ZeroShot	10
Student Housing	gemini-2.5-flash	Default	20
		Reflexion	20
		ZeroShot	20
	gemini-2.5-pro	Default	10
		Reflexion	10
		ZeroShot	10

4.2 Syntactic Validity

The first criterion for the pipeline's utility is the generation of syntactically valid code. A run is considered "Syntactically Valid" only if the LLM produces a Turtle (`.ttl`) file that can be parsed by *RDFLib* *and* contains SPARQL constraints that successfully compile without syntax errors.

4.2.1 Success Rate

Figure 4.1 illustrates the success rates across all configurations.

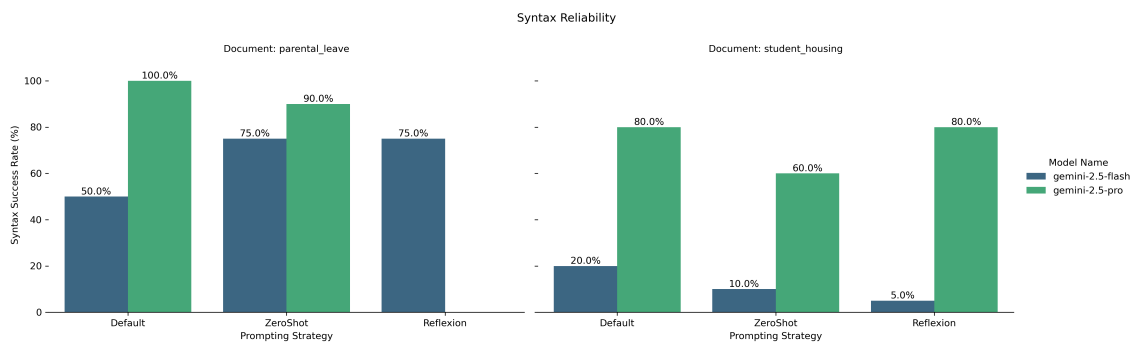


Figure 4.1: Syntactic Validity Rates by Configuration

Impact of Document Complexity

As was expected, the complexity of the source document served as a strong predictor of failure.

In the case of the Parental Leave document (Intermediate complexity), both models performed adequately. Even the weaker Flash model achieved a 75% validity rate using the Reflexion and ZeroShot strategies.

On the contrary, the Student Housing document (High complexity) acted as a strict filter. Flash failed to produce valid code in the vast majority of attempts (36 out of 60 runs failed syntax checks), whereas Pro proved strong enough to handle the increased logical depth, though it still suffered a 20% degradation compared to the simpler use case.

Impact of Model Class

The data reveals a disparity in coding capability between the two model variants.

The Gemini 2.5 Pro model demonstrated high reliability, achieving a 100% success rate on the Parental Leave document and maintaining an 80% success rate on the complex Student Housing document (under Default prompting).

In contrast, the Gemini 2.5 Flash model struggled significantly with syntactic precision. While it achieved moderate success on the simpler Parental Leave document (ranging from 50% to 75%), its performance collapsed on the complex Student Housing document, with

success rates dropping as low as 5% (Reflexion) to 20% (Default).

Impact of Prompting Strategy

The impact of prompting strategies varied by model architecture.

For Flash, the *Reflexion* strategy provided a significant boost on the simpler document (improving validity from 50% to 75%). However, this benefit vanished on the complex document, where Reflexion actually performed worse (5%) than the Default prompt (20%).

For Pro, the *Default* and *Reflexion* strategies performed identically (80% on Housing), while the *ZeroShot* strategy resulted in a notable drop in stability (falling to 60% on Housing).

4.2.2 Failure Mode Analysis

To better understand the mechanisms of failure, the invalid runs were categorized by error type: *RDF Syntax Errors* (invalid Turtle file structure) and *SPARQL Syntax Errors* (malformed queries within valid Turtle). Figure 4.2 presents the error rates normalized by the total number of runs for each model-document pair.

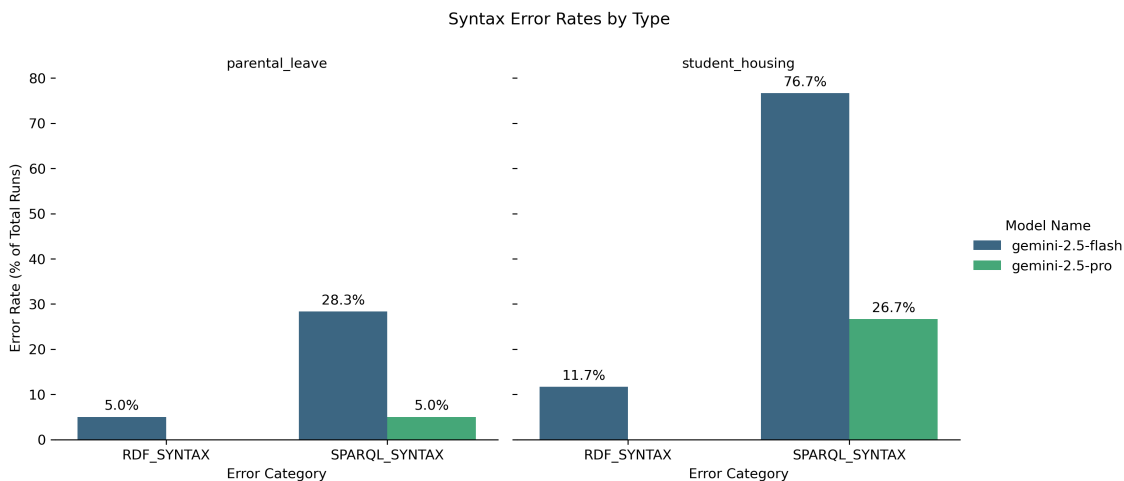


Figure 4.2: Distribution of Syntax Error Types by Model and Document

The data indicates that SPARQL Syntax Errors were the dominant failure mode across all configurations.

Gemini 2.5 Flash exhibited a high frequency of SPARQL errors, particularly on the complex Student Housing document, where 76.7% of all runs failed due to query syntax. Notably, Flash also produced a non-negligible rate of RDF Syntax errors (5.0% on Parental Leave, 11.7% on Student Housing), indicating occasional failures in generating even the fundamental file structure.

Gemini 2.5 Pro demonstrated significantly higher stability. It produced zero RDF syntax errors across all 50 experiments. Its failures were exclusively confined to SPARQL syntax, with error rates of 5.0% on the simpler document and 26.7% on the complex document.

4.3 Logic Validity

For the subset of runs that successfully produced syntactically valid code, the focus shifts to *Functional Logic Accuracy*. A run is classified as having "Perfect Logic" if and only if the generated SHACL shapes correctly identify the expected number of violations for *every single scenario* in the test suite (both the baseline Golden Citizen and all edge cases). Runs that crashed during execution or failed syntax checks were excluded from this analysis to isolate the reasoning capability of the models.

It is critical to note that 'Logic Accuracy' is evaluated as a holistic, end-to-end performance metric. A failure to correctly validate a citizen scenario may stem from errors at any stage of the neuro-symbolic pipeline: a missed precondition during the initial summarization (Stage 1), a malformed mapping in the Information Model (Stage 2), or an incorrect SHACL generation (Stage 4). Consequently, a 'Logic Failure' indicates that the system, as a whole, failed to enforce the regulation, regardless of which specific component was the root cause.

4.3.1 Success Rate

Figure 4.3 illustrates the rate of flawless logical execution across configurations.

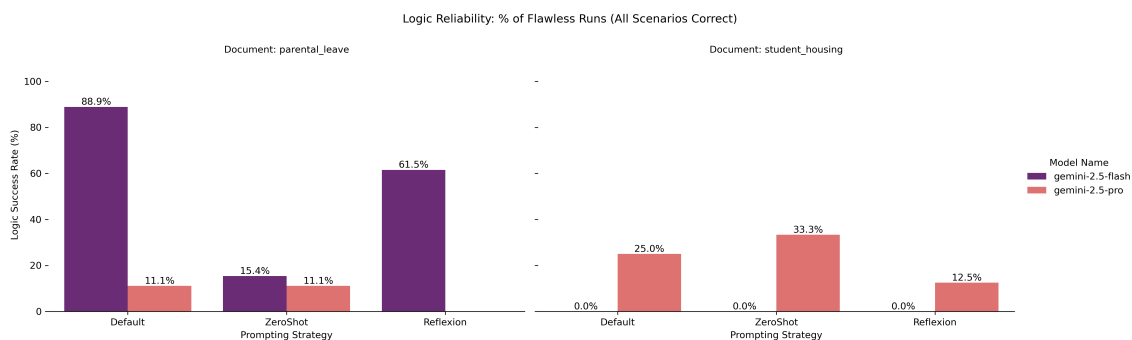


Figure 4.3: Logic Validity: Percentage of Flawless Runs (All Scenarios Correct)

Impact of Document Complexity

Consistent with the syntax results, the complexity of the document was the primary determinant of success.

Parental Leave, having simpler logic, allowed for high performance, with the best-performing configuration (Flash/Default) achieving an 88.9% perfect run rate.

Student Housing and its complex logic requirements caused a near-total collapse in functional accuracy. Across all models and prompts, the highest achieved success rate was only 33.3%, with many configurations failing to produce a single logically correct run.

Impact of Model Class

The performance relationship between models *inverted* depending on the task.

On the simple document, Flash significantly outperformed Pro, achieving 88.9% accuracy (Default) compared to Pro's 11.1%. However, on the complex document, Flash failed completely, with a 0.0% success rate across all 60 attempts.

While Pro underperformed on the simple task, it was the only model capable of solving the complex Student Housing logic, achieving success rates between 12.5% and 33.3%.

Impact of Prompting Strategy

Removing examples (ZeroShot) caused a sharp drop in accuracy from 88.9% to 15.4% for Flash on the simple document. Conversely, for Pro on the complex document, ZeroShot unexpectedly yielded the highest accuracy (33.3%).

The self-correction strategy (Reflexion) did not yield consistent improvements. For Flash, it reduced accuracy from 88.9% to 61.5% on the simple document. For Pro, it performed roughly equivalent to the Default strategy.

4.3.2 The Syntax-Logic Gap

A comparison between the syntax validity rates (Figure 4.1) and logic accuracy rates (Figure 4.3) reveals a distinct degradation in performance as the evaluation becomes stricter. This is apparent if we isolate the complex Student Housing use case. While the Pro model generated valid syntax in $\approx 80\%$ of runs, only $\approx 25\%$ of those valid runs contained correct logic. The Flash model struggled at both levels, with low syntax validity (20%) and zero functional correctness (0%).

4.4 Overall Pipeline Reliability

Beyond specific syntax and logic metrics, this section evaluates the system's viability as an end-to-end automated service. The analysis considers two perspectives: the operational stability of the pipeline and its reliability inside the broader context of this work, which is public service recommendations.

4.4.1 Pipeline Feasibility and Attrition

Figure 4.4 presents the distribution of final outcomes for all 170 experimental runs. This "Waterfall Analysis" categorizes every attempt into a single mutually exclusive outcome, revealing the attrition rate of the system.

The data indicates a high attrition rate:

- **Syntax Failures:** The majority of runs failed early. SPARQL Syntax Errors accounted for the largest share of failures (N=72), followed by RDF Syntax Errors (N=10).
- **Logic Failures:** Of the runs that compiled, a significant portion (N=55) produced code that executed but failed to correctly validate all test scenarios.

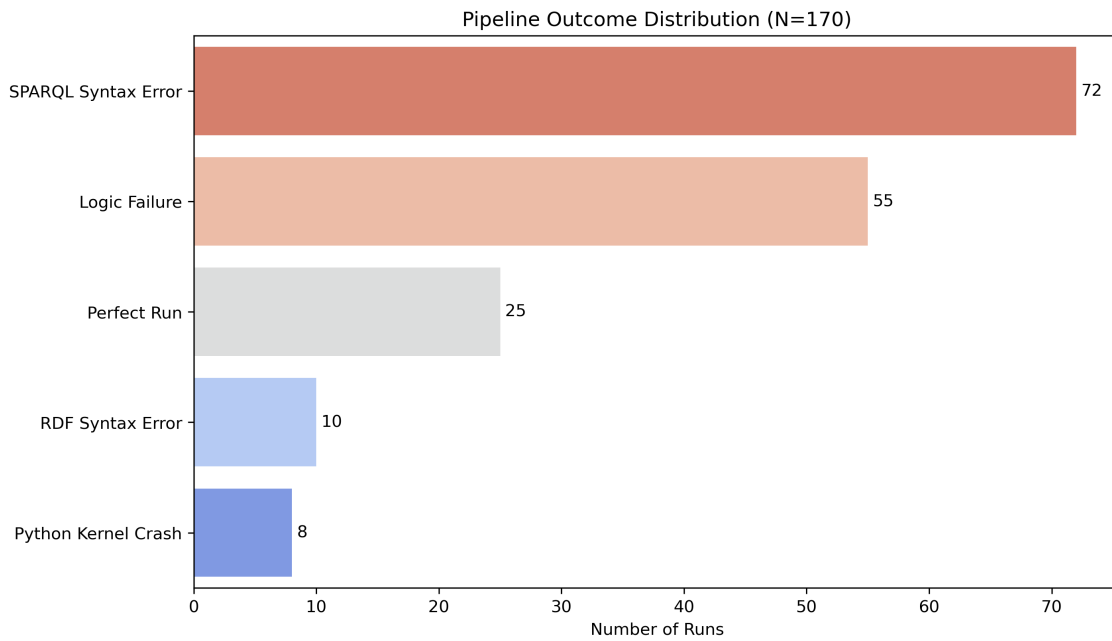


Figure 4.4: Distribution of Final Pipeline Outcomes.

- **Success:** Only 25 runs (14.7% of the total) achieved the status of a "Perfect Run," generating both valid syntax and flawless logic across all edge cases.
- **System Stability:** Operational crashes (Python/API errors) were rare (N=8), indicating that the underlying infrastructure, retry mechanisms and exception handling were largely sufficient.

4.4.2 In-context (Recommender System) reliability

To evaluate the system's utility as a public service recommender, the validation outcomes were aggregated into a Confusion Matrix (Figure 4.5). In this context, the classes are defined based on the goal of recommending eligible services:

- **Positive Class (Recommendation):** The system validates the citizen as Eligible.
- **Negative Class (Rejection):** The system flags at least one Violation.

To interpret the confusion matrix in the specific context of public service recommendations, the standard machine learning classifications were mapped to domain-specific service outcomes, as defined in Table 4.2.

The confusion matrix is then created based on this terminology.

The matrix reveals the system's risk profile:

- **True Positives (10.5%):** In 59 cases, the system correctly identified and recommended the service to an eligible citizen ("Correct Recommendation"). This confirms the system's ability to successfully validate legitimate claims when the generated logic is sound.
- **False Positives (10.1%):** In 57 cases, the system erroneously recommended the service to an ineligible citizen ("Bad Recommendation"). This represents a "Trust

Table 4.2: Definition of Classification Outcomes in the Recommender Context

	System: "Violation" (Rejection) <i>Triggered Violations > 0</i>	System: "Conforms" (Recommendation) <i>Triggered Violations = 0</i>
Citizen is Ineligible <i>Expected Violations > 0</i>	True Negative (TN) <i>Correct Rejection</i> (System works)	False Positive (FP) <i>Bad Recommendation</i> (Trust Risk)
Citizen is Eligible <i>Expected Violations = 0</i>	False Negative (FN) <i>Missed Opportunity</i> (Service Failure)	True Positive (TP) <i>Correct Recommendation</i> (System works)

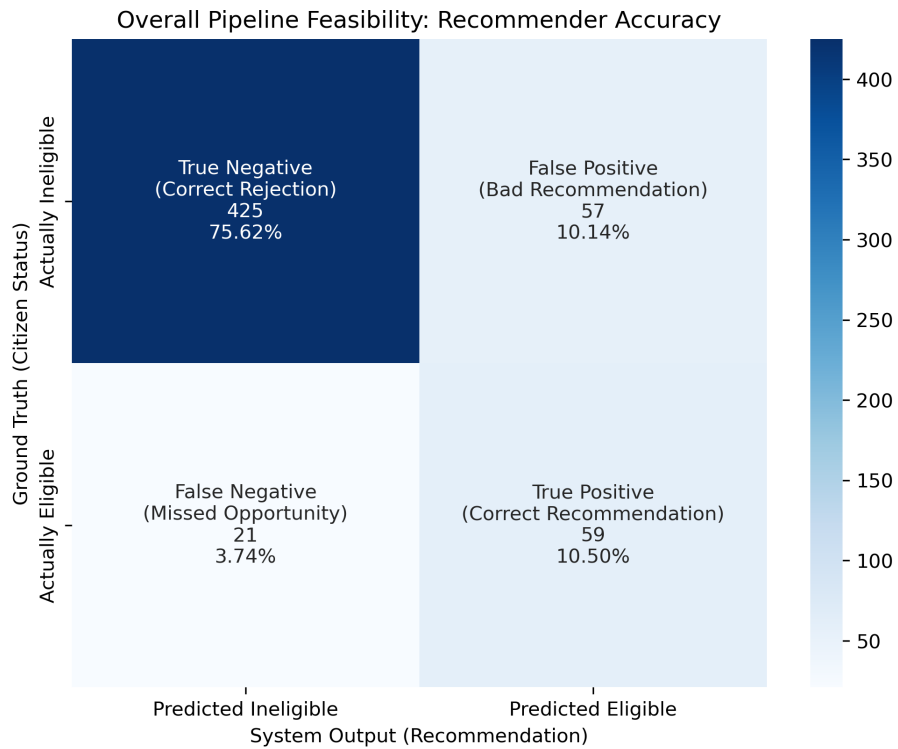


Figure 4.5: Confusion Matrix of Eligibility Recommendations

Risk," where users might be guided to apply for benefits they cannot receive.

- **True Negatives (75.6%):** The system correctly rejected ineligible applicants in the majority of cases.
- **False Negatives (3.7%):** In 21 cases, the system incorrectly rejected an eligible applicant ("Missed Opportunity"). While this number is low, it represents a "Service Failure," denying access to entitled benefits.

5 Discussion

This chapter synthesizes the quantitative results presented in Chapter 4 to evaluate the broader implications of using Large Language Models for automated public service eligibility checks and recommendation. By analyzing the patterns of failure, ranging from syntactic hallucinations to logical paradoxes, this discussion aims to characterize the fundamental limitations of current neuro-symbolic architectures. The analysis moves beyond simple performance metrics to address the core challenges of semantic fidelity, algorithmic determinism and the operational sovereignty required for deployment in public administration.

5.1 Cognitive Dissonance in Code Generation

The most pervasive pattern observed throughout the experimental campaign was a fundamental disconnect between the Large Language Model's ability to generate valid *syntax* and its ability to construct valid *logic*. This phenomenon, termed here as "Cognitive Dissonance", underscores the architectural limitation of Transformer-based models when applied to formal reasoning tasks: they operate as approximate pattern matchers in a domain that requires exact symbolic execution.

5.1.1 The Illusion of Fluency

The results from the "Student Housing" use case serve as the primary evidence for this dissonance. The Gemini 2.5 Pro model achieved an 80% syntactic validity rate, successfully producing well-formed Turtle files with structurally correct SPARQL queries. To a human reviewer, this code appeared indistinguishable from expert-written logic. However, the functional accuracy of this "valid" code was only $\approx 25\%$.

This discrepancy reveals that the model has successfully memorized the *grammar* of SHACL (e.g., correct brackets, prefixes, keywords) but failed to grasp the *semantics* of the query it was constructing. The model "knows" how to write a query, but it does not "understand" well enough what the query actually calculates.

5.1.2 Structural Logic Failures

The model's inability to properly understand graph topology led to recurring critical failures in the generated SPARQL constraints. Below we analyze some of the most prominent mistakes discovered upon human inspection of the runs flagged by the system as not perfect.

The Double-Counting Trap

In scenarios involving family units (for example, two parents with two children), the model consistently failed to apply set-theoretic distinctness. By generating queries that traversed from `:Parent` to `:Child` without the `COUNT(DISTINCT ?child)` modifier, the model inadvertently created a multiplicative join. Since both parents are linked to the same children, the query counted each child twice (once per parent path), artificially inflating the "Child Count" variable. This subsequently distorted calculations, leading to false validations where ineligible families were approved due to miscalculated thresholds.

The Cartesian Product Trap

Similarly, when aggregating income, the model frequently joined Income patterns and Child patterns in a single WHERE clause without sub-query separation. This caused the SPARQL engine to generate a Cartesian product of the two datasets, effectively multiplying every income record by every child record. This resulted in erroneous rejections where families were flagged with violations due to massive over-estimations of total family incomes.

Recursive Loops and Infinite Regression

A more catastrophic failure mode was observed in the Flash model's handling of bidirectional relationships. The Student Housing ontology defines inverse relationships (e.g., a `:Parent` has a `:Child`, and that `:Child` has a `:Parent`). In several runs, the model generated SPARQL property paths that traversed these links cyclically (e.g., `:hasParent :hasChild :hasParent ...`) without a terminating condition. This created infinite recursion loops during execution, causing the PySHACL validation engine to crash entirely (logged as "Python Kernel Crash" in Section 4.4.1). This demonstrates that the model treats graph traversal as a linguistic association task ("Parents are related to Children") rather than a directed graph walk, failing to anticipate the computational consequences of such cycles.

5.2 Syntax Hallucination and Language Bleed

While the Pro model's failures were primarily logical, the Flash model struggled to maintain the boundaries of the language itself. The experimental data reveals a phenomenon of "Language Bleed," where the model, optimized for high-throughput generalized text generation, conflated the syntax of semantically similar languages.

5.2.1 SQL Contamination

The most frequent syntax error was the appearance of illegitimate keywords, such as `FILTER NOT (...)`. This construct is valid in SQL (`WHERE NOT`) but invalid in SPARQL (which requires `FILTER (! ...)` or `FILTER NOT EXISTS`). This suggests that the model's

training data contains significantly more SQL examples than SPARQL, leading it to default to the more dominant syntax when the probability distribution for the next token is ambiguous.

5.2.2 Token-Level Hallucinations

The model also exhibited errors that reveal its nature as a token predictor rather than a parser. The insertion of a dot (.) after BIND statements violates SPARQL grammar (where dots act as triple delimiters).

Furthermore, the model frequently attempted to use dot notation (e.g., `?s.hasChild`) or complex property path slashes (e.g., `?s /:hasChild`) in contexts where explicit triple patterns were required. While property paths exist in SPARQL 1.1, the specific context in which such syntax was generated often mimicked Object-Oriented programming accessors rather than valid RDF graph traversal.

5.2.3 Namespace Invention

A distinct class of errors involved the hallucination of ontology definitions. Despite being provided with a fixed set of prefixes, the model occasionally invented new namespaces (e.g., using the deprecated 2007 SHACL draft URI or inventing an `ex:Citizen` ontology). This behavior was the primary cause of all recorded "RDF Syntax Errors" (as distinct from SPARQL errors), confirming that smaller models struggle to adhere to strict "Negative Constraints" (i.e., "Do not use any other prefix").

5.3 The Trade-off of Abstraction vs. Fidelity

A widely held assumption in Large Language Model research is that "Model Capability" (size, reasoning power) correlates linearly with performance across all tasks. However, the experimental results from the "Parental Leave" use case reveal a critical inversion of this principle.

5.3.1 The "Smart Model" Trap

The extraction of eligibility preconditions requires extreme fidelity to the source text. Legal constraints often rely on specific enumerations that define the scope of the law. Such a case was presented when models came across the following precondition in the Parental Leave use case: *"The applicant must be employed under a dependent employment regime, in the Private or Public sector."*

In this task, the Gemini 2.5 Flash model (theoretically less capable model) significantly outperformed the Gemini 2.5 Pro model. Flash, lacking the capacity for deep abstraction, tended to "copy-paste" the precondition literally. When presented with the employment requirement, it preserved the disjunction ("Private OR Public").

Pro, optimized for high-level reasoning and helpfulness, attempted to "summarize" the requirement. It interpreted "Private or Public" as a generic concept ("Employed"), effectively deleting the exclusion of other sectors (e.g., Freelancers). This finding suggests that for compliance tasks, "Smart" models may be fundamentally misaligned with the goal. Their training bias towards summarization and abstraction leads to Semantic Drift, where the gist of the rule is preserved but the legal boundary is lost.

5.3.2 The Deterministic Superiority

This trade-off extends to the architectural design of the pipeline itself. A stark contrast was observed between the error rates of the neural components and the symbolic components.

Notably, zero syntax errors were recorded in the generation of the "Citizen-Service Graph" (Stage 3). This stage relied on deterministic Python code to serialize the graph based on the LLM-derived Information Model, rather than asking the LLM to generate the Turtle syntax directly. The perfect stability of this stage, contrasted with the high failure rate of the LLM-generated SHACL shapes (Stage 4), empirically validates the architectural decision to offload structural tasks to deterministic code wherever possible.

This leads to a key design principle for Neuro-Symbolic systems in public administration: LLMs are necessary for interpretation (Extraction), but they are suboptimal for serialization (Code Generation). A robust pipeline must treat the LLM as a "Translator" of natural language, but never as an "Architect" of the final system structure.

5.4 The Complexity Ceiling

The divergence in performance between the "Student Housing" and "Parental Leave" use cases identifies a distinct Complexity Ceiling for current LLM-based logic generation. While the pipeline demonstrated high viability for administrative tasks involving categorical classification (Parental Leave), it experienced a near-total collapse when tasked with recursive arithmetic (Student Housing).

This failure mode correlates strongly with the Dependency Depth of the required logic:

- **Shallow Dependencies (Success):** Constraints that rely on single-node checks (e.g., *Nationality* depends only on *Applicant*) or flat Boolean logic (e.g., *isValid* is True OR False) were handled with high accuracy (88.9% logic success).
- **Deep Dependencies (Failure):** Constraints that require multi-hop traversal (e.g., *Applicant* → *Parent* → *Residence*) or recursive variable modification (e.g., *Income Limit* changes based on *Dependent Child Count*) consistently triggered the "Cartesian Product" bug or infinite recursion errors.

This suggests that while the tested LLMs can successfully act as "Semantic Parsers" for straightforward bureaucracy, they lack the internal "Working Memory" required to maintain the state of multi-variable equations throughout the code generation process.

5.5 The Contribution of Prompt Engineering

The experimental results challenge the prevailing narrative that "better prompting" is a universal solution to model limitations. Instead, the data reveals complex trade-offs where techniques that improve syntactic stability may inadvertently degrade logical reasoning.

5.5.1 The "Copy-Paste" Bias

The Default (Few-Shot) strategy proved essential for stabilizing syntax in the Pro model, boosting syntactic validity from 60% to 80% on the complex document. However, this stability came at a cost to logical accuracy. Zero-Shot strategy, despite producing broken code more often, achieved the highest logical accuracy (33.3%) when it *did* compile.

This suggests a "Copy-Paste Bias": when provided with examples, the model seems to over-fit to the logic of the few-shot template, attempting to force the new problem into the old structure. Without examples (Zero-Shot), the model is forced to reason from first principles, leading to messier syntax but potentially more original (and correct) logical derivations.

5.5.2 The Failure of Self-Correction

The Reflexion strategy failed to deliver the expected performance gains. For the less capable Flash model on the complex document, Reflexion actually *degraded* performance, dropping syntax validity from 20% to 5%. This indicates that a model incapable of solving a problem in the first pass is equally incapable of critiquing its own solution. Asking a confused model to "double-check" its work merely introduces a second opportunity for hallucination, compounding errors rather than resolving them.

5.5.3 The Engineering Ceiling

These findings imply an "Engineering Ceiling": one cannot prompt their way out of a fundamental reasoning deficit. While Prompt Engineering can guide a capable model (Pro) to follow syntactic rules, it cannot bestow reasoning capabilities upon a smaller model (Flash) that physically lacks them. For high-stakes logic generation, architectural scale remains the dominant variable.

5.6 Feasibility and Sovereignty

The final dimensions of analysis concern the operational viability of deploying such a system within a public administration context. The experimental campaign revealed critical vulnerabilities in the reliance on proprietary Model-as-a-Service (MaaS) infrastructure.

5.6.1 The Semantic Drift of "Eligibility"

Feasibility is first challenged at the point of ingestion. The extraction and summarization phase (Stage 1) demonstrated a persistent ambiguity in defining "Eligibility." Despite explicit prompt instructions to ignore administrative steps, the model frequently conflated procedural requirements (e.g., "Log in to TaxisNet") with substantive facts (e.g., "Be employed"). This semantic drift creates a system that validates paperwork rather than reality. While acceptable for a pilot, a production system would require a stricter, legally-grounded ontology of "Evidence" vs. "Conditions" to prevent the digitization of bureaucracy from becoming merely the automation of red tape.

5.6.2 Replication Crisis

The most severe threat to feasibility, however, emerged from the infrastructure itself. During the experimental window, unannounced changes to the Google Gemini API rate limits and model availability caused a sudden, catastrophic degradation in pipeline throughput. This event serves as a potent case study for Digital Sovereignty.

A public administration pipeline that relies on opaque, third-party endpoints is fundamentally fragile. The inability to guarantee consistent latency, availability, or even model behavior (version drift) renders MaaS solutions unsuitable for critical government infrastructure. The findings of this study strongly advocate for a shift towards Sovereign AI: deploying open-weights models (e.g., Llama 3, Mistral) on government-controlled infrastructure. Only by owning the compute can the administration guarantee the reproducibility and stability required for legal automation.

5.7 Risk Asymmetry in Public Administration

The evaluation of the system's "Recommender Accuracy" (Section 4.4.2) reveals a critical insight into the deployment readiness of these neuro-symbolic agents. While standard machine learning models optimize for balanced F1 scores, the operational context of public administration imposes an asymmetric cost of error.

The experimental data showed a 10.1% False Positive Rate, which corresponds to instances where the system erroneously recommended a service to an ineligible citizen. In a commercial context (e.g., movie recommendations), such errors are trivial. However, in digital governance, a False Positive actively generates bureaucratic friction. Specifically, it compels a citizen to gather documents and submit an application that is destined to fail. This wastes public resources and also erodes trust in the automated system. Conversely, the 3.7% False Negative Rate (Missed Opportunities), while statistically undesirable, represents a "safer" failure mode that maintains the status quo.

Consequently, the current pipeline's bias towards 'over-recommending' presents a

significant barrier to unsupervised deployment in a real-world administrative setting.

6 Limitations & Future Work

The development and evaluation of the proposed Neuro-Symbolic pipeline has highlighted several boundaries that currently exist between generative AI capabilities and the requirements of administrative automation. This chapter synthesizes both the technical and the conceptual constraints encountered during the pilot study with suggestions for future development. The goal of this last part is to define the necessary trajectory for moving from pilot-scale experimentation toward production-grade digital governance.

6.1 Infrastructure Dependencies & Digital Sovereignty

The experiments conducted in this study were significantly shaped by their reliance on proprietary, cloud-based "Model-as-a-Service" (MaaS) infrastructure. While the use of the Google Gemini ecosystem provided reasoning capabilities adequate for a pilot-scale prototype, it introduced a significant *infrastructural dependency* that reveals an important limitation for public sector applications.

6.1.1 Model and Resource Constraints

A primary limitation of this work is the narrow diversity of the models used. Due to the reliance on free-tier access, the study was restricted to just two models within a single vendor's ecosystem. This prevents a broader comparative analysis of how different architectural families (such as GPT or Llama) handle the specific nuances of administrative logic. Future research must expand to include a statistically significant variety of LLMs. This could be a way to determine if any of the inadequacies, such as the "Semantic Drift" observed in this study, are universal traits of generative AI or a specific characteristic of the utilized models.

6.1.2 Operational Fragility and the Replication Crisis

The reliance on external APIs introduced another roadblock. Throughout the implementation phase, unannounced updates to the underlying models and shifting rate-limiting policies resulted in regressions in throughput and occasionally also logic. This lack of control over the "versioned state" of the model created a significant barrier to reproducibility as well. In a digital governance context, such operational instability is unacceptable. A legal pipeline must be idempotent and resilient to the business decisions of third-party "Big Tech" providers.

6.1.3 Model Specialization

A significant limitation of relying on models from a "foreign" general-purpose ecosystem (like Gemini or GPT) is their lack of alignment with specific legislative drafting styles. Future research should prioritize *Domain-Specific Fine-Tuning*. By taking an open-weights local model and training it on a curated dataset of administrative texts, together with their formal corresponding logic, we can create a specialized LLM. This transition from few-shot (or zero-shot) prompting to a fine-tuned model is expected to significantly reduce the occurrence of observed phenomena such as the "Language Bleed" and more importantly improve the model's understanding of concepts like administrative hierarchy.

6.1.4 A Roadmap Toward Sovereign AI

To address these dependencies, the future roadmap for this research prioritizes the transition to *Digital Sovereignty*. By migrating the Neuro-Symbolic pipeline from cloud-based APIs to local, open-weights models hosted on private government infrastructure, several strategic goals can be achieved:

- **Data Privacy:** Full compliance with GDPR by ensuring citizen data never leaves a secure environment.
- **Operational Stability:** Eliminating the risk of API-related "logic shifts" or downtime.
- **Fine-tuning:** The ability to perform Domain-Specific Fine-tuning on actual legislative corpuses, which is often restricted or cost-prohibitive on proprietary cloud platforms.

The move toward local infrastructure is not merely a technical upgrade but a political and legal necessity for the "No-Stop Government" vision, ensuring that the "code" of the law remains under the exclusive control of the state.

6.2 Scalability

While the pilot study successfully demonstrated the technical feasibility of the Neuro-Symbolic pipeline, the scope of the experimental campaign was intentionally narrowed to prioritize analytical depth over statistical breadth. A clear path exists for scaling the framework in future research.

6.2.1 Document Corpus and Prompting Strategies

The primary limitation regarding generalizability is the limited sample size of the included document corpus. By focusing the evaluation on two administrative document sets, the study provided a high-resolution view of the pipeline's behavior in a complex scenario. However, to validate the framework's usefulness across the entire spectrum of public administration,

future work must scale the experimental setup to include hundreds of heterogeneous documents across different domains (e.g., healthcare, transportation, business licensing etc.).

Furthermore, the prompt engineering strategies employed were primarily focused on instruction following. Future research should investigate more sophisticated strategies, such as deeper Chain-of-Thought (CoT) prompting or more complex and specific "Self-Refine" loops, to determine if these techniques can improve the success rates identified in the Pilot Study.

6.2.2 From Pilot to Interoperability

A fundamental strength of this dissertation's methodology is the decision to ground the symbolic components in established European standards, turning the synthesized graphs from isolated artifacts to being inherently interoperable with the broader European semantic ecosystem. Currently, the pipeline utilizes a subset of these vocabularies. The next phase of development should involve:

- **Full Vocabulary Integration:** Expanding to cover the entirety of the CCCEV and CPSV-AP classes, allowing for the representation of the whole spectrum of the Public Service and its preconditions.
- **Cross-Border Interoperability:** Testing the pipeline on legislative texts from different EU member states to evaluate if the neuro-symbolic bridge can handle multilingual semantics.
- **Automated Ontology Mapping:** Implementing a dynamic mapping stage where the LLM can identify and utilize newly added classes from the official EU vocabularies without manual updates.

By leaning heavily on these standards, the framework serves as a modular component of the "Once-Only Principle", ensuring that once a rule is synthesized and validated, it can be shared and understood by any administrative system across the European Union.

6.3 Methodological Refinement

Beyond scaling the quantity of data, future iterations of this research must refine the qualitative evaluation of the pipeline's artifacts. The current study relied on functional execution success, but if we are to gain a deeper understanding of the "Neuro-Symbolic Gap", we require more granular metrics.

6.3.1 Root Cause Analysis

While our automated testing logs provide a baseline for failure rates, they cannot fully explain the linguistic nor the logical nuances that lead to a hallucinated constraint. A necessary next step is the introduction of a human expert at the role of auditor. According

to this idea, domain experts (possibly legal and administrative professionals) will review the specific cases where the pipeline failed either in syntax or logic, to categorize the errors. This will provide valuable insights such as if failures occur primarily during the precondition extraction (interpreting the text) or text-to-logic (generating the code). Understanding this divide is very important for pinpointing which stage of the pipeline requires more intensive engineering.

6.3.2 Artifact Evaluation

Currently, the evaluation of the pipeline artifacts is primarily functional. Future work should implement a method to evaluate them in depth.

- **Topology vs. Content:** Instead of comparing node names or descriptions, the analyzer will use metrics such as Graph Edit Distance (GED) to evaluate how different the LLM-generated structures are between different iterations (runs) using the same configuration.
- **Logic Similarity Metrics:** Future frameworks should explore similarity metrics for SPARQL as well, such as Tree Edit Distance on Abstract Syntax Trees (AST), to assess how logically different the generated queries are.
- **Natural Language Similarity:** The stability of natural language summaries can be assessed through *embeddings* to test for consistency.
- **Artifact Benchmarking:** Another approach to facilitate the same goal is for human experts to draft "perfect" reference graphs, summaries and SPARQL queries for a control set of documents. By comparing against these gold standards, we can measure semantic and structural drifts with mathematical precision, identifying patterns where the LLM consistently over-simplifies or over-complicates.

It is worth noting that since this work employs a meticulous persistence approach, keeping all artifacts on file, many of these ideas can be implemented already using the stored artifacts without needing to re-run experiments.

6.4 Trust-Centric Optimization

As discussed previously, in the context of public service eligibility the cost of a False Positive (incorrectly recommending a benefit) is often higher than the cost of a False Negative (missing an opportunity for plausible eligibility). Future iterations of the pipeline must be refined to prioritize *Precision* (Trustworthiness) over *Recall* (Coverage).

The prompts should be engineered and the intermediate representations should be calibrated to be "conservative by default". If the model encounters an ambiguous clause, it should be instructed to be strict or flag the constraint for human review rather than attempting a probabilistic "best guess".

Through this lens, the JSON Information Model can be expanded to include "Confidence

Attributes" where the model self-reports its certainty for each extracted precondition, allowing the symbolic stage to automatically reject any logic that falls below a specific trust threshold.

Bibliography

- [1] I. Konstantinidis, I. Magnisalis, and V. Peristeras, “A framework for a public service recommender system based on neuro-symbolic ai,” *Applied Sciences (Switzerland)*, vol. 15, no. 20, 2025. doi: 10.3390/app152011235
- [2] I. Oranekwu, L. Elluri, R. Yus, and A. Kotal, “Scalable automation for iot cybersecurity compliance: Ontology-driven reasoning for real-time assessment,” *Computers and Security*, vol. 161, 2026. doi: 10.1016/j.cose.2025.104711
- [3] A. Z. Spyropoulos and V. D. Tsiantos, “Interoperable semantic systems in public administration: Ai-driven data mining from law-enforcement reports,” *Computers*, vol. 14, no. 9, 2025. doi: 10.3390/computers14090376
- [4] M. G. Hanuragav and V. Gopinath, “Graph-driven validation of csr (tfls) using semantic technologies,” vol. 4085, 2025.
- [5] P. Agarwal, N. Kumar, and S. J. Bedathur, “Symkgqa: Few-shot knowledge graph question answering via symbolic program generation and execution,” vol. 1, 2024, pp. 10 119–10 140. doi: 10.18653/v1/2024.acl-long.545
- [6] C. V. S. Avila, V. M. P. Vidal, W. Franco, and M. A. Casanova, “Few-shot learning or rag in llm-based text-to-sparql? why not both?,” 2025, pp. 76–79. doi: 10.1109/ICSC64641.2025.00016
- [7] L. Jiang, J. Huang, C. Moller, and R. Usbeck, “Ontology-guided, hybrid prompt learning for generalization in knowledge graph question answering,” 2025, pp. 28–35. doi: 10.1109/ICSC64641.2025.00010
- [8] M. Shah et al., “Improving llm-based kgqa for multi-hop question answering with implicit reasoning in few-shot examples,” 2024, pp. 125–135.
- [9] S. Walter and H. Bast, “Grasp: Generic reasoning and sparql generation across knowledge graphs,” *Lecture Notes in Computer Science*, vol. 16140, pp. 271–289, 2026. doi: 10.1007/978-3-032-09527-5_15
- [10] A. Soularidis, K. I. Kotis, M. Lamolle, Z. Mejdoul, G. Lortal, and G. A. Vouros, “Llm-assisted generation of swrl rules from natural language,” 2024, pp. 7–12. doi: 10.1109/AIxDKE63520.2024.00008

- [11] J. F. Lehmann, P. Gattogi, D. Bhandiwad, S. Ferré, and S. Vahdati, “Language models as controlled natural language semantic parsers for knowledge graph question answering,” *Frontiers in Artificial Intelligence and Applications*, vol. 372, pp. 1348–1356, 2023. DOI: 10.3233/FAIA230411
- [12] L. Kovriguina, R. Teucher, D. Radyush, and D. I. Muromtsev, “Sparqlgen: One-shot prompt-based approach for sparql query generation,” vol. 3526, 2023.
- [13] M. Mountantonakis and Y. Tzitzikas, “Generating sparql queries over cidoc-crm using a two-stage ontology path patterns method in llm prompts,” *Journal on Computing and Cultural Heritage*, vol. 18, no. 1, 2025. DOI: 10.1145/3708326
- [14] J. G. Ongri, E. Tjitrahardja, F. Darari, and F. J. Ekaputra, “Towards an open nli llm-based system for kgs: A case study of wikidata,” 2024, pp. 44–49. DOI: 10.1109/ISRITI64779.2024.10963661
- [15] L. M. Vieira da Silva, A. Köcher, F. Gehlhoff, and A. Fay, “On the use of large language models to generate capability ontologies,” 2024. DOI: 10.1109/ETFA61755.2024.10710775
- [16] V. Emonet, J. T. Bolleman, S. Duvaud, T. M. de Farias, and A. C. Sima, “Llm-based sparql query generation from natural language over federated knowledge graphs,” vol. 3953, 2025.
- [17] N. Mashhaditafreshi, A. Textor, P. Rubel, N. Moarefvand, and A. Wagner, “Samm copilot: Bootstrapping semantic models with the eclipse semantic modeling framework from domain data in json using large language models,” vol. 4020, 2025, pp. 36–52.
- [18] J. F. Sequeda, D. Allemang, and B. Jacob, “Knowledge graphs as a source of trust for llm-powered enterprise question answering,” *Journal of Web Semantics*, vol. 85, 2025. DOI: 10.1016/j.websem.2024.100858
- [19] D. Allemang and J. F. Sequeda, “Increasing the accuracy of llm question-answering systems with ontologies,” *Lecture Notes in Computer Science*, vol. 15233, pp. 324–339, 2025. DOI: 10.1007/978-3-031-77847-6_18
- [20] A. Gashkov, M. Eltsova, A. Perevalov, and A. Both, “Instruction-tuned language models as judges for sparql query correctness in knowledge graph question answering,” vol. 4020, 2025, pp. 177–194.
- [21] D. Adam and T. Kliegr, “Traceable llm-based validation of statements in knowledge graphs,” *Information Processing and Management*, vol. 62, no. 4, 2025. DOI: 10.1016/j.ipm.2025.104128

- [22] L. P. Meyer et al., “Llm-kg-bench 3.0: A compass for semantic technology capabilities in the ocean of llms,” *Lecture Notes in Computer Science*, vol. 15719, pp. 280–296, 2025. DOI: 10.1007/978-3-031-94578-6_16
- [23] C. Kosten, F. Nooralahzadeh, and K. Stockinger, “Evaluating the effectiveness of prompt engineering for knowledge graph question answering,” *Frontiers in Artificial Intelligence*, vol. 7, 2024. DOI: 10.3389/frai.2024.1454258
- [24] D. M. Schmidt, R. Schubert, and P. Cimiano, “Compost: A benchmark for analyzing the ability of llms to compositionally interpret questions in a qald setting,” *Lecture Notes in Computer Science*, vol. 16140, pp. 3–22, 2026. DOI: 10.1007/978-3-032-09527-5_1
- [25] N. Tufek et al., “Validating semantic artifacts with large language models,” *Lecture Notes in Computer Science*, vol. 15344, pp. 92–101, 2025. DOI: 10.1007/978-3-031-78952-6_9
- [26] C. Kosten, P. Cudré-Mauroux, and K. Stockinger, “Spider4sparql: A complex benchmark for evaluating knowledge graph question answering systems,” in *2023 IEEE International Conference on Big Data (BigData)*, IEEE, Dec. 2023, pp. 5272–5281. DOI: 10.1109/bigdata59044.2023.10386182 [Online]. Available: <http://dx.doi.org/10.1109/BigData59044.2023.10386182>
- [27] Z. Lin, S. Trivedi, and J. Sun, “Generating with confidence: Uncertainty quantification for black-box large language models,” *Trans. Mach. Learn. Res.*, vol. 2024, 2023. DOI: 10.48550/arxiv.2305.19187
- [28] E. Stani, F. Barthélemy, K. Raes, M. Pittomvils, and M. A. Rodriguez, “How data vocabulary standards enhance the exchange of information exposed through apis: The case of public service descriptions,” *Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance*, 2020. DOI: 10.1145/3428502.3428626
- [29] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, “Unifying large language models and knowledge graphs: A roadmap,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 7, pp. 3580–3599, Jul. 2024, ISSN: 2326-3865. DOI: 10.1109/tkde.2024.3352100 [Online]. Available: <http://dx.doi.org/10.1109/TKDE.2024.3352100>
- [30] E. Nuyts, M. Bonduel, and R. Verstraeten, “Comparative analysis of approaches for automated compliance checking of construction data,” *Adv. Eng. Informatics*, vol. 60, p. 102443, 2024. DOI: 10.1016/j.aei.2024.102443
- [31] N. Ferranti, J. F. de Souza, S. Ahmetaj, and A. Polleres, “Formalizing and validating wikidata’s property constraints using shacl and sparql,” *Semantic Web*, 2024. DOI: 10.3233/sw-243611

- [32] P. Pareti and G. Konstantinidis, “A review of shacl: From data validation to schema reasoning for rdf graphs,” pp. 115–144, 2021. DOI: 10.1007/978-3-030-95481-9_6
- [33] P. Hagedorn, P. Pauwels, and M. König, “Semantic rule checking of cross-domain building data in information containers for linked document delivery using the shapes constraint language,” *Automation in Construction*, 2023. DOI: 10.1016/j.autcon.2023.105106
- [34] S. Ouyang, J. Zhang, M. Harman, and M. Wang, “An empirical study of the non-determinism of chatgpt in code generation,” *ACM Transactions on Software Engineering and Methodology*, vol. 34, pp. 1–28, 2023. DOI: 10.1145/3697010
- [35] M. Schäfer, S. Nadi, A. Eghbali, and F. Tip, “An empirical evaluation of using large language models for automated unit test generation,” *IEEE Transactions on Software Engineering*, vol. 50, pp. 85–105, 2023. DOI: 10.1109/tse.2023.3334955
- [36] B. Donato, L. Mariani, D. Micucci, and O. Riganelli, “Studying how configurations impact code generation in llms: The case of chatgpt,” *2025 IEEE/ACM 33rd International Conference on Program Comprehension (ICPC)*, pp. 442–453, 2025. DOI: 10.1109/icpc66645.2025.00055
- [37] M. Papadakis, M. Kintis, J. M. Zhang, Y. Jia, Y. L. Traon, and M. Harman, “Chapter six - mutation testing advances: An analysis and survey,” *Adv. Comput.*, vol. 112, pp. 275–378, 2019. DOI: 10.1016/bs.adcom.2018.03.015
- [38] F. Tip, J. Bell, and M. Schäfer, “Llmorpheus: Mutation testing using large language models,” *IEEE Transactions on Software Engineering*, vol. 51, pp. 1645–1665, 2024. DOI: 10.1109/tse.2025.3562025
- [39] A. Gupta, M. Lu, K. Zhu, S. O’Brien, and V. Sharma, *Novelhopqa: Diagnosing multi-hop reasoning failures in long narrative contexts*, 2025. arXiv: 2506.02000 [cs.CL]. [Online]. Available: <https://arxiv.org/abs/2506.02000>
- [40] Y.-C. Chen et al., “Enhancing responses from large language models with role-playing prompts: A comparative study on answering frequently asked questions about total knee arthroplasty,” *BMC Medical Informatics and Decision Making*, vol. 25, 2025. DOI: 10.1186/s12911-025-03024-5
- [41] L. Wang et al., “Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models,” pp. 2609–2634, 2023. DOI: 10.48550/arxiv.2305.04091
- [42] Z. Chu, Y. Wang, L. Li, Z. Wang, Z. Qin, and K. Ren, “A causal explainable guardrails for large language models,” *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 2024. DOI: 10.1145/3658644.3690217

- [43] S. Sivarajkumar, M. Kelley, A. Samolyk-Mazzanti, S. Visweswaran, and Y. Wang, “An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: Algorithm development and validation study,” *JMIR Medical Informatics*, vol. 12, 2024. DOI: 10.2196/55318
- [44] S. Kresevic, M. Giuffré, M. Ajčević, A. Accardo, L. Crocè, and D. Shung, “Optimization of hepatological clinical guidelines interpretation by large language models: A retrieval augmented generation-based framework,” *NPJ Digital Medicine*, vol. 7, 2024. DOI: 10.1038/s41746-024-01091-y
- [45] L. Liu et al., “Instruct-of-reflection: Enhancing large language models iterative reflection capabilities via dynamic-meta instruction,” pp. 9956–9978, 2025. DOI: 10.48550/arxiv.2503.00902
- [46] M. Chen et al., *Evaluating large language models trained on code*, 2021. arXiv: 2107.03374 [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2107.03374>

A Appendix placeholder

Extracts of the generated SHACL shapes (both valid and broken examples).

Samples of the YAML Mutation Scenarios.

Samples of RDFS Ontologies used.