

# Exploring Neuro-symbolic Pipelines for Structured Knowledge Extraction

Nikolaos Laoutaris

December 10, 2025

## **Abstract**

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

# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
1.1	Background and Motivation . . . . .	2
1.2	Problem Statement . . . . .	2
1.3	Objectives . . . . .	2
1.4	Dissertation Structure . . . . .	2
<b>2</b>	<b>Systematic Literature Review</b>	<b>3</b>
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 . . . . .	5
2.3.2	Data Extraction . . . . .	5
2.4	Thematic Analysis . . . . .	7
2.4.1	From Text to Structured Models . . . . .	7
2.4.2	Automated Logic Generation . . . . .	7
2.4.3	Validation and Hallucination Control . . . . .	7
2.5	Discussion and Research Gap . . . . .	7
<b>3</b>	<b>Pilot Study</b>	<b>8</b>
3.1	Overview . . . . .	8
3.2	Methodology and System Architecture . . . . .	8
3.2.1	Setup Environment . . . . .	8
3.2.2	Semantic Data Modelling . . . . .	9
3.2.3	The Extraction and Generation Pipeline . . . . .	10
3.2.4	The Validation Engine . . . . .	12
3.3	Experimental Design . . . . .	12
3.3.1	The Mutation Testing Framework . . . . .	12
3.3.2	Experimental Configurations . . . . .	12
3.3.3	Evaluation Metrics . . . . .	12

<b>4</b>	<b>Results</b>	<b>13</b>
4.1	Syntax Validity . . . . .	13
4.2	Logic Validity . . . . .	13
4.3	Pipeline Reliability . . . . .	13
4.4	Semantic Stability (nice-to-have) . . . . .	13
<b>5</b>	<b>Discussion</b>	<b>14</b>
5.1	The Logic Traps . . . . .	14
5.2	The Syntactic Failures . . . . .	14
5.3	Domain Influence . . . . .	14
5.4	Prompt Engineering . . . . .	14
5.5	Model Size Trade-offs . . . . .	14
5.6	Replication Crisis . . . . .	14
<b>6</b>	<b>Limitations &amp; Future Work</b>	<b>15</b>
6.1	Limitations . . . . .	15
6.2	Future Research Directions . . . . .	15
<b>A</b>	<b>Appendix placeholder</b>	<b>16</b>

# 1 Introduction

## 1.1 Background and Motivation

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

## 1.2 Problem Statement

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

## 1.3 Objectives

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

## 1.4 Dissertation Structure

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

# 2 Systematic Literature Review

## 2.1 Introduction

This chapter details the Systematic Literature Review (SLR) conducted to establish the theoretical foundations of Neuro-Symbolic AI. We approach the current state of research in Neuro-Symbolic AI, specifically focusing on how Large Language Models (LLMs) and Knowledge Graphs (KGs) are combined. We aim to identify existing approaches for extracting rules from text and generating formal logic (SPARQL/SHACL), as well as methods of evaluating the results of such a process.

## 2.2 Methodology

To ensure scientific rigor and reproducibility, the review adheres to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. The process was structured into four distinct 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 titles and abstracts, followed by a rigorous quality assessment of full texts. This phase utilized specific inclusion/exclusion criteria. Finally, data was extracted from the selected primary studies into a standardized matrix to synthesize key themes, directly related to the RQs.

### 2.2.1 Research Questions

To achieve our objective, we defined three specific Research Questions (RQs) that guide 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 critically analyze how existing studies ensure 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**. A search query was constructed 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/Language:** ("SHACL" OR "SPARQL")
3. **Symbolic Context:** ("Semantic Web" OR "Knowledge Graph")

These blocks were combined using the AND operator. 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 applied some metadata filters during this phase:

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

Interestingly, despite the Date Range not being restricted, all results fell in the range of **2023–2026**. This could be explained by the fact that the application of Large Language Models to formal constraint languages (like SHACL) is a nascent field that emerged primarily after the widespread adoption of GPT-4 class models.

The described search strategy yielded an initial set of candidates which were then subjected to the screening process described in the following section.

## 2.2.3 Inclusion/Exclusion Criteria

Next, 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. Table 2.1 summarizes the criteria used.

Of the papers sought, some could not be retrieved due to access restrictions (paywall). The remaining ones were downloaded assessed for eligibility by reading the full text. In this "Quality Screening" phase, we applied a second set of quality exclusion criteria (QE):

- **QE1 (Name):** We ?
- idea: Specify "Schema-aligned extraction" or "Constraint/Rule extraction". This ensures we are looking for papers that deal with complexity (like eligibility rules), not just connectivity.
- idea: No OpenIE. (same as above?)

The next section summarizes the results following this quality assessment.

**Table 2.1:** Inclusion and Exclusion Criteria

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

## 2.3 Results

From an initial set of 125 records, 14 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).

**Picture 2.1:** PRISMA Flow Diagram of the selection process.

### 2.3.2 Data Extraction

Table 2.2 presents the data extraction summary for the 14 included studies. The studies are categorized by their primary theme: (1) Domain-Specific Pipelines, (2) Automated Logic Generation, (3) Validation Frameworks, and (4) Retrieval (GraphRAG).

**Table 2.2:** Summary of Included Studies (Data Extraction)

Study	Domain / Input	Task	Target Logic	Validation Method
<i>Category 1: Domain-Specific Neuro-Symbolic Pipelines</i>				
Konstantinidis (2025) <b>Konstantinidis2025</b>	Public Service Regulations	Framework Proposal	RDF + SHACL	Conceptual Prototype (No regression testing)
Hanuragav (2025) <b>Hanuragav2025</b>	Clinical Study Reports	Compliance Check	SHACL + SPARQL	Deterministic Rule Execution
Oranekwu (2026) <b>Oranekwu2026</b>	IoT Security (NIST)	Compliance Check	Ontology SWRL	Ontology-driven Reasoning
Spyropoulos (2025) <b>Spyropoulos2025</b>	Police Reports	Text-to-Graph	RDF Triples	Human-in-the-loop Verification
<i>Category 2: Automated Logic Generation (Text-to-Logic)</i>				
Walter (2026) <b>Walter2026271</b>	General (Wiki-data)	Text-to-SPARQL	SPARQL	Execution Accuracy (Zero-shot)
Souliaridis (2024) <b>Souliaridis2024</b>	NL Rules	Text-to-SWRL	SWRL	LLM-assisted Generation
Jiang (2025) <b>Jiang202528</b>	Scholarly QA	Text-to-SPARQL	SPARQL	Ontology-Guided Prompting
Mashhaditafreshi (2025) <b>Mashhaditafreshi202536</b>	JSON Data	Modeling	SHACL Shapes	Human Evaluation of Models
Avila (2025) <b>Avila2025223</b>	General QA	Text-to-SPARQL	SPARQL	Benchmark Execution (Auto-KGQA)
<i>Category 3: Validation &amp; Hallucination Control</i>				
Perevalov (2025) <b>Perevalov2025563</b>	Multilingual QA	Query Filtering	SPARQL	LLM-based Probabilistic Filtering
Gashkov (2025) <b>Gashkov2025177</b>	QA Systems	Query Judging	SPARQL	LLM-as-a-Judge
Tufek (2025) <b>Tufek202592</b>	Industrial Standards	Requirement Translation	SPARQL	F1 Score on Logic Translation
<i>Category 4: Retrieval Frameworks (GraphRAG)</i>				
Ongiris (2025) <b>Ongiris2025116</b>	General (Wiki-data)	GraphRAG	SPARQL	Jaccard Similarity
Ahmed Khan (2026) <b>AhmedKhan2026</b>	Data Center Telemetry	Text-to-Query	SPARQL	Execution Accuracy vs NoSQL

## **2.4 Thematic Analysis**

### **2.4.1 From Text to Structured Models**

Current research demonstrates that LLMs are highly effective at the initial 'extraction' phase, successfully mapping unstructured text into RDF or SHACL skeletons.

### **2.4.2 Automated Logic Generation**

Several studies focus on translating natural language directly into query languages. Walter et al. achieved state-of-the-art results in zero-shot SPARQL generation, while Soularidis et al. explored generating SWRL rules. However, these approaches often struggle with complex, nested logic without guidance.

### **2.4.3 Validation and Hallucination Control**

A critical challenge is ensuring the generated logic is correct. Perevalov et al. propose using an LLM to 'judge' or filter the SPARQL queries. In contrast, Tufek et al. use F1 scores against a gold standard. Crucially, most existing validation methods are probabilistic (LLM-based) rather than deterministic.

## **2.5 Discussion and Research Gap**

While the literature shows success in extraction (2.4.1) and generation (2.4.2), there is a gap in deterministic validation. Papers like Konstantinidis propose the theoretical framework for public services, and Hanuragav applies similar logic to clinical reports. However, no study has yet implemented a comprehensive Mutation Testing framework to rigorously test the structural stability of LLM-generated SHACL shapes for public service eligibility. This dissertation aims to fill that gap.

# 3 Pilot Study

## 3.1 Overview

Present the end-to-end pipeline design in high-level, detailing the data flow from document ingestion to the final eligibility decision and the mutation framework. Flow chart here? Unlike "end-to-end" approaches that attempt to generate code directly from text, this pipeline utilizes a chain of intermediate structured representations to stabilize the Large Language Model's reasoning process.

## 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 local processing for semantic graph operations and cloud-based APIs for Large Language Model inference.

### System Architecture

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

1. **The Core Module Library:** Contains the reusable logic, such as API communication, graph operations, parsing and testing utilities. It also contains the *pipeline core*, which encapsulates the end-to-end extraction-generation workflow.
2. **The Orchestration Layer (The "Cockpit"):** 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 batch processing.
3. **The Persistence Layer:** To ensure auditability and reproducibility, the system employs a strict "Artifact Preservation" strategy. Every experimental run generates a dedicated directory locally, containing all intermediate outputs of the core pipeline. Testing metrics and metadata are saved in a Master CSV file for post-hoc analysis.

## Technologies and Libraries

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

- **RDFLib:** Used for parsing, manipulating and serializing RDF graphs (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.
- **Pandas:** Used for the post-hoc aggregation and statistical analysis of the testing logs.

### 3.2.2 Semantic Data Modelling

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 a 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:** Connected to Constraint nodes via `cccev:constrains`, these nodes 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. By anchoring the pipeline's output in the CPSV-AP and CCCEV ecosystems, the generated graphs are natively 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

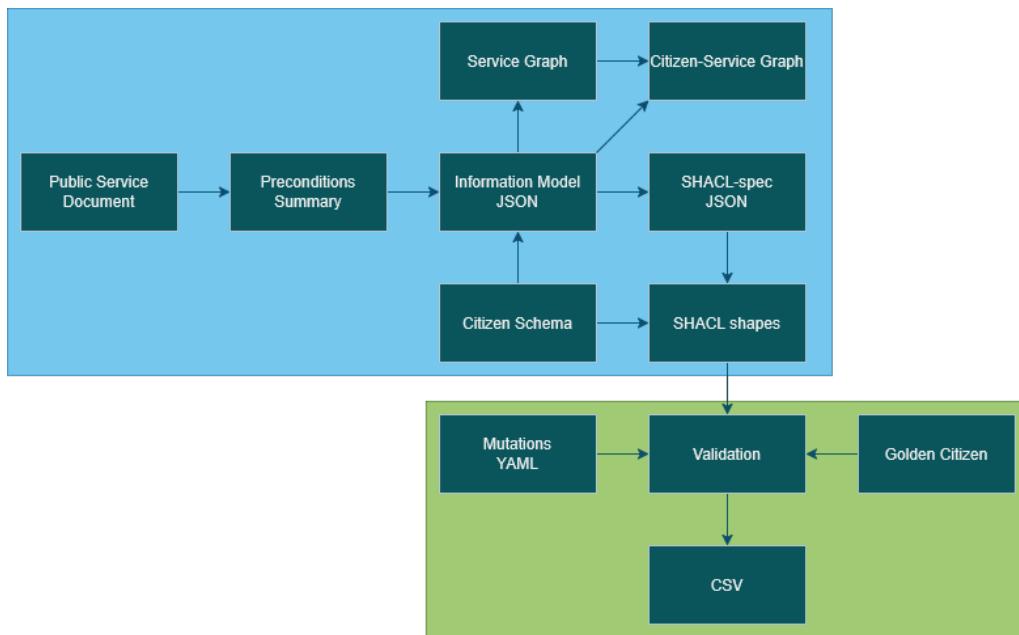
granular instead of aggregate data as nodes (e.g. prefer "Date of Birth" rather than "Age") and is encouraged to use abstract and reusable classes.

It has been demonstrated that the generation of such schemas can be automated as part of the pipeline (cite Konstantinidis). 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 relies on injecting specific faults into the citizen graph (e.g., modifying property values to trigger violations). This requires a deterministic, known-in-advance schema structure; had the schema been generated dynamically during each run, it would be impossible to define a static 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.



**Picture 3.1:** Flow Chart of the core pipeline.

#### Stage 1: Document Summarization and Precondition Extraction

The pipeline begins with the ingestion of the raw public service document (PDF). Using a Large Language Model (LLM), 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. Summarization reduces the cognitive load required for the 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 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 rules 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).

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 CPSV/CCCEV triples in the subsequent stage, while ensuring that all data references are grounded in the controlled vocabulary of the Citizen Schema.

## 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 (Explainability Layer):** 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 legal requirement.

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.

#### **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.

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.

The LLM is then invoked to translate this specification into RDF triples (Turtle format). For atomic constraints involving single-hop properties and literal comparisons (e.g., `Citizenship = 'GR'`), the model generates standard `sh:property` shapes. For requirements involving arithmetic, aggregations, date comparisons, or cross-referenced data (e.g., `now() - birthDate > 18`), the model encapsulates the logic within `sh:sparql` constraints. This allows for the expression of complex conditional logic that exceeds the expressivity of the SHACL Core vocabulary. 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.

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 following section.

#### **3.2.4 The Validation Engine**

### **3.3 Experimental Design**

#### **3.3.1 The Mutation Testing Framework**

Describe the methodology of creating "Golden Citizens" and the YAML-based mutation engine used to systematically generate failing edge cases.

#### **3.3.2 Experimental Configurations**

Justify the selection of the specific documents, models, prompting strategies.

#### **3.3.3 Evaluation Metrics**

Describe what kind of data was collected from the experiments. Define the specific quantitative metrics used, justify choosing them.

# 4 Results

Here we will only give facts, numbers, graphs. No explanations and opinions.

## 4.1 Syntax Validity

Present statistical data on how often the pipeline generated syntactically valid RDF and SPARQL code under different configurations.

## 4.2 Logic Validity

Report the scores and figures for the pipeline logic accuracy, maybe highlighting the discrepancy between "Syntax Success" and "Logic Success." Should further elaborate on failed cases.

## 4.3 Pipeline Reliability

An aggregation of overall pipeline feasibility combining syntax and logic.

## 4.4 Semantic Stability (nice-to-have)

If we have time to do it, this chapter will use data from the semantic similarity metrics drawn from past run artifacts on file.

# 5 Discussion

Here we will interpret the results and give subjective opinions and other observations.

## 5.1 The Logic Traps

Interpret the recurring logical failures in logically sound SPARQL generation, discussing the LLM's struggle with aggregations, joins, recurrency and closed loops.

## 5.2 The Syntactic Failures

Interpret the models' usage of wrong syntax, using other languages (perhaps more training data) despite being explicitly told to avoid it.

## 5.3 Domain Influence

Discuss the impact of complex vs simple domains or documents, semantic web wise.

## 5.4 Prompt Engineering

Discuss the differences the tested prompting techniques made (or didn't make).

## 5.5 Model Size Trade-offs

Discusses the counter-intuitive finding where the smaller model outperformed the larger model in literal constraint extraction and summarization (where else?)

## 5.6 Replication Crisis

Critique the reliance on proprietary Model-as-a-Service infrastructure, arguing that operational instability renders them unsuitable for critical pipelines. Maybe this can be in the Limitations chapter.

# 6 Limitations & Future Work

## 6.1 Limitations

Acknowledge the limited document sample size, API restrictions and the reliance on a specific vendor's ecosystem, limited prompt engineering, limited models (free tier). Maybe critique the reliance on proprietary infrastructure/software/models.

## 6.2 Future Research Directions

Propose a roadmap for using local open-source models to ensure sovereignty, the implementation of iterative "Self-Correction" agents to fix syntax errors, ideas for more robust testing of this kind of pipeline, ideas not implemented by this work for scoping reasons.

The decision to base the schema on an existing vocabulary was with good reason. This design allows the graphs generated by the pipeline to include more classes of the used ontologies, for future integration with more sophisticated systems. The pipeline itself could also be expanded upon to include more classes.

# 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.