

From Rules as Executed to Citizen Explanation

Evaluating LLM-Generated Explanations in Rule-Based Government Systems

Flora Magdalena Johanna BOS ^{a,b,1}, Anne SCHUTH ^{a,2} and Jieying CHEN ^b

^a*Dutch Digital Service, Ministry of Economic Affairs and Climate*

^b*Vrije Universiteit Amsterdam*

Abstract. Governments increasingly use automated decision-making systems to apply legislation in domains such as social benefits, taxation, and licensing. While Rules as Code (RaC) approaches enable legally consistent and transparent decision-making through executable legislation, they do not automatically generate explanations that are understandable for citizens. Large language models (LLMs) can support natural-language explanation generation, but may hallucinate facts or exclude legally relevant information. This paper investigates how Graph-based LLM architectures can generate citizen-focused explanations from deterministic rule-engine traces. A rule engine evaluates synthetic citizen profiles against formalised Dutch legislation and produces execution traces, which are transformed into structured decision representations for explanation generation. The approach is evaluated using automated metrics and human annotation. Results show that explicitly exposing decisive legal conditions improves faithfulness, and that the full Graph configuration, which combines a structured decision graph with a constrained prompt, further improves alignment with the rule-execution trace. Human evaluation shows that explanations perceived as understandable by citizens are not always considered legally correct by experts. The findings suggest that structured representations of deterministic legal reasoning can support legally faithful and citizen-focused explanations of automated government decisions. The source code is published.

Keywords. Rules as Code, Large Language Models, Explainable AI

1. Introduction

Governments increasingly use automated decision-making systems to apply legislation in domains such as taxes, social benefits, and licensing. Such systems support consistent, at-scale decision-making, but they also increase the need for explanations that citizens can understand. Rules as Code (RaC) approaches translate legislation into machine-executable representations, enabling reproducible decision processes [1,2]. We use the term Rules as Executed for the executable rule representations that actually produce gov-

¹Work done while being an intern at the Ministry of the Interior and Kingdom Relations.

²Work done while at the Ministry of the Interior and Kingdom Relations.

ernment decisions [3]. While RaC systems can support legally consistent outcomes, they do not automatically produce explanations that are accessible to non-expert citizens.

Large language models (LLMs) offer strong natural-language generation capabilities and therefore appear promising for producing citizen-friendly explanations. However, LLMs may hallucinate facts, exclude legally relevant conditions, or generate explanations that appear plausible without accurately reflecting the underlying decision process [4,5]. In legal and governmental contexts, these limitations are particularly problematic, since explanations must remain faithful to the actual reasoning behind a decision.

Retrieval-Augmented Generation (RAG) aims to improve factual reliability by incorporating external information during generation [6], and recent Graph-based and knowledge-augmented approaches extend this principle [7,8]. Within the AI&Law community, Nguyen and Satoh’s ConsRAG similarly uses aspect-based constraints in retrieval and generation, with iterative backtracking, to reduce legal hallucination [9], and Wehnert et al. combine legal knowledge graphs with textual retrieval for legal norm retrieval [10]. However, such architectures primarily address outputs grounded in textual documents or knowledge graphs, rather than in deterministic rule-execution traces produced by government decision systems, where the complete set of legally relevant facts is already available, and the challenge is one of structuring and communicating executable reasoning rather than retrieving evidence.

At the same time, explainable AI (XAI) research emphasises that explanations must be understandable to affected stakeholders, not only technically correct [11]. Within AI&Law, Zuurmond et al. demonstrate a graph-database, question-driven approach to explaining rule-based legal decision systems [12], Nouwens et al. apply case-based reasoning to decision support in Dutch administrative law [13], and Bouche-Pillon et al. bridge rule formalisation with justification-oriented reasoning for law enforcement data processing [14]. Citizen-focused legal explanations therefore involve a trade-off: they must remain aligned with the underlying legal decision process while being accessible to non-expert users.

This paper investigates how graph-structured representations of rule-engine execution traces can support LLM-generated, citizen-focused explanations of automated government decisions. The rule engine remains solely responsible for deterministic legal decision-making, the language model generates explanations from a structured representation of the decision trace rather than from open-ended legal text. In this paper, we ask the following research question: how can graph-structured representations of rule-based government decision processes be integrated with LLMs to generate citizen-focused explanations that are legally grounded and supported by correct evidence?

We develop a graph-based explanation pipeline for three Dutch legal domains: the Health Care Benefit Law [15], the Social Assistance Law [16], and the Alcohol Law [17]. To answer this research question, generated explanations are evaluated along two dimensions: legal faithfulness to the execution trace and citizen-focused comprehensibility, combining automated metrics with human validation by legal experts and citizens across five LLMs and three regulations.

This paper contributes (i) a graph-based architecture that integrates deterministic RaC execution traces with LLM-based explanation generation, (ii) an evaluation framework that verifies explanations against executable legal reasoning traces rather than only textual reference sources, and (iii) empirical evidence, across five LLMs and three regulations, of a faithfulness and accessibility trade-off relevant to the deployment of AI-

generated explanations in public administration. The source code for experiments in this paper is published.³

2. Related Work

2.1. Rules as Code and executable legislation

RaC approaches translate legislation into machine-readable formats, enabling deterministic rule execution [1,2]. Work on formal representation of norm interpretation [18] and on deriving executable services from legal sources [19] underlies the Dutch executable-legislation tradition our rule engine builds on. A recent systematic assessment of RaC languages provides a framework for situating such systems [20]. Normware approaches conceptualise computational artefacts that explicitly encode norms and governance constraints, treating explainability as an architectural property [21]. Machine-readable legislation must remain transparent and contestable, since public trust depends on citizens understanding the rules applied to them [2]. Rule-based systems provide deterministic, verifiable decision-making [22,2], but do not inherently support natural-language explanation generation, creating a gap between decision logic and citizen communication [1,23].

2.2. From rule execution to explanation generation

Computational law research shows that legal texts can be transformed into executable logic while preserving traceability [22]. Structured NLP combined with formal logic supports rule extraction and explainability [23,24]. Explanation faithfulness, however, is not guaranteed by self-consistency between predictions and explanations [25], and remains one of the central, insufficiently defined properties in NLP explainability research [26]. Within AI&Law, Prakken and Ratsma’s case-based argumentation model for explaining classifier decisions [27] and Zuurmond et al.’s human-centred, question-driven explanation method [12] are the closest published to the architecture proposed here. Górski and Ramakrishna evaluate XAI methods (e.g., LIME, SHAP) from the perspective of practising lawyers, underscoring that technical explainability does not equal usability for legal professionals [28]. Human-focused XAI research more broadly emphasises that explanations should support stakeholder understanding rather than only technical correctness [11,29].

2.3. LLM-generated legal explanations and faithfulness evaluation

Recent AI&Law work directly investigates LLMs for legal explanation and judgment. Jiang and Yang propose legal syllogism prompting for interpretable legal judgment prediction [30], and Blair-Stanek et al. investigate the limits of LLM statutory reasoning [31]. Most relevant to our contribution, Lindsay et al. propose argumentative LLMs that generate legal-entailment decisions together with explanations explicitly designed to be faithful and contestable [32]. On the evaluation side, Gultekin et al. propose a two-dimensional (truthfulness and reasoning-soundness) LLM-as-judge framework for legal question answering [33], and Sanchi and Novotná combine automatic metrics with

³Available at <https://github.com/Bosfloortje/Thesis-Rule-to-Citizen-Explanation>

human-grounded evaluation for LLMs in the legal domain [34]. Blair-Stanek and Van Durme further document output instability in LLM legal question answering [35]. More broadly, NLI-based methods outperform QA-, embedding-, and n-gram-based factual consistency metrics [36,37], atomic claim decomposition improves fine-grained factual verification [38], and efficient models such as MiniCheck approach larger-model hallucination-detection performance [39]. Existing metrics, however, assume textual reference sources rather than deterministic computational traces [36,40], motivating the adapted, trace-based evaluation framework used in this paper.

2.4. Knowledge representation and public-sector deployment

Structured knowledge representations improve traceability and factual grounding when combined with LLMs [41,8], building on foundational legal ontology and rule representation work such as LegalRuleML [42,43] and LKIF-Core [44]. Beyond technical faithfulness, deploying such systems in government raises questions of algorithmic transparency and citizen trust. Grimmelikhuijsen shows experimentally that transparency and explainability of automated decision-making increase perceived trustworthiness [45], a finding directly relevant to citizen-focused government explanations, while the Dutch *toeslagenaffaire* childcare-benefits scandal illustrates the societal cost of opaque algorithmic risk profiling [46]. These developments motivate evaluating not only whether LLM explanations are legally faithful, but whether structured, trace-grounded generation can reconcile faithfulness with citizen comprehensibility, the gap this paper addresses.

3. Dataset

Existing factual-consistency and legal NLP datasets do not contain executable decision traces [38,24], and administrative datasets (e.g., CBS microdata, EU-SILC) lack law-specific eligibility variables. We therefore construct a dedicated synthetic dataset of Dutch citizen profiles, using aggregated CBS statistics [47] to guide realistic distributions of age, income, and household composition while avoiding identifiable personal data. Profiles are generated through deterministic balancing across demographic and socio-economic categories (gender, age group, cultural background, work status, income level, household composition), combined with controlled randomisation within categories to avoid repetitive profiles. Each profile is evaluated by the rule engine to produce a legal outcome and execution trace against the formalised Health Care Benefit Law [15], Social Assistance Law [16], and Alcohol Law [17], which together form the input/output structures used throughout the evaluation. The dataset contains 200 profiles, each evaluated under all three laws; with five models and three prompting conditions this yields 9,000 explanations (3,000 per condition), plus the additional Llama 3.1:70B runs.

4. Methodology

4.1. System overview

The pipeline (Figure 1) has three stages: (1) decision generation, a deterministic rule engine evaluates each citizen profile against YAML-formalised legislation, (2) trace extrac-

tion, a structured trace records the outcome, decisive conditions, calculations, and legal references, (3) explanation generation, the trace is transformed into a structured decision representation passed to an LLM.

Three prompting conditions are compared, differing only in how much of the trace is exposed to the LLM and whether it is structured. In the Open baseline, the LLM receives the complete, unfiltered rule-engine output together with the raw citizen profile in YAML. No fact or condition is flagged as decisive, so the model itself has to identify what is legally relevant within the trace and the profile. In the Flat condition, the LLM instead receives only the pre-extracted decisive conditions, the citizen facts each condition is evaluated against, and the calculated amount, presented as an unordered list of plain-text statements. This is a subset of the information given to the Graph condition, presented without explicit relations between individual facts, conditions, and the outcome. In the Graph condition, this same information is instead encoded as a directed graph with five typed node categories: Decision (the final outcome or awarded amount), Rule (an individual legal condition evaluated by the engine), Fact (a citizen-specific data point), Calculation (an intermediate computed value, if applicable), and Law (the statutory source), connected by typed edges: HAS_FACT (the citizen fact it is evaluated against), SATISFIED / FAILED (whether that condition was met by the given facts), and USED_IN (the calculation or decision it feeds into). This graph is passed to the LLM with an explicit pass/fail status for every evaluated condition, so that the legally relevant facts and their relation to the decision are made explicit, rather than left for the model to infer as in Open, or left unstructured as in Flat. These node and edge types are derived directly from the execution trace itself rather than from an external legal ontology such as LKIF-Core or LegalRuleML [44,42]. The Graph condition differs from Flat in more than structure: it lists all evaluated conditions and their thresholds rather than only the decisive condition, uses a constrained system prompt that forbids facts outside the graph, and normalises rounded monetary amounts in the output to the exact engine value. Comparing Open versus Flat therefore isolates the effect of explicitly surfacing decisive conditions, while comparing Flat versus Graph measures the effect of the full Graph configuration, not of graph structure alone.

4.2. Rule-based decision engine and graph construction

Legal decisions are produced by a deterministic rule engine executing YAML-formalised regulations, treated as the authoritative source of legal reasoning for this proof-of-concept, formalisation errors would propagate into both explanations and evaluation (see Section 7). For each profile, a minimal decision subgraph is extracted containing only the decisive conditions, relevant facts, calculation results, and legal references needed to justify the outcome, reducing irrelevant context and limiting opportunities for hallucination.

5. Evaluation

5.1. Evaluation setup

Five LLMs are evaluated across all three regulations and prompting conditions under the same settings: GPT-4o [48], Claude Haiku 4.5 [49], Llama 3.1:8B [50], Mis-

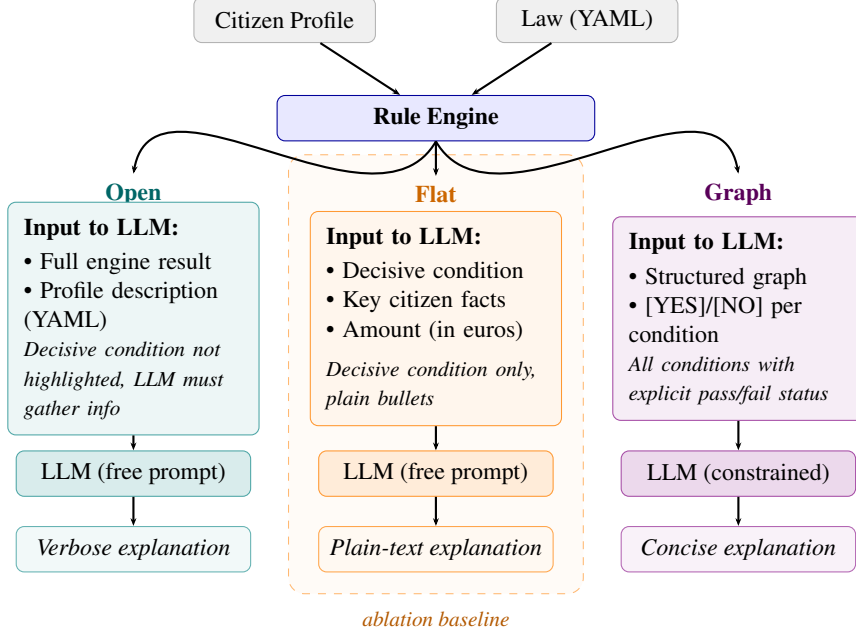


Figure 1. Overview of the System: three LLM explanation approaches.

tral:7B [51], and DeepSeek-R1:8B [52]. Llama 3.1:70B [53] is added to assess the effect of model scale independently of prompting condition. Low-temperature generation is used throughout to reduce variance. In the Graph condition, the prompt explicitly prohibits introducing facts, rules, or conclusions that are not in the graph.

5.2. Counterfactual robustness

For a subset of profiles per law and model, one decisive variable (income, age, insurance status, household composition) is modified across a legally relevant threshold, and the model is asked, as a follow-up in the same conversation, to explain the updated trace, to assess whether explanations update consistently with it.

5.3. Human annotation

A subset is annotated by three legal experts, each with more than ten years of experience in interpreting and applying Dutch administrative law, and three citizens with varied backgrounds. For each law, two fixed citizen cases are evaluated across all five models and both the Open and Graph conditions, so each annotator rates 60 explanations (360 ratings in total), independently and in randomised order. Legal experts additionally receive the execution trace and legal references and assess legal correctness, completeness, and factual consistency. Citizens assess readability, trust, and actionability. Inter-annotator agreement uses Cohen’s κ (categorical) and Kendall’s τ [54] (ordinal). Furthermore, alignment between automated metrics and human judgement is assessed via Pearson’s r , Spearman’s ρ , and Kendall’s τ [54].

5.4. Metrics

Legal grounding and faithfulness. Following evidence that NLI-based evaluation and claim-level verification outperform document-level scoring [36,38,55], each explanation is decomposed into up to three verifiable claims derived from the execution trace: the decision outcome, the awarded amount (where applicable), and the decisive condition. Outcome and amount are checked by exact matching, the decisive condition by keyword matching. Each claim i is scored as $x_i = 1$ if it is correctly represented in the explanation and $x_i = 0$ otherwise, and faithfulness is the proportion of correctly represented claims:

$$\text{Faithfulness} = \frac{1}{N} \sum_{i=1}^N x_i, \quad x_i \in \{0, 1\}, \quad (1)$$

with $N \in \{2, 3\}$ depending on whether an awarded amount applies to the case.

Because string matching is fragile to paraphrasing, it is complemented by a model-based estimate: an NLI entailment probability per claim is computed with mDeBERTa [56,57], treating the generated explanation as the premise and the trace-derived claim statement as the hypothesis, so that a claim is credited whenever it is semantically entailed by the explanation, independently of exact wording. This entailment probability is compared against the binary string-matching faithfulness labels via ROC AUC [58], following the TRUE framework for evaluating factual-consistency metrics [36]. The faithfulness scores reported in Tables 1 and 2 use these NLI judgements, counting a claim as correctly represented when its entailment probability is at least 0.5.

Readability and contestability. Readability uses the Dutch Flesch Reading Ease (FRE) score [59,60]. Contestability scores the presence of the decisive condition, a counterfactual, and a next step ($\frac{1}{3} \sum e_i, e_i \in \{0, 1\}$), reflecting calls for explanations to support contestability rather than only technical correctness [11,12].

6. Results

6.1. Performance across models and laws

Across all evaluated models, the Graph condition achieves the highest faithfulness (Table 1). Claude Haiku, Llama 3.1, and Mistral all show substantial improvements from Open to Flat and again from Flat to Graph, indicating that much of the improvement comes from explicitly exposing decisive conditions, with the Graph configuration adding a further increase. The largest gains occur for Llama 3.1 and Mistral (Open $\approx 0.50 \rightarrow$ Graph ≈ 0.95), suggesting smaller open-source models benefit most from explicit structure. Open prompting is generally most readable and richest in counterfactual statements, while Graph explanations are less readable but more focused on the executed decision path. Scaling Llama 3.1 from 8B to 70B improves readability and slightly improves faithfulness for Open/Flat, but does not change the ranking of prompting conditions, indicating the proposed architecture contributes more than model size alone.

The effect of prompting differs by legal domain (Table 2). For the Alcohol Law, all conditions score high faithfulness, likely reflecting a relatively simple decision structure and a dataset skew toward straightforward rejections (many synthetic profiles lack

Table 1. Faithfulness, readability (FRE), and contestability by prompting approach and model (mean $\pm$ SD across three laws). Highest mean per approach/metric in bold.

Approach	Model	Faithfulness	FRE	Contestability
Open	DeepSeek-R1:8b	0.66$\pm$0.18	52.2 $\pm$ 3.0	0.42$\pm$0.01
	Claude Haiku	0.59 $\pm$ 0.26	61.1 $\pm$ 1.1	0.33 $\pm$ 0.01
	GPT-4o	0.59 $\pm$ 0.29	61.5 $\pm$ 0.9	0.32 $\pm$ 0.02
	Llama3.1:8b	0.52 $\pm$ 0.38	65.4 $\pm$ 2.1	0.33 $\pm$ 0.00
	Mistral:7b	0.50 $\pm$ 0.40	49.2 $\pm$ 5.5	0.34 $\pm$ 0.00
	Llama3.1:70b	0.60 $\pm$ 0.33	70.0$\pm$6.0	0.36 $\pm$ 0.17
Flat	DeepSeek-R1:8b	0.74 $\pm$ 0.17	54.6 $\pm$ 4.4	0.34 $\pm$ 0.03
	Claude Haiku	0.91$\pm$0.06	58.9 $\pm$ 4.9	0.26 $\pm$ 0.01
	GPT-4o	0.85 $\pm$ 0.21	60.4 $\pm$ 3.3	0.33 $\pm$ 0.01
	Llama3.1:8b	0.73 $\pm$ 0.30	62.3 $\pm$ 4.1	0.34 $\pm$ 0.01
	Mistral:7b	0.78 $\pm$ 0.30	42.1 $\pm$ 9.0	0.34 $\pm$ 0.00
	Llama3.1:70b	0.82 $\pm$ 0.28	65.8$\pm$6.8	0.36$\pm$0.12
Graph	DeepSeek-R1:8b	0.78 $\pm$ 0.15	53.6 $\pm$ 10.5	0.19 $\pm$ 0.08
	Claude Haiku	0.96$\pm$0.06	49.1 $\pm$ 8.5	0.28 $\pm$ 0.06
	GPT-4o	0.96$\pm$0.05	53.4 $\pm$ 11.5	0.20 $\pm$ 0.14
	Llama3.1:8b	0.95 $\pm$ 0.03	54.5$\pm$13.7	0.16 $\pm$ 0.14
	Mistral:7b	0.96$\pm$0.05	44.6 $\pm$ 10.1	0.36$\pm$0.02
	Llama3.1:70b	0.87 $\pm$ 0.22	50.2 $\pm$ 14.3	0.17 $\pm$ 0.17

Table 2. Faithfulness, ROC AUC, readability, and contestability per law and approach (mean $\pm$ SD across GPT-4o, Claude Haiku, Llama3.1:8b, and Mistral:7b). Highest mean per law/metric in bold.

Law	Approach	Faithfulness	ROC AUC	FRE	Contestability
Alcohol	Open	0.97 $\pm$ 0.02	0.15 $\pm$ 0.21	56.8 $\pm$ 7.9	0.33$\pm$0.00
	Flat	1.00$\pm$0.00	0.22$\pm$0.09	58.2$\pm$7.2	0.32 $\pm$ 0.03
	Graph	1.00$\pm$0.00	— ^a	36.9 $\pm$ 2.0	0.21 $\pm$ 0.10
Social Assistance	Open	0.50 $\pm$ 0.01	0.07 $\pm$ 0.03	59.1$\pm$6.7	0.33$\pm$0.01
	Flat	0.94 $\pm$ 0.05	0.42 $\pm$ 0.17	48.6 $\pm$ 11.1	0.32 $\pm$ 0.04
	Graph	0.97$\pm$0.03	0.69$\pm$0.08	52.1 $\pm$ 8.6	0.18 $\pm$ 0.15
Health Care	Open	0.17 $\pm$ 0.14	0.69 $\pm$ 0.13	62.1 $\pm$ 4.0	0.32 $\pm$ 0.02
	Flat	0.52 $\pm$ 0.22	0.89$\pm$0.05	61.0 $\pm$ 6.0	0.31 $\pm$ 0.04
	Graph	0.89$\pm$0.02	0.80 $\pm$ 0.07	62.1$\pm$3.8	0.35$\pm$0.02

^aUndefined: all string-matching labels are positive, so only one class is present.

a Chamber-of-Commerce registration, since profiles are drawn from general rather than horeca-specific CBS statistics). For the Health Care Benefit Law, the most complex regulation, involving multiple thresholds and a numerical calculation, faithfulness rises from 0.17 (Open) to 0.52 (Flat) to 0.89 (Graph), the largest gain observed, indicating that structured representations become more valuable as legal complexity increases. For the Social Assistance Law, most of the improvement already occurs at the Flat layer. ROC AUC values below 0.5 (Social Assistance, Open: 0.07) indicate disagreement between string matching and NLI entailment, not a separate faithfulness failure.

Table 3. Explanations for one synthetic Health Care Benefit case (Claude Haiku, translated from Dutch, abridged). Trace: entitled, €1,630.66/year; assessed income €25,386.88 ≤ €39,719; assets €15,479.70 ≤ €141,896; age 68 ≥ 18; insured.

Open	“Good news! You meet <i>all conditions</i> . You are retired, so you have a fixed income. You are single. <i>Your income is probably not too high</i> . You receive €1,630.66 per year.”
Flat	“Your income of €25,386.88 is below the maximum of €39,719. <i>You are 65+, so you receive extra support</i> . You receive €1,630.66 per year.”
Graph	“You are entitled to health care benefit. The calculated amount is 1,630.66 euros per year. Your assessed income of 25,386.88 euros is below the maximum of 39,719.00 euros. Your assets of 15,479.70 euros are below the maximum of 141,896.00 euros. You are 68, so you meet the minimum age of 18. You are insured under the Health Insurance Act.”

6.2. Example

Table 3 shows one case. Open gets the outcome and amount right but replaces the decisive conditions with plausible, unchecked reasons; Flat adds an unsupported claim about age; Graph names each evaluated condition with its threshold.

6.3. Counterfactual robustness

Across evaluated counterfactual cases (GPT-4o, Claude Haiku, Llama 3.1, Graph condition), 55.4% of explanations explicitly incorporated the modified input variable and 62.2% clearly signalled the resulting outcome change, indicating that the pipeline is sensitive to changing legal conditions but does not yet consistently communicate counterfactual reasoning to citizens.

6.4. Human annotation

Citizens generally rated explanations more positively than legal experts, especially when explanations were readable but not legally decisive. This pattern was strongest for the Health Care Benefit Law, where explanations were often perceived as understandable despite legal inaccuracies, while Social Assistance showed closer agreement between citizens and experts. Experts rated Graph explanations as more legally correct than Open explanations for the Alcohol Law (3.90 vs. 2.93 on a 0–5 scale) and the Social Assistance Law (3.60 vs. 2.70), but not for the Health Care Benefit Law (2.03 vs. 2.38), where both conditions scored low. Inter-annotator agreement was stronger for ordinal dimensions (legal correctness: Kendall’s $\tau = 0.55$, $p < 0.01$, comprehensibility: $\tau = 0.37$, $p < 0.001$) than for categorical dimensions (missing information: Cohen’s $\kappa = 0.44$, actionability: $\kappa = 0.01$), suggesting citizen-focused quality remains partly subjective even under structured guidelines. Automated faithfulness correlates positively with expert-rated legal correctness ($\rho = 0.59$, $p < 0.001$), and contestability correlates positively with citizen scores and clarity ($\rho = 0.38$ – 0.44 , $p < 0.05$). Flesch readability correlates negatively with both legal correctness ($\rho = -0.68$, $p < 0.001$) and citizen clarity ($\rho = -0.50$, $p < 0.01$), indicating that easier-to-read explanations are not necessarily more correct or more useful: legally precise explanations often require more explicit detail, which can reduce readability while improving faithfulness.

7. Discussion

The results show that structuring rule-engine traces improves the faithfulness of LLM-generated explanations, with the largest improvements for regulations involving multiple conditions and calculations. Several limitations qualify this conclusion. First, synthetic profiles generated from aggregated CBS statistics [47] cannot capture the ambiguity and edge cases of real administrative data, most visibly for the Alcohol Law, where the absence of a horeca-specific population inflates the share of trivial rejection cases. Second, the framework assumes the YAML formalisation correctly implements the underlying legislation, any formalisation error propagates into both explanations and evaluation, a limitation shared with other RaC systems [2]. Third, ROC AUC scores reflect agreement between string-matching and NLI-based verification rather than an externally validated ground truth, unlike the human-annotated TRUE benchmark [36]. Fourth, the Graph condition combines graph structure with a constrained prompt and amount normalisation, so its gain over Flat cannot be attributed to structure alone; normalisation in particular favours the amount claim. Fifth, faithfulness checks only the outcome, amount, and decisive condition, and does not penalise additional unsupported statements; completeness is assessed only through the expert judgement of missing information. This is most visible for the Health Care Benefit Law, where the large gain in automated faithfulness did not carry over to expert-rated legal correctness, and experts frequently reported missing information. Sixth, we report means and standard deviations without significance tests or confidence intervals, so small differences, such as between models within the Graph condition, should not be over-interpreted. Finally, scaling to hundreds of regulations would require a law-selection layer before rule execution, for example along the lines of HippoRAG [61].

Beyond these technical limitations, the human annotation results reveal a trade-off between legal faithfulness and citizen perception: Graph explanations were more faithful and, for two of the three laws, rated more legally correct, yet citizens sometimes preferred the more elaborate Open explanations. Improving legal faithfulness alone is therefore unlikely to be sufficient. Future work should investigate explanation strategies that preserve alignment with deterministic legal reasoning while meeting citizens' informational needs. The results should not be interpreted as evidence that such systems are ready for operational deployment without human oversight, legal validation, and governance procedures for evolving legislation [29,12].

8. Conclusion

Structuring deterministic rule-engine traces improves faithfulness relative to less constrained generation, and explicitly exposing decisive legal conditions accounts for much of this improvement, with the Graph configuration providing a further gain across five LLMs and three Dutch regulations. At the same time, explanations most closely aligned with the execution trace were not always the most readable or citizen-preferred, and explanations rated as understandable by citizens were not always judged legally correct by experts. Explanation quality in this domain therefore cannot be reduced to faithfulness or readability alone, but requires balancing legal precision, transparency, and citizen understanding.

References

- [1] Fraser H, Barraclough T. Governing Digital Legal Systems: Insights on Artificial Intelligence and Rules as Code. MIT Computational Law Report. 2024. Available from: <https://law.mit.edu/pub/governingdigitallegalsystems>.
- [2] Mohun J, Roberts A. Cracking the Code: Rulemaking for Humans and Machines; 2020. OECD Working Papers on Public Governance, No. 42, OECD Publishing.
- [3] Hotting E, Schuth A. Rules as Executed: Publishing Machine-Executable Law to Rebalance the Powers; 2026. Working paper, Zenodo. Available from: <https://doi.org/10.5281/zenodo.22201094>.
- [4] Zhang Z, Fang M, Chen L, Namazi-Rad MR, Wang J. How Do Large Language Models Capture the Ever-Changing World Knowledge? A Review of Recent Advances. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; 2023. p. 8289-311.
- [5] Han S, Zhang Y, Huang Y, Li X, Liu C, Guo Y. Reimagining Legal Fact Verification with GenAI: Toward Effective Human-AI Collaboration. In: Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems. ACM; 2026. p. 1-15.
- [6] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*. 2020;33:9459-74.
- [7] Peng B, Zhu Y, Liu Y, Bo X, Shi H, Hong C, et al. Graph retrieval-augmented generation: A survey. *ACM Transactions on Information Systems*. 2025;44(2):1-52.
- [8] Nguyen HT, Satoh K. KRAG Framework for Enhancing LLMs in the Legal Domain; 2024. arXiv:2410.07551.
- [9] Nguyen HT, Satoh K. ConsRAG: Minimize LLM Hallucinations in the Legal Domain. In: *Legal Knowledge and Information Systems*. vol. 395 of *Frontiers in Artificial Intelligence and Applications*. IOS Press; 2024. p. 327-32.
- [10] Wehnert S, Padmanabhan V, De Luca EW. Hybrid Legal Norm Retrieval: Leveraging Knowledge Graphs and Textual Representations. In: *Legal Knowledge and Information Systems*. vol. 395 of *Frontiers in Artificial Intelligence and Applications*. IOS Press; 2024. p. 191-203.
- [11] Wu X, Zhao H, Zhu Y, Shi Y, Yang F, Hu L, et al. Usable XAI: 10 Strategies Towards Exploiting Explainability in the LLM Era; 2024. arXiv:2403.08946.
- [12] Zuurmond S, Borg A, Van Kempen M, Wieten R. Human-Centred Explanation of Rule-Based Decision-Making Systems in the Legal Domain; 2023. arXiv:2310.16704.
- [13] Nouwens J, Borg A, Prakken H. An Application of Case-Based Reasoning to Decision-Making in Dutch Administrative Law. In: *Legal Knowledge and Information Systems*. vol. 395 of *Frontiers in Artificial Intelligence and Applications*. IOS Press; 2024. p. 131-41.
- [14] Bouche-Pillon J, Aussenac-Gilles N, Chevalier Y, Zaraté P. Decision Support in Law: From Formalizing Rules to Reasoning with Justification. In: *Legal Knowledge and Information Systems*. vol. 395 of *Frontiers in Artificial Intelligence and Applications*. IOS Press; 2024. p. 25-36.
- [15] Wet op de zorgtoeslag; 2026. BWBR0018451. Available from: <https://wetten.overheid.nl/BWBR0018451/2026-01-01/0>.
- [16] Participatiewet; 2026. BWBR0015703. Available from: <https://wetten.overheid.nl/BWBR0015703/2026-04-03/0>.
- [17] Alcoholwet; 2025. BWBR0002458. Available from: <https://wetten.overheid.nl/BWBR0002458/2025-02-12>.
- [18] van Doesburg R, van Engers TM. Perspectives on the Formal Representation of the Interpretation of Norms. In: *Legal Knowledge and Information Systems*. vol. 294 of *Frontiers in Artificial Intelligence and Applications*. IOS Press; 2016. p. 183-6.
- [19] van Engers TM, van Doesburg R. At Your Service: On the Definition of Services from Sources of Law. In: *Proceedings of the 15th International Conference on Artificial Intelligence and Law (ICAIL 2015)*. ACM; 2015. p. 221-5.
- [20] Ali S, Sileno G, van Engers TM. A Systematic Approach to Assess Languages Proposed for Rules as Code. In: *Legal Knowledge and Information Systems*. vol. 416 of *Frontiers in Artificial Intelligence and Applications*. IOS Press; 2025. p. 264-9.
- [21] Sileno G, Boer A, Van Engers T. The Role of Normware in Trustworthy and Explainable AI; 2018. arXiv:1812.02471.
- [22] Bertl M, Price S, Draheim D. Transforming legal texts into computational logic: Enhancing next generation public sector automation through explainable AI decision support. *International Journal of Cogni-*

tive Computing in Engineering. 2025;7:40-57.

- [23] Price S, Bertl M. Extracting Computational Logic from Legal Text: A Decision Support Approach for Public Sector Automation; 2024. TechRxiv preprint.
- [24] Bakker R, van Drie RA, De Boer M, van Doesburg R, van Engers T. Semantic Role Labelling for Dutch Law Texts. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference; 2022. p. 448-57.
- [25] Parcalabescu L, Frank A. On measuring faithfulness or self-consistency of natural language explanations. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2024. p. 6048-89.
- [26] Lyu Q, Apidianaki M, Callison-Burch C. Towards Faithful Model Explanation in NLP: A Survey. Computational Linguistics. 2024;50(2):657-723.
- [27] Prakken H, Ratsma R. A top-level model of case-based argumentation for explanation: formalisation and experiments. Argument & Computation. 2022;13(2):159-94.
- [28] Górski Ł, Ramakrishna S. Explainable Artificial Intelligence, Lawyer's Perspective. In: Proceedings of the 18th International Conference on Artificial Intelligence and Law (ICAIL 2021). ACM; 2021. p. 60-8.
- [29] Hummel A, Burden H, Stenberg S, Steghöfer JP, Kühl N. The Bidirectional Relationship Between XAI and Regulation: Operationalizing XAI for the AI Act; 2026. Available from: <https://arxiv.org/abs/2505.20311>.
- [30] Jiang C, Yang X. Legal Syllogism Prompting: Teaching Large Language Models for Legal Judgment Prediction. In: Proceedings of the 19th International Conference on Artificial Intelligence and Law (ICAIL 2023). ACM; 2023. p. 417-21.
- [31] Blair-Stanek A, Holzenberger N, Van Durme B. Can GPT-3 Perform Statutory Reasoning? In: Proceedings of the 19th International Conference on Artificial Intelligence and Law (ICAIL 2023). ACM; 2023. p. 22-31.
- [32] Lindsay C, Toni F, Nguyen TD, Nguyen TM, Pham NAT, Chu QH, et al. Argumentative LLMs for Legal Information Entailment. In: Legal Knowledge and Information Systems. vol. 416 of Frontiers in Artificial Intelligence and Applications. IOS Press; 2025. p. 121-32.
- [33] Gultekin S, Rossi Reich M, Galloni F, Lagioia F, Consiglio E, Sartor G, et al. A Two-Dimensional Evaluation Framework for Factual and Reasoning Assessment of LLMs in Legal Question Answering. In: Legal Knowledge and Information Systems. vol. 416 of Frontiers in Artificial Intelligence and Applications. IOS Press; 2025. p. 86-97.
- [34] Sanchi M, Novotná T. Towards Hybrid Evaluation Methodologies for Large Language Models in the Legal Domain. In: Legal Knowledge and Information Systems. vol. 395 of Frontiers in Artificial Intelligence and Applications. IOS Press; 2024. p. 389-92.
- [35] Blair-Stanek A, Van Durme B. LLMs Provide Unstable Answers to Legal Questions. In: Proceedings of the 20th International Conference on Artificial Intelligence and Law (ICAIL 2025). ACM; 2025. p. 425-9.
- [36] Honovich O, Aharoni R, Herzig J, Taitelbaum H, Kukliansy D, Cohen V, et al. TRUE: Re-evaluating factual consistency evaluation. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2022. p. 3905-20.
- [37] Gekhman Z, Herzig J, Aharoni R, Elkind C, Szpektor I. TrueTeacher: Learning factual consistency evaluation with large language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; 2023. p. 2053-70.
- [38] Min S, Krishna K, Lyu X, Lewis M, Yih Wt, Koh P, et al. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; 2023. p. 12076-100.
- [39] Tang L, Laban P, Durrett G. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; 2024. p. 8818-47.
- [40] Quan X, Valentino M, Dennis LA, Freitas A. Faithful and robust LLM-driven theorem proving for NLI explanations. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2025. p. 17734-55.
- [41] Cai L, Kang Y, Yu C, Fu Y, Zhang H, Zhao Y. Bringing two worlds together: The convergence of large language models and knowledge graphs. In: 2024 3rd International Conference on Automation, Robotics and Computer Engineering (ICARCE). IEEE; 2024. p. 207-16.

- [42] Athan T, Governatori G, Palmirani M, Paschke A, Wyner A. LegalRuleML: Design Principles and Foundations. In: Reasoning Web. Web Logic Rules. vol. 9203 of Lecture Notes in Computer Science. Springer; 2015. p. 151-88.
- [43] Athan T, Boley H, Governatori G, Palmirani M, Paschke A, Wyner A. OASIS LegalRuleML. In: Proceedings of the 14th International Conference on Artificial Intelligence and Law (ICAIL 2013). ACM; 2013. p. 3-12.
- [44] Hoekstra R, Breuker J, Di Bello M, Boer A. LKIF Core: Principled Ontology Development for the Legal Domain. In: Law, Ontologies and the Semantic Web: Channelling the Legal Information Flood. vol. 188 of Frontiers in Artificial Intelligence and Applications. IOS Press; 2009. p. 21-52.
- [45] Grimmelikhuijsen S. Explaining Why the Computer Says No: Algorithmic Transparency Affects the Perceived Trustworthiness of Automated Decision-Making. Public Administration Review. 2023;83(2):241-62.
- [46] Autoriteit Persoonsgegevens. Werkwijze Belastingdienst in strijd met de wet en discriminerend; 2020. Available from: <https://www.autoriteitpersoonsgegevens.nl/actueel/werkwijze-belastingdienst-in-strijd-met-de-wet-en-discriminerend>.
- [47] CBS. Dashboard Bevolking; 2026. Available from: <https://www.cbs.nl/nl-nl/visualisaties/dashboard-bevolking>.
- [48] OpenAI. Hello GPT-4o; 2024. Available from: <https://openai.com/index/hello-gpt-4o/>.
- [49] Anthropic. Claude Haiku 4.5; 2025. Available from: <https://www.anthropic.com/claude/haiku>.
- [50] Ollama. llama3.1:8b; 2025. Available from: <https://ollama.com/library/llama3.1:8b>.
- [51] Mistral. Mistral 7B; 2023. Available from: <https://mistral.ai/news/announcing-mistral-7b>.
- [52] Deepseek. Deepseek-R1:8B; 2025. Available from: <https://ollama.com/library/deepseek-r1:8b>.
- [53] Ollama. Llama3.1:70B; 2025. Available from: <https://ollama.com/library/llama3.1:70b>.
- [54] Kendall MG. A New Measure of Rank Correlation. Biometrika. 1938;30(1-2):81-93.
- [55] Sansford H, Richardson N, Maretic HP, Saada JN. GraphEval: A Knowledge-Graph Based LLM Hallucination Evaluation Framework; 2024. Available from: <https://arxiv.org/abs/2407.10793>.
- [56] He P, Gao J, Chen W. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing; 2021. arXiv:2111.09543.
- [57] Laurer M, Van Atteveldt W, Casas A, Welbers K. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and BERT-NLI. Political Analysis. 2024;32(1):84-100.
- [58] Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006;27(8):861-74.
- [59] Bansal S. GitHub - textstat/textstat: Python package to calculate readability statistics of a text object - paragraphs, sentences, articles; 2018. Available from: <https://github.com/textstat/textstat>.
- [60] de Bock H, Debeuckeleare G, Deggeler L, van Hauwermeiren P, Jansen CJM, van der Noort W, et al. Leesbaarheidsformules. Massacommunicatie. 1982;10(3):115-23.
- [61] Gutiérrez BJ, Shu Y, Gu Y, Yasunaga M, Su Y. HippoRAG: Neurobiologically inspired long-term memory for large language models. Advances in Neural Information Processing Systems. 2024;37:59532-69.