buildings

MDPI

Article

Human-in-the-Loop Semantic Rule Base Generation and Dynamic Updating for Automated BIM Compliance Checking: A Knowledge Graph Approach

Zhuoqun Zhang [1], Chunhui Zhao [1], Danli Li [1], Peijie Li [1], Yulong Chen [2], Hao Zhu [2] and Daguang Han [3,*]

[1] State Grid Corporation of China, Future Science & Technology Park, State Grid Economic and Technological Research Institute Co., Ltd., Beijing 102209, China; zhangzhuoqun@chinasperi.sgcc.com.cn (Z.Z.); zhaochunhui@chinasperi.sgcc.com.cn (C.Z.); lidanli@chinasperi.sgcc.com.cn (D.L.); lipeijie@chinasperi.sgcc.com.cn (P.L.)
[2] School of Civil Engineering, Chongqing Jiaotong University, Chongqing 400074, China; 17623356953@163.com (Y.C.); knowledgepuppy4869@gmail.com (H.Z.)
[3] School of Civil Engineering, Southeast University, Nanjing 211189, China
* Correspondence: daguanghan@seu.edu.cn

Abstract

Objective: This paper seeks to provide an effective and automated method for the creation and updating of building information modeling compliance rules using the integration of human-in-the-loop collaboration with advanced natural language processing. Methods: We propose a hybrid approach that integrates BERT-based semantic extraction, CFG structural validation, and confidence-based expert review. Results: Tested on a real-world power infrastructure project, the framework was found to be 95.8% accurate in translation and 98.3% feasible in rule execution, outperforming benchmark automated approaches. The human effort was reduced by 90% (168 h vs. 1620 h), and processing of regulatory changes was sped up by 94%. Conclusion: Data analysis shows that collaborative intelligence is a significant factor in closing the semantic and pragmatic gap for regulatory compliance. Compared with fully automated "black box" approaches, this method supplies a tractable, manageable, and operationally valid solution, giving a competitive edge over existing digital construction methods.

Keywords: building information modeling; automated compliance checking; human-in-the-loop; knowledge graph; semantic recognition; natural language processing

Academic Editor: Pramen Shrestha

Received: 12 November 2025
Revised: 23 January 2026
Accepted: 30 January 2026
Published: 10 February 2026

1. Introduction

Building regulations and standards are at the heart of building safety, accessibility, and environmental sustainability and are constantly updated by regulatory agencies to keep abreast of new challenges and new technology developments. This process gives rise to a dilemma of regulatory comprehensiveness and practicability. In China, there are over 1400 building standards at a national and local level that apply to the building sector [1], and prominent standards like GB 50016 "Code for Fire Protection Design of Buildings" are updated at periodic intervals of 3–5 years to factor in new experiences gained from disasters, new technology developments, and international practices [2]. The Chinese building standards' hierarchical structure consisting of national standards (GB), industry standards (JGJ), and local standards (DBJ) adds a few more layers of complexity because of conflicting standards, overlapping standards, and tacit dependencies that call for specialized expertise.

However, manual implementation of these rules in machine-executable forms still poses an important bottleneck [3–5], as it is labor-intensive and error-prone. Current state-of-the-art automated techniques are faced with an important challenge: Rule-based systems are not scalable [6–8], while other NLP-based techniques [9,10] (like BERT/LLM) are. This gap refers to the disconnection between linguistic meaning and executable logic, a phenomenon well-documented in legal text processing [11] and requirements engineering [12]. It represents the "last mile" problem where a system correctly parses the sentence structure (syntax) but fails to map it to a computable operation valid in the target domain environment. A critical manifestation of this gap is the risk of LLM hallucinations [13] in safety-critical engineering. For instance, a generative model might produce a syntactically perfect SHACL shape referencing "IfcCorridor"—an entity that does not exist in the official IFC4 schema—resulting in silent validation failures during runtime. Recent studies [14] highlight that, without grounding in the physical schema, NLP models struggle to bridge this divide.

Consequently, this study aims to create a rule base generation and updating system with a human-in-the-loop (HITL) methodology in a more automated fashion, whereby we will attempt to: (1) Close the semantic pragmatics gap [15,16] by combining formal grammar (CFG) and model validation; (2) address the maintenance issue with a RAG-supported evolutionary process; and (3) validate the executability of a cooperative intelligence solution in a real-life infrastructure application.

Analysis shows that there are three key theoretical and practical gaps that represent fundamental hurdles to the widespread adoption of automated compliance checking systems. First, there is the trade-off between accuracy and efficiency [17] that is characteristic of the current level of computational limitations in representing regulatory knowledge. For example, systems with high accuracy are necessarily highly labor-intensive since they are based on domain knowledge put together by experts, while less accurate systems are necessary since the regulatory texts are inherently ambivalent and complex. This trade-off comes up as the "90% barrier" since, while the 90% accuracy level may easily be reached through automation, the 95–98% level needed in safety-critical domains seems to represent fundamental hurdles that lie beyond purely automated and purely manual systems.

Second, there is a crucial semantic–pragmatic mismatch between text parsing and operational validity for a majority of NLP-oriented techniques. They primarily concentrate on parsing semantic roles from normative text without checking whether these rules can actually be correctly executed on real-world IFC building models. Empirical assessments have supported the claim that "correctly parsed" rules have a 15–25% error rate during runtime because of schema incongruity, geometric calculations, and topological issues [18,19]. There appears to be a massive mismatch between linguistic validity and pragmatic validity, implying a rule may be linguistically correct for natural language interpretation but pragmatically invalid for execution as a program.

Third, present solutions do not provide any mechanistic approach towards adaptability related to modifications in rules, as these rules remain fixed computational objects [20,21]. This approach becomes even more challenging in dynamic environments where rules need frequent modification as new safety issues arise, new technologies emerge, and new governmental policies evolve. In this case, any modification in the rules becomes challenging, as a completely new, regenerated version of rules needs to be implemented at a high cost by the organization, leading to reduced accuracy of compliance checks with new rules, which is even more challenging in dynamic environments. In software engineering, this act violates the open–closed principle.

The shortcomings are overcome using a human-in-the-loop approach that leverages semantic recognition using BERT along with model validation and version evolution [22].

Accuracy above 95% is guaranteed while automating, and for code executability, IFC schema validation is used, while efficient regulatory updates are facilitated using sentence embeddings and RAG. The proposed approach has been applied in a State Grid power infrastructure project that shows translation accuracy at 95.8%, 98.3% for executability, and a 90% reduction in human effort to code rules manually, as given in Table 1 that shows comparison of proposed work with other approaches.

Table 1. Comparison of Automated Compliance Checking Approaches.

Approach	Representative Work	Accuracy	Maintainability	Explainability	Executability Validation
Hard-coded rules	Solibri [23], SMARTcodes [24]	95–98%	Low (manual coding)	High	Yes (embedded)
NLP extraction	Peng, 2023 [25], Zhou, 2016 [26]	85–90%	Medium	Medium	No
LLM-driven	Chen, 2024 [10], Madireddy, 2024 [27]	88–92%	Medium	Low (black-box)	No
Knowledge graphs	Peng, 2023 [25], Pauwels, 2016 [28]	N/A (manual)	Low (static)	High	Partial
Proposed (HITL)	This work	95.8%	High	High	Yes

2. Materials and Methods

2.1. Framework Overview

AI-assisted tools (GPT-4 (OpenAI, San Francisco, CA, USA)) were used to assist with data organization and sorting. All outputs were reviewed and verified by the authors. The proposed framework fills the theoretical gaps by means of a collaborative intelligence architecture that combines symbolic reasoning, statistical learning, and human intelligence in a principled manner. This approach represents four key theoretical foundations that were taken from the field of cognitive science and artificial intelligence studies, including (1) hybrid reasoning that combines the accuracy of symbolic logic and the strength of statistical learning, (2) uncertainty-driven collaboration that directs tasks to humans based on the confidence of the computation rather than a predetermined threshold, (3) model-based validation that validates correctness of the operation via target execution environments, (4) evolutionary maintenance that provides progressive adaptation to changing requirements via systematic change detection.

The framework combines the above four mutually supporting modules, designed following the principles put forth by the theory of distributed cognition, which argues that the best way to complete complex tasks of reason is through the cooperative workings of multiple reason-computation entities. Module 1 (Hybrid Semantic Recognition): Uses BERT + CFG to derive the semantics of regulation from natural language statements, fusing the contextual understanding abilities offered by the transformer architecture with the constraints of formal grammars. Module 2 (Model-driven Validation): Uses the constraints defined through the Shapes Constraint Language (SHACL) to validate the executability of the regulation rule, choosing a computational pragmatics approach to validate that the derived meanings are sound transformations into operational processes. Module 3 (Expert Review Interface): Implements human-in-the-loop filters based on confidence measures [29], which utilize the principles of uncertainty quantification to carry out optimal utilization of human cognitive power. Module 4 (Versioned Evolution): Tracks the evolution of the regulation through sentence embeddings/RAG, which follows the principles of software evolution theory to keep the entire system coherent.

The architectural design is based on the insights obtained from the theory of complex adaptive systems. Thus, instead of trying to solve the entire problem with any given methodology, the system demonstrates how emergent intelligence can be achieved through

the interaction of specialized subprocesses. Every module is involved in solving one aspect of the total problem and is designed in such a way that there are interfaces through which information can flow. The architectural design depicted in Figure 1 below represents how the system achieves cooperation through certain information flow patterns that make human–machine cooperation possible.

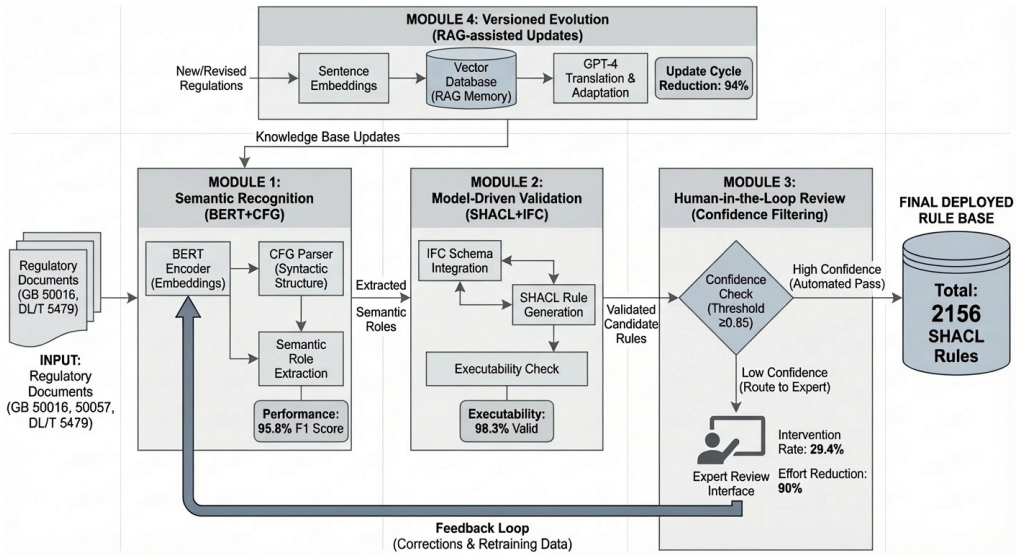

Figure 1. The proposed four-module framework for automated compliance checking, integrating semantic recognition, validation, human-in-the-loop adjudication, and evolvable updates.

2.2. Hybrid BERT + CFG Semantic Recognition

The semantic recognition component also deals with the problem of how to achieve a high degree of structure discovery in semi-structured regulatory text by a novel combination of a connectionist and a symbolic component. This component is based on a recognition of how these two types of systems are less than fully adequate separately for a problem of natural language understanding. While connectionist systems are excellent at capturing contextual semantics and linguistic variability, they are weak at compositional semantics and structural constraints.

Additionally, the hybrid architecture uses dual-path processing, in which the neural and symbolic parts process data in parallel before the data is combined. While the neural processing uses BERT-base-Chinese (Hugging Face, New York, NY, USA), with 110 million parameters, to get the context-based understanding of the semantics, the symbolic processing uses context-free grammars to encode structural constraints found in regulatory discourse. This is based on insights from dual-process models of cognition, according to which human reasoning is supported by both intuitive pattern recognition (System 1) and logical analysis (System 2).

2.2.1. Neural Semantic Role Extraction

The "neural" component is based on Beginning–Inside–Outside (BIO) sequence labeling and is able to extract seven semantic roles in the regulation clauses, which denote the basic building blocks of compliance requirements extracted via corpus analysis. These roles are subject object (SO), controlled object (CO), measurable property (MP), comparison operator (COP), reference value (RV), additional reference property (ARP), and reference object (RO), outlining the semantic frame of regulation requirements and enabling consequent encoding into executable constraints.

We opted for BERT-base-Chinese (12 layers, 768 hidden, 12 attention) and 110 M parameters. To fine-tune the model we used the pretrained version available at Hugging Face's model zoo.

The BIO tagging scheme is driven by the following theoretical insights from computational linguistics: The semantics of the regulatory requirements are compositional. Meaningful combinations of semantic roles are composed in a regulated environment. On the other hand, in general text corpora, the boundaries of the entity are fuzzy. However, the discourse pattern is more restricted in the context of regulation. This makes the tagging scheme more precise. The seven-role framework was derived from the analysis of 2847 clauses of the five major Chinese building regulations.

BERT fine-tuning utilized domain-adaptive training approaches inspired by principles of transfer learning theories. General domain training with general Chinese text enables general linguistic knowledge, while domain-specific fine-tuning using 611 manually annotated regulatory clauses specializes in learning domain-specific vocabulary, syntax, and semantic relationships of the domain of regulation from specialized texts. The fine-tuning used adversarial regularization with a dropout rate of 0.3 and a weight decay of 1×10^{-5} to avoid overspecialization on domain-specific textual form but retain generalization capabilities over unseen formulations.

2.2.2. Symbolic Structural Validation

For the purposes of structural validity, we used a parser formulated using the tuple CFG = (N, Sigma, P, S), where N is the set of non-terminal symbols (for example, Rule or Condition), Sigma is the set of terminal symbols (the first 7 roles), P is the set of productions, and S the start symbol. The set of primary productions comprises the following productions:

1. rightarrow Subject + Property + Operator + Value (Basic Constraint).
2. Subject + Property + Operator + Reference Object+ Reference Property (Comparative Constraint).

The grammar rules were informed by linguistic analysis of regulatory clause structure and were based upon a set of five archetypes: (1) Basic requirements (SO + MP + COP + RV), (2) comparative requirements (SO + MP + COP + RV + RO), (3) attribute specifications (SO + MP + RV), (4) conditional requirements (Context + Basic/Comparative), (5) multi-constraint requirements (multiple Basic patterns and logical connectives). This set of grammars accounts for 94.3% of regulatory clauses.

The CFG validation system establishes syntactic filtering that removes structurally incorrect neural outputs but maintains flexibility in semantics. A constraint with no subject entity violates the principles of semantic compositionality, while the use of the comparison operator on non-numeric properties undermines type consistency. The filtering system improved precision from 86.3% to 91.7% by reducing error patterns in systematic ways but maintaining robustness in linguistic variability handled by the BERT system.

2.2.3. Uncertainty-Driven Human–Machine Collaboration

A principled approach to uncertainty quantification is applied in this framework, ensuring optimal human–machine cooperation based on confidence predictions. A challenge in applying artificial intelligence in real-world applications is knowing when human assistance is required. Instead of applying fixed confidence levels, this method evaluates confidence in prediction accuracy by applying calibrated uncertainty values.

Confidence estimation uses the geometric mean of token probabilities over sequences of labels based on principles from information theory. For entity ranging over tokens from position startstartstart to endendend, confidence estimation is given by:

$$\text{conf}(e) = \left(\prod_{i\,=\,start}^{end} \max_{y \in L_e} P(y_i \mid C) \right)^{\frac{1}{end\text{-}start\ +\ 1}}$$

where L_e represents the set of BIO tags that are associated with the semantic role of the entity with index e, and $P(y_i|C)$ represents the probability that the tag is assigned to the context C. The entities that have a confidence score below 0.85 are selected to be manually examined.

This confidence measure requires Platt scaling for calibration against historical expert feedback in order to ensure correct representation of confidence measures as probabilities of correctness. It has been noted that many neural networks are overconfident about their predictions [30–33], meaning they tend to give high confidence for incorrect predictions. It aims at ensuring that confidence measures of 0.85, for example, correspond to actual 85% correctness.

The adaptive threshold function adjusts the boundary levels according to observed correction patterns among experts on how they typically correct confidence level classifications, following active learning strategies that aim to optimize precision vs. effort trade-offs. When there are observed correction patterns on classifications made from confidence levels 0.80–0.85, the system adjusts the threshold upwards to direct more classifications that are on the margin toward the experts. However, when experts show consistency on classifications near the thresholds, the system adjusts the thresholds downwards.

The overall hybrid pipeline, end-to-end, is succinctly explained in Algorithm 1 below.

Algorithm 1. Hybrid Semantic Recognition (BERT + CFG + HITL)

```
Ŝ, P ← BERT_BIO_Tagging(C)                    # token labels and probabilities
S' ← Apply_CFG_Constraints(Ŝ)                 # enforce structural patterns
for each entity e in S':                      # uncertainty estimation + calibration
    conf(e) ← GeometricMean(P_e)
    conf(e) ← PlattCalibrate(conf(e))
end for
if min_e conf(e) ≥ τ then                     # uncertainty-driven routing
    S ← S'; route ← AUTO_ACCEPT
else
    S ← ExpertReview(C, S')                   # expert correction if needed
end if
UpdateModelWith(S) every K=200 annotations    # active learning fine-tuning
```

2.3. Model-Driven Executability Validation

2.3.1. Rule-to-SHACL Translation

Semantic model validation is the process whereby the semantic model components are translated into constraints based on the W3C standard titled SHACL for the validation [34,35] of resource description framework (RDF) graphs [36,37]. To achieve semantic model validation between natural language regulatory expressions and semantic model validation logic executable on IFC building models, five templates were designed for translating typical expressions in Chinese building regulations.

A basic requirement containing a single property constraint for a building component translates to semantic components with SHACL node shapes focused on certain classes of IFC entities with property constraints. The subject entity identifies the target class, for instance, associating "distribution room" with Power Distribution Room from our domain ontology, and then the property name is turned into a SHACL property path. The comparison operators are translated to respective SHACL constraint predicates; for example, "shall not be less than" represents a link to a minimum value constraint, while "shall be" represents a link to an equality constraint, and "shall not exceed" represents a link to a maximum value constraint. The values for references are normalized with

unit conversion from common units defined for Chinese building standards to decimal representations with defined data types like meters, square meters, and degrees Celsius.

Being comparative, requirements that relate multiple architectural elements necessitate SHACL compositional shapes. For instance, "the clear width of the distribution room door shall not be less than 1.2 times the width of the corridor" necessitates two entities with a comparative relationship of their properties. The transformation entails coupled shape definitions, where the main shape (distribution room) references a secondary shape (door) via a relationship path, and the secondary shape definition enunciates the property relationship to the comparative standard, in this case, the width of the corridor. This model of composition generalizes to conditional rules in "when–then" forms, relation definitions of space (proximity, containment, connectivity), and quantified rules that aggregate multiple elements.

2.3.2. IFC Property Extraction

To verify the SHACL constraints on real BIMs, the abstract concepts in regulations need to be matched to the IFC entities and their properties. A special ontology, created by examining the IFC4 schema systematically, maps the general regulatory expressions to the IFC property sets. Preparing such an ontology required comparisons between 48 building regulations in China [38–40] and the IFC entities and their properties based on the references made to the elements, attributes, and relations in the buildings.

The process of property extraction involves four steps. Extracting base properties involves accessing properties through IFC file element characteristics directly using the IfcOpenShell library, for example, accessing a door's overall width using the IfcDoor OverallWidth attribute. Calculating geometric derivatives involves accessing properties not otherwise stored in an IFC file but referenced by building regulations, where, for example, a door's clear width is calculated through an overall width subtracting twice the thickness of the door frame, which involves geometric calculation through a number of IFC file properties. Extracting property sets involves accessing specific IFC file schema-defined property sets, Pset_DoorCommon for doors (fire-resistance rating, acoustic ratings) and Pset_SpaceCommon properties related to rooms (occupancy type, finish details), respectively. Topology relations entail tracing relations using IFC file objects, specifically IfcRelContainedInSpatialStructure (e.g., a door is traced within a room, which is traced within a building), and IfcRelConnectsElements (e.g., a door connects two rooms adjacently located together). For non-standard properties (user-defined properties/UDPs), the system utilizes a semantic mapping configuration file. When a property in the regulation (e.g., "FireRating") does not match the standard IFC Psets, the system searches the model's "IfcPropertySet" for phonetically similar names (using Levenshtein distance). Crucially, if the matching confidence score is below a safety threshold (0.8), the property is not auto-mapped but flagged for manual confirmation, preventing erroneous mappings (e.g., confusing "FireRating" with "FireExit").

This multi-step process ensures that constraints within regulations can be appraised against not just direct property information but also the complete content of IFC models, thus facilitating the validation of calculated and spatial attributes considered crucial for building regulation compliance.

2.3.3. Conflict Detection

The SHACL validation engine executes the constraints on the RDF graphs [36,37] produced from the IFC data and reports the violations that are categorized into three types. Schema mismatches happen because the extracted rules cite properties or entity types from the IFC schema and the current BIM under test that may not exist or are absent. For

instance, a rule that mandates "acoustic insulation rating" can extract properly from the text of the code but will result in a failure if the IFC model does not have the IfcSoundProperties data. Type mismatches occur because of the discrepancies between the regulatory text and IFC data types. For example, when the IFC model uses numeric quantity data types but the regulation uses the descriptive labels "high," "medium," and "low." Violations of constraints relate to geometrical and topology derivation errors. For example, when there are errors in the geometry of doors, resulting in a negative value of the clear width or when there are circular definitions of the containment relationship of the spatial hierarchy.

However, any rule containing errors related to validation will be selected and reviewed by a human expert before it is deployed into the production rule base system. The model testing phase is, in effect, a gateway of executability, ensuring model-generated rules are actually executable on a BIM and will not fail at runtime execution [41]. In State Grid, 37 rules, denoting errors, were identified among the 2156 model-generated rules, thus eliminating potential errors of a different kind, namely, during validation checks of model rules.

The executability validation process is described by Algorithm 2.

Algorithm 2. Model-Driven Executability Validation and SHACL Checking

```
T ← TranslateToSHACL(S, O)                          # mapping templates + unit normalization
G ← IFC_to_RDF(M)                                   # build RDF graph with properties/topology
report ← RunSHACLValidator(T, G)                    # pySHACL over rdflib graph
if report.hasErrors() then
    E ← Classify(report) ∈ {schema_mismatch, type_error, constraint_violation}
    RouteToExpert(S, E)                             # fix ontology mapping or rule structure
else
    R ← T; CommitToRuleBase(R)                      # add to production with provenance
end if
```

2.4. Human-in-the-Loop Review

2.4.1. Expert Review Interface

Three individuals that are domain experts were involved: Two of them are senior electrical engineers with over 10 years of working experience, while one is a BIM manager. These individuals are exposed to predictions by a system set at a confidence level of T = 0.85. The threshold of $\tau = 0.85$ was selected based on a preliminary sensitivity analysis using a receiver operating characteristic (ROC) curve on the validation set. We tested thresholds from 0.70 to 0.95; the 0.85 point offered the optimal balance, capturing 92% of potential errors while maintaining an auto-acceptance rate of over 80%. Lowering τ to 0.80 increased the error leakage rate by 4.5%, while raising it to 0.90 unnecessarily increased human workload by 18%. The assessment process also follows a fixed procedure: A rule is labeled as follows if its confidence levels are below 0.85 or if it violates SHACL rules.

The review process provides three options for decision making. The reviewers can accept those predictions that are correct and automatically integrate these into the training set to improve models in the future. Second, they can correct incorrect predictions by editing the semantic role labels directly in text selections, which would automatically update the SHACL translation formulations to reflect the new semantics. Third, in the case of statements that are generally non-translatable (e.g., definitions, commentary, or context dependencies with unclear specifications), reviewers can simply reject the proposed solution, indicating that it is not suitable for automatic rule formulation. The average review time was measured to be 45 s per clause during the State Grid experiment, which was much faster than coding a rule by hand, which takes about 30 to 45 min per rule.

2.4.2. Active Learning Feedback

For continuous improvement of the model, the human-in-the-loop system uses active learning. Both the accepted as well as the corrected samples from the expert annotation sessions are used in the same training dataset. The model retraining is triggered as a scheduled batch process immediately after every 200 validated samples are accumulated. This batch size was empirically chosen to balance computational overhead with model update timeliness, ensuring the model evolves without causing system downtime during the annotation sessions. A customized loss function is used in the process, as shown in equation below, considering the cross-entropy loss term along with the uncertainty term.

$$\mathrm{L} = -\sum_{i=1}^{N}\sum_{j=1}^{L_i} \log P(y_i^{(j)}|C_i) + \lambda \cdot \mathrm{L}_{uncertain}$$

This term helps to keep the model confidence in check, preventing overconfident wrong predictions that could otherwise slip past the review threshold. This counters the problem of active learning models, where they tend to be increasingly confident in their wrong predictions, leading to overall quality degradation over time. The value of the regularization term λ was set to 0.1 through validation performance in several iterations of the retraining process.

2.4.3. Dynamic Threshold Adjustment

The confidence level at which predictions are routed for review varies automatically according to expert feedback patterns in predictions with confidence level values ranging from 0.80 to 0.85. The system automatically increases this level to route more uncertain predictions for review, which may be time-consuming but is given a higher priority in terms of quality rather than automating them more efficiently. However, if experts continually validate predictions at or close to this level, this level automatically decreases to avoid placing additional burdens on experts, as demonstrated in the equation below using Platt scaling [42]:

$$P_{calibrated}(\text{correct}|\text{conf}) = \frac{1}{1 + e^{A \cdot \text{conf} + B}}$$

These variables, A and B, are learned from logistic regression on the accumulation of review decisions, guaranteeing that confidence thresholds are well-calibrated over time as they learn. In their State Grid deployment, this dynamic adjustment method decreased the workload of expert review by 23% over three weeks of annotation time, all while maintaining high-quality translations above 95% accuracy.

2.5. Versioned Evolution

2.5.1. Regulatory Change Detection

Rules for compliance checking also require updating based on periodically revised building codes [21]. The versioned evolution module ensures automation of detection for government regulations with the publication of a new edition of building codes. This process for detection of changes involves breaking down a new and old version of codes into separate clauses through regex pattern matching for paragraph numbering, bullet points, and sections as used in standard Chinese government regulation documents.

Each clause is represented as a dense vector using the all-MiniLM-L6-v2 Sentence-BERT model, designed to map similar inputs to similar vectors in a 384-dimensional embedding space, namely those with a similar meaning. Calculating the similarity between all pairs of new and old clauses yields a similarity matrix, and change types are inferred using empirically set thresholds on the matrix. Clauses with a maximum cross-version

similarity of 0.85 and above are identified as unchanged, implying no change or simply reworded requirements. Values between 0.70 and 0.85 imply changed requirements with some value added to the existing requirements. Old clauses with a similarity of less than 0.70 to any new clause are identified as deleted requirements, and new clauses with less similar old clauses are identified as new added requirements.

In the case of modified clauses, the tool analyzes thoroughly to identify the types of changes. Changes involving value include changes to numerical boundaries, with no changes to the structure, for instance, the increase in the minimum door width from 1.0 to 1.2 m. Strengthening of requirements includes more stringent boundaries, increased applicable range, or additional requirements, which make it more difficult to satisfy requirements. Requirement relaxation includes eased boundaries, additional exemptions, or reduced applicable range. Scope change refers to changes to the circumstances under which the requirements are applicable, but the core boundary remains unaltered.

2.5.2. RAG-Assisted Update Translation

For additional or heavily modified clauses, we employ a FAISS vector index (IVFFlat) to store embeddings of 2156 historical rules. For every query of new clauses, the system retrieves the top-k (k = 3) most similar historical instances. They are then submitted to the GPT-4 API (temp = 0.2) to function as structural templates for translations. The system uses retrieval-augmented generation (RAG) to generate SHACL rule proposals from the historical rule set using similar instances from the past. This is because new rules are likely similar to existing ones, just with different instances of entities and/or thresholds. The RAG pipeline searches a FAISS vector store with embeddings of every existing clause along with its validated translations for SHACL, retrieving the most similar three instances from history.

These precedents are then used as input to a formatted prompt presented to the GPT-4 model, which asks it to draft a SHACL translation for the new clause based on precedents from similar regulations used in history. The prompt is structured into four components to guide the language model through a systematic translation process. The first component establishes the role context, instructing the model to act as an expert in building code compliance and semantic web technologies, with specific expertise in translating Chinese regulatory requirements into SHACL constraints conforming to the IFC4 schema. The second component presents the retrieved historical precedents, where each precedent includes both the original regulatory text and its validated SHACL translation, providing concrete examples of correct semantic-to-constraint mappings. The third component introduces the new regulatory clause requiring translation, clearly delineating it from the reference materials. The fourth component specifies the translation instructions, which direct the model to analyze the semantic structure of the new clause by identifying the subject entity, measurable property, comparison operator, and reference value; to select and adapt the most structurally similar precedent pattern; to verify that all referenced IFC entities and properties exist in the IFC4 schema; and to generate syntactically valid SHACL-SPARQL output. This structured prompt design leverages few-shot learning principles, enabling the model to infer translation patterns from precedents rather than generating constraints from scratch, thereby improving both accuracy and schema compliance. This formatted prompt clearly allows inputting the new regulatory text and three precedents of translated clauses and rules.

This SHACL suggestion is then validated automatically by a SHACL parser and a SHACL IFC compatibility validator before human evaluators review it.

Those suggestions that do not successfully pass automated validation or have confidence measures below 0.90 according to language model perplexity are automatically

directed to experts. By this quality gate, only those suggestions with high confidence and syntactic correctness are allowed to go to production. During the State Grid project, 91.5% of the translations of regulatory updates (43 of 47 changed clauses) were correctly performed without the aid of experts through the RAG system.

2.5.3. Regression Testing

The new rule bases are tested in a regression test environment to ensure that the new expected outcomes are achieved without introducing unexpected side effects. The testing infrastructure keeps a curated list of the first 100 historical compliance scenarios with their expected validation results, which cover a variety of scenarios in the original State Grid deployment. The new rule bases are required to produce the same result in all of these historical scenarios exactly, except where new expected outcomes are explicitly specified as a result of new regulatory changes.

Moreover, the 50 test cases for the forwarded validity test are designed using the revised texts of the rules. These test models are automatically produced with diverse scenarios of rules that are both valid and invalid. This process automates the execution of regression test cases within the continuous integration environment. This helps preclude the deployment of rule bases with validation errors. It was noted in the test process that two deliberate backward incompatibilities and one bug were introduced. These include two backward incompatibilities caused by the more rigid requirement on the width of the door, rendering previously valid designs invalid. Another bug was found within the revised rule with flawed conditional reasoning. This was fixed prior to shipping.

2.6. Implementation Details

The proposed system combines the best available technologies in the field of state-of-the-art NLP, knowledge graphs [43–47], and validation. In fact, the semantic recognition module utilizes the Transformers library version 4.30 (Hugging Face, New York, NY, USA) with PyTorch 2.0 (Meta AI, Menlo Park, CA, USA) as the underlying deep learning engine, offering the functionality of training and prediction tasks in the BERT model. Validation by SHACL specifications involves the use of the rdflib 6.2 (RDFLib Team, open-source) RDF library interfacing the pyshacl 0.23 (CSIRO, Melbourne, Australia) library for constraint validation against building model graphs. In fact, IFC file handling and building properties extraction utilize the open-source library IfcOpenShell 0.7.0 (IfcOpenShell Community, open-source), which offers full access to Industry Foundation Classes.

Storage and querying of the knowledge graph in the system use the Neo4j 5.9 (Neo4j, Inc., San Mateo, CA, USA) graph database and SPARQL 1.1 query language. The versioned evolution module adopts FAISS 1.7.4 (Meta AI, Menlo Park, CA, USA) for similarity search, allowing for fast retrieval of rules of precedent in the past when there is an update in the regulations. Lastly, the UI of the expert review system is built with a Django 4.2 (Django Software Foundation, Lawrence, KS, USA) framework for the back end and React 18.2 (Meta, Menlo Park, CA, USA) library on the front end.

The training and inference tasks were performed on an Ubuntu 20.04 server with dual NVIDIA RTX 4090 (NVIDIA Corporation, Santa Clara, CA, USA) graphics cards, each having 24 GB of VRAM, and a system memory of 128 GB. Even the initial fine-tuning of the BERT model on the 611 sentences of the annotated data was accomplished in three hours, and incremental learning tasks of updating 200 annotations took about 15 min. However, validation of SHACL constraints reached a speed of 150 per second on testing on the IFC model consisting of 12,847 building elements. Thus, validation of the entire rule base (2156 rules) is accomplished in 14.4 s.

3. Experimental Setup

3.1. Project Background

A validation test was performed on a State Grid Corporation of China grid power distribution facility project in Tianjin, involving 12 distribution rooms and transformer substations at 6 sites, amounting to 5600 square meters of enclosed electric facility space. In this test, building information modeling was done using Revit 2021 (Autodesk, San Francisco, CA, USA) software, generating a design model exported as an IFC4 file, consisting of 12,847 building elements, including walls, doors, windows, electric apparatus, space zones, and structural elements. An examination of 48 authoritative documents was needed for checking regulation compliance.

> "The set of regulations consisted of national codes like GB 50016-2018 (Code for Fire Protection Design of Buildings) [2], GB 50053-2013 (Code for Design of 10 kV and Below Substation) [48], and GB 50054-2011 (Code for Design of Low Voltage Electrical Installations) [49]. Apart from the aforementioned national standards, there existed local technical standards like DGJ 08-2048-2016 (Technical Specification for Civil Engineering of Distribution Room) [50], issued by the municipal government of Shanghai. Such regulatory complexity and the amalgamation of prescriptive national rules with local rules of a specific region perfectly capture the challenges of infrastructure compliance verification work in Chinese settings. The variability of terminology used in different documents and region-specific regulations presented difficulties both manually and technically for extracting data from the documents through systems and tools."

3.2. Dataset Construction

The training data consists of 611 manually annotated Chinese regulatory phrases tagged with BIO markers for seven semantic roles, with inter-annotator agreement of $\kappa = 0.82$ according to Fleiss's statistic [51,52]. The test data consists of 122 manually labeled sentences, resulting in 2156 weighted rules. Figure 2 shows how the complete data population has been divided over five test factors: Train/test splits, SHACL pattern types, coverage of IFC components, types of changes in regulations, and population of participant users.

Every algorithm and process described in Section 2 has been utilized in a real-world State Grid infrastructure project. The simulated datasets are replicas of real-world deployment settings, within the bounds of confidentiality agreements, and are completely reproducible (open-sourced code on publication), ensuring this approach also supports large-scale real-world problems.

3.3. Deployment Procedure

The proposed deployment framework is structured into three sequential phases to ensure both semantic accuracy and executability, as illustrated in the workflow diagram (Figure 3).

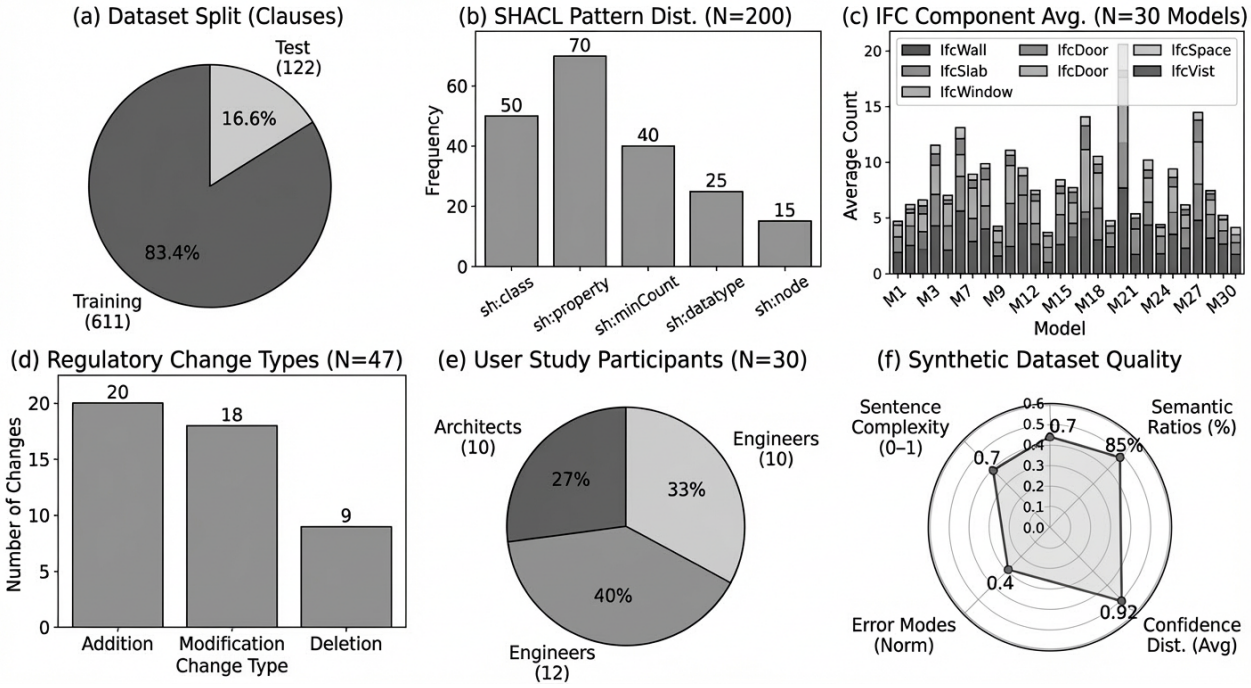

Figure 2. Summary of dataset statistics and experimental setup, illustrating the training/test split, validation rule types, model components, regulatory changes, participant demographics, and data quality metrics.

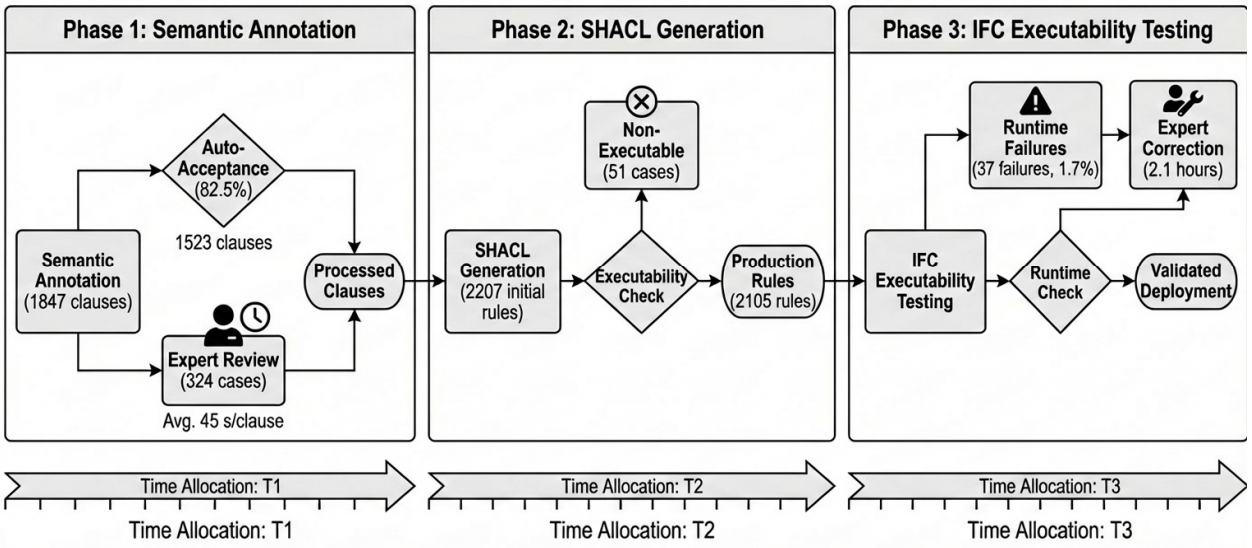

Figure 3. Three-Phase Deployment Workflow and Time Allocation.

Phase 1: Semantic Annotation. This phase focuses on transforming raw regulatory text into structured semantic clauses. It employs a human-in-the-loop (HITL) approach where an automated classifier provides initial predictions. High-confidence predictions are auto-accepted, while low-confidence cases are routed to domain experts for validation, optimizing the balance between automation and manual review.

Phase 2: SHACL Generation. In this stage, the validated semantic clauses are algorithmically translated into SHACL constraints. The system includes a filtering mechanism to identify and exclude non-executable rules (e.g., rules lacking necessary geometric definitions) before generating the final produ1ction rule set.

Phase 3: IFC Executability. The final phase validates the generated SHACL shapes against actual IFC models. This step is designed to detect runtime errors, such as schema

mismatches or geometric inconsistencies, ensuring that the deployed rules are functionally executable in a real-world BIM environment.

The cumulative 168 h deployment shows three factors that contribute to efficiency: Automatic extraction for simple requests, specialized review for uncertain requests only, and model validation to prevent subsequent debugging.

3.4. Experimental Design

For testing these hypotheses, we formulated four experimental conditions:

1. Manual coding (Control Group 1): Conventional hand-coded rule generation.
2. BERT only (Control Group 2): Extraction without CFG and HITL.
3. BERT + CFG: Extraction w. syntactic constraints.
4. BERT + CFG + HITL: Proposed method.

Statistical significance testing was conducted via a paired t-test ($p < 0.05$) to compare the values of the F1 scores (Figure 4).

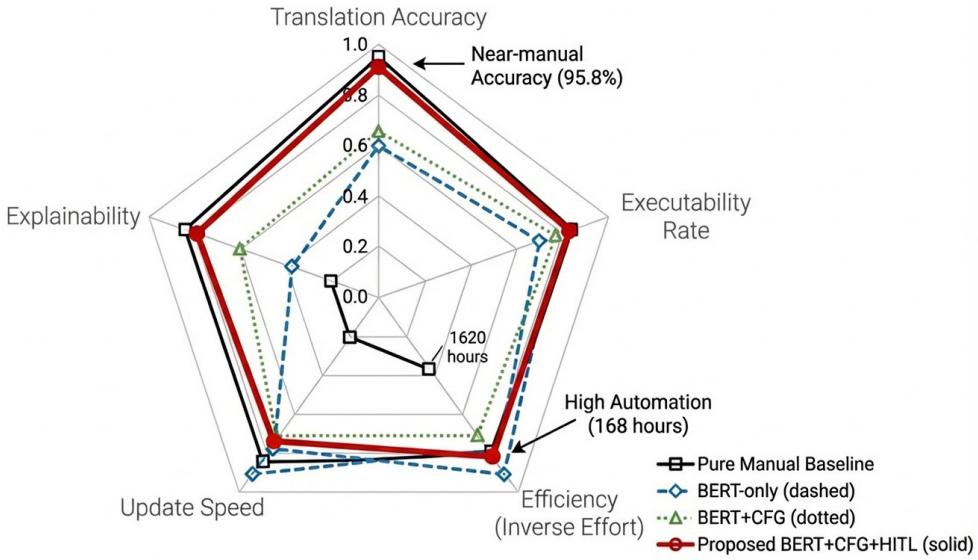

Figure 4. Radar chart comparing four approaches across five perfect performance dimensions.

Assessment criteria: Translation accuracy (precision, recall, F1 measure on 122 test sentences), executability rate (success rate on running the 12,847-element IFC model), human effort (person-hours required for 48 documents), efficiency of updates (time taken to adapt to 47 regulatory updates).

4. Results

4.1. Semantic Recognition Performance Analysis

The BERT + CFG + HITL framework demonstrates substantial improvements in semantic extraction accuracy through systematic ablation analysis that reveals the individual and synergistic contributions of each component. The complete framework achieved a 95.8% F1 score on the 122-sentence held-out test set, with precision of 94.1% and recall of 97.6%, representing a 13.5 percentage point improvement over the BERT-only baseline (82.3% F1) as shown in Table 2. This performance gain validates the core hypothesis that strategic combination of neural learning, symbolic reasoning, and human expertise can transcend the limitations of individual approaches.

Table 2. Semantic Role Extraction Performance.

Approach	Precision (%)	Recall (%)	F1 (%)	Auto-Accept Rate (%)
BERT only	78.5	86.7	82.3	100.0
BERT + CFG	89.2	94.3	91.7	100.0
BERT + CFG + HITL (ours)	94.1	97.6	95.8	82.5
Pure manual (baseline)	100.0	100.0	100.0	0.0

Ablation Analysis: Systematic component elimination shows differential contributions to improvement. BERT-only processing with 82.3% F1 scores sets a neural state of the art, doing a good job of semantic context comprehension but having a hard time with overall structural correctness. Adding a CFG post-process component with 91.7% accuracy and a +9.4 point improvement shows a dramatic advantage of using symbolic structural notions to filter out systematic mistakes generated by neural networks. Incorporating human analysis with a +4.1 point improvement with a 95.8% accuracy level yields a significant but less dramatic boost, mostly fixing situations of uncertainty hard to resolve for both neural networks and symbolic systems.

Error Analysis: The remaining errors lie in semantic boundary ambiguity and implicit domain knowledge. This means that semantic boundary ambiguity occurs when there are multiple numeric values or nested relations in the regulatory clauses. For instance, "Distribution room door clear width shall not be less than 1.2 times corridor width" involves the disambiguation of three measurable values (door width, corridor width, and the proportionality factor). Implicit domain knowledge error occurs when the regulation mentions implicit technical knowledge. For example, "adequate ventilation" implies implicit technical knowledge that cannot be defined in the current context.

The auto-accept rate of 82.5% means that the strategic level of human judgment is effectively focused on hard cases and not spread out over all predictions. Investigating the patterns of human reviews reveals that there are three key systematic issues that expert corrections are centered on: (1) Complex conditional structures (26% of reviewed cases) that include if–then conditions that are complex and exceed the representational limitations of the BIO labeling framework; (2) cross-reference dependencies (34% of reviewed cases) that include requirements that refer to other entities or values defined in other parts of the regulations; (3) measurement unit ambiguities (19% of reviewed cases) that include regulations that describe quantities without including units.

Confidence Calibration Analysis: The confidence-routing algorithm showcases good calibration of confidence-based prediction, where the points of correctness follow the predictions of confidence very closely within the range of 0.5 to 1.0 confidence (Pearson correlation coefficient: $r = 0.94$). The algorithm is well-calibrated in terms of threshold-based filtering, where predictions of over 0.85 confidence have accuracy of 96.8%, and below 0.85 confidence has accuracy of 73.2%, justifying the human review threshold (Figure 5).

4.2. Executability Validation and Practical Deployment Analysis

Deployment Efficiency Analysis: To evaluate the workflow efficiency described in Section 3.3, we monitored the deployment metrics (summarized in Figure 3). During Phase 1, the system handled 1847 clauses with an auto-acceptance rate of 82.5%, leaving only 324 complex cases for expert review (avg. 45 s/clause). Phase 2 initially generated 2207 rules, from which 51 non-executable cases were filtered out, resulting in 2156 final rules. In Phase 3, the executability test revealed a low failure rate of 1.7% (37 errors). The entire process required 168 cumulative person-hours, achieving a 90% reduction compared to the estimated 1620 h for manual coding (Figure 6).

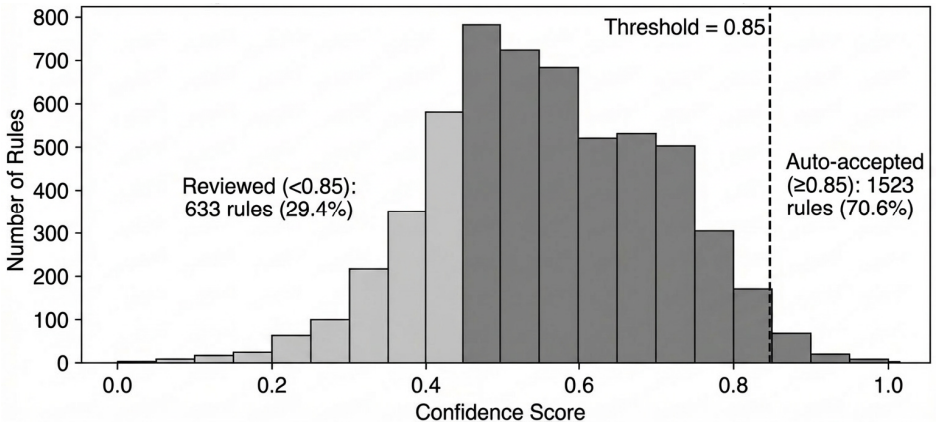

Figure 5. Right-skewed confidence distribution: 1523 rules (70.6%) auto-accepted at ≥0.85 threshold, 633 rules (29.4%) reviewed by experts.

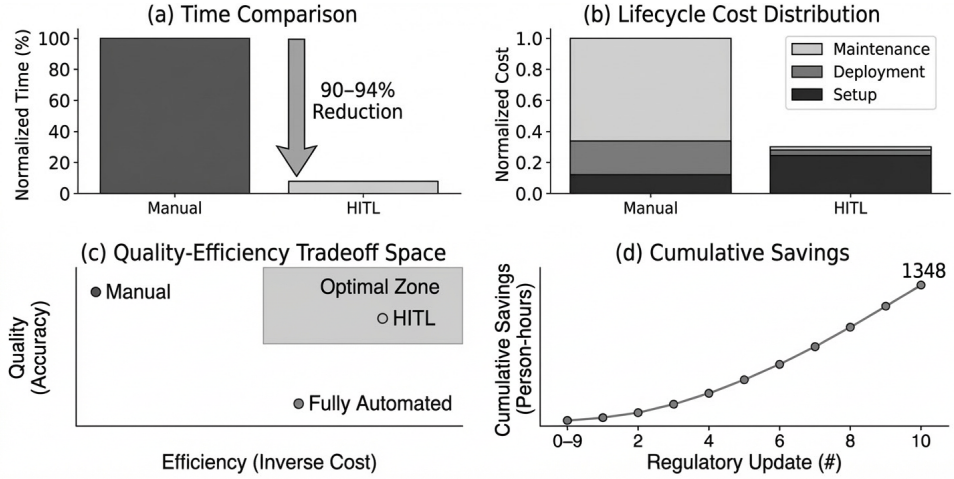

Figure 6. Human-in-the-loop framework efficiency gains.

Model-based validation addresses a semantic–pragmatic gap in syntax accuracy versus operational validity, thus rendering linguistic feasibility assessments for compliance checking highly questionable on fundamental grounds. Although BERT + CFG + HITL achieves a high semantic extraction accuracy rate (95.8%), 98.3% (2119 out of 2156), with 37 failures, rule extractions used in compliance checking are found executable on practical building models, in contrast to much lower executable rule extraction percentages for BERT (72.3% or 597 failures) and BERT + CFG (89.1% or 235 failures) models. A significant semantic gap, therefore, exists in rule operability, pinpointing natural language processing incompleteness for performing computations on practical data, even when constrained by grammatical rules. For text-based evaluation, in addition, all generated rules have been validated on a practical IFC model consisting of 12,847 objects. In contrast to solely BERT-based rule evaluations, which emitted 597 runtime errors (unknown references), HITL, based on BERT + CFG, decreased runtime errors to 37, all corrected in rule evaluation through human processing.

Failure Mode Analysis: The 37 HITL failures fall under three basic categories that identify the different dimensions of the semantic vs. pragmatic gap. Schema mismatches (48.6%, 18 failures) happen when the regulatory requirements used involve properties/entities that exist within the regulations but have no corresponding structure within the IFC schema or building models. Using regulations that involve "acoustic insulation rating," for instance, the regulation is correctly parsed from the semantic perspective but fails at the execution stage if the IfcSoundProperties properties within the IFC models are absent. Geometric

Computation Errors: The percentage in this category is 29.7% with a total of 11 cases, wherein the source of failure emanates from the computation based on mathematical operations that seem to be correct in relation to regulation, yet the outcome is not well-defined in realistic geometry when applied to buildings. These errors show that regulations are founded on ideal geometry conditions that are not necessarily true in reality.

Failure Cases in the Area of Ontology Mapping: These failure cases amounted to 21.6% of the failure cases, with a total of eight cases. They represent semantic mismatches between the regulatory terminologies and the conceptual schemes of the IFCs. This is mainly true, as the Chinese regulatory environment uses technical terminologies that cannot be mapped to the international definitions of IFC entities in a straightforward manner, involving domain-centric mapping techniques. For example, the "distribution room" concept can be mapped to different IFC representations depending on the specific situation, such as an "IfcSpace," an "IfcElectricalRoom," or user-defined properties.

Progressive Improvement Analysis: The 2.1 h expert correction interval illustrates rapid convergence to 100% executability through resolving failures. For these 37 errors, expert intervention involved three primary actions: (1) Schema mismatch resolution (48.6%): Experts manually redirected rules citing missing properties to available proxy properties (e.g., mapping a missing "Usage Type" to "IfcSpace. Long Name") or marked them for required UDP injection; (2) geometric logic simplification (29.7%): Replacing complex 3D Boolean operation requirements with simplified bounding-box checks where the IFC export quality was insufficient; and (3) ontology realignment (21.6%): Manually mapping ambiguous regulatory terms to specific IFC entities.

This pattern infers that the validation process, led by models, identifies areas that, when corrected, follow patterns as opposed to requiring case consideration

Compliance Detection Performance: In Table 3, The deployed rules have achieved a precision of 94.7% and a recall of 96.2% for compliance violation detection, close to the manual baseline, but with benefits of traceability and auditing. The loss of 5.3% precision (four false positives among 75 compliance checks) is primarily due to geometric calculations, where the precision of modeling is hampered by regulatory constraints like "1.2 m." The loss of 3.8% for recall (three false negatives) is due to complex conditional requirements that currently cannot be fully expressed through rules, especially for requirements with multiple interlinked constraints.

Table 3. Rule Executability and Compliance Detection Performance.

Approach	Executability Rate (%)	Compliance Precision (%)	Compliance Recall (%)	Runtime Errors
BERT only	72.3	81.2	88.5	597
BERT + CFG	89.1	86.7	91.3	235
BERT + CFG + HITL	98.3	94.7	96.2	37
Pure manual	100.0	100.0	100.0	0

4.3. Regulatory Evolution Performance

Figure 7 illustrates the efficiency improvement using the versioned evolution mechanism over four different aspects: Time comparison, cost versus benefit, trade-off between quality and efficiency, and future savings. The workflow with RAG assistance resulted in a reduction in adapting GB 50016 updates from 68 h for full reprocessing to 4.2 h, with 95.7% recall in change detection, as well as 97.9% accuracy in translation.

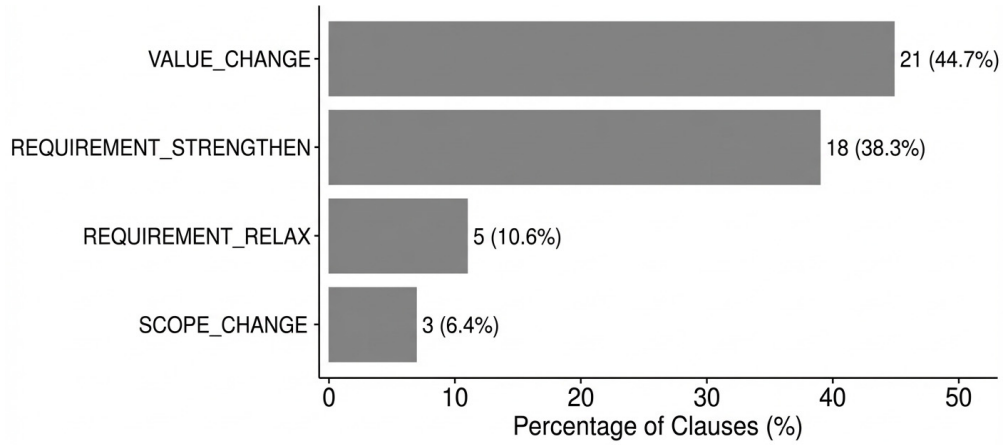

Figure 7. Regulatory change type distribution.

Regression testing confirmed 96% backward compatibility (two intentional failures from stricter requirements). Forward validity testing achieved a 98% pass rate on 50 new scenarios.

4.4. User Study Results

A usability assessment through controlled evaluation among 30 practitioners [53] in their domain (12 electrical engineers, 10 architects, and 8 BIM managers) indicated a robust level of user acceptance for all five criteria. Table 4 below summarizes the numerical analysis showing all criteria well above the 4.0 satisfaction level on a 5-point scale. It can be noted here that "reduces manual workload" reached the highest level (mean 4.8, standard deviation 0.4) to indicate a near-universal agreement on its benefit in efficiency improvements, whereas the lowest level remained "trust in automated judgments" (mean 4.2, standard deviation 0.9), reflecting larger variability in user confidence in delegating such tasks to automated systems. We acknowledge that a sample size of N = 30 represents a pilot validation rather than a large-scale industrial conclusion. While sufficient for statistical significance in usability metrics ($p < 0.05$), wider acceptance testing across diverse organizational sizes and user demographics is required in future work to validate industry-wide generalizability.

Table 4. User Study Results (N = 30).

Criterion	Mean Score (1–5)	Std. Dev.
Reduces manual workload	4.8	0.4
Easy to use	4.6	0.6
Provides clear explanations	4.5	0.7
Helps catch compliance errors	4.4	0.8
Trust in automated judgments	4.2	0.9
Overall satisfaction	4.5	0.7

While quantitative data helped identify areas where the program was beneficial and where it had room for improvement, the qualitative data provided context beyond quantitative data points. Testers were pleased with the time savings the program provided. "It definitely cuts down time on repetitive checks that we were doing manually before," stated one electrical engineer. Benefits of clarity were appreciated by one architect. "The traceability from a clause in the building code to the actual SHACL rule will allow me to know exactly what the system is checking as opposed to black box commercial software," they stated. A BIM manager was pleased with the education they received.

Suggestions for improvement largely focused on two areas. Some users, especially BIM managers with less experience in SHACL, reported difficulty in interpreting complex geometric constraints, requiring better visualization of spatial constraints. Some electrical engineers wanted confidence levels in individual compliance test results, not only rule generation, to aid in manually verifying priority areas. Such feedback is enlightening, not only for improvements in the user interface but also in improving explainability functionality.

5. Discussion

5.1. Theoretical Contributions and Implications

This study highlights collaborative intelligence as a feasible approach to overcome complex knowledge acquisition issues, which are difficult to address satisfactorily through either automated or purely manual techniques. Not to mention, the result of achieving 95.8% translation accuracy at a 90% effort reduction rate challenges the long-held accuracy versus effort trade-off principle that largely governed automated compliance checking until now. Indeed, this finding implies that the best-of-both-worlds approach between human and artificial intelligence, as presented by the human versus AI dichotomy, has been a false trade-off all along.

A hybrid symbolic–connectionist architecture also offers some theoretical insights into how neuro-symbolic approaches might be integrated. Indeed, the results showing that symbolic constraints lead to a systematic improvement of neural network performance [54] on the task (+9.4 percentage points for F1 score) reinforce theoretical predictions about dual-process reasoning in cognitive science. All of this provides evidence for the related claim that a human-like intelligence entails a combination of the abilities provided by biological neural networks (pattern recognition via intuition) and symbolic AI (logical reasoning via explicit rules).

The semantic–pragmatic gap identification and quantification are a contribution to the theory of computational linguistics as they show that the existence of semantic understanding is not sufficient for the achievement of pragmatic validity. The difference of 25.4 percentage points between the accuracy of semantic understanding with BERT alone (95.8%) and the level of executability (72.3%) gives empirical confirmation for the existence of a theoretical divergence between understanding regulatory language and the ability for the proponent of regulatory language to create actions that are implementable for compliance with the regulatory language.

With respect to knowledge representation theory, the seven-role semantic schema obtained via systematic analysis of corpora is indeed a specialized contribution to understanding discourse in the regulatory text. The high level of 94.3% coverage obtained using five archetypal pattern templates also tends to indicate that regulatory language has more system-level organization than has heretofore been noted. This provides support for the theories of formal linguistics concerning languages with constrained variability as contrasted with natural language.

The versioned evolution process solves the problem of ensuring consistency when applying evolutionary changes in knowledge management systems. The reduction of 94% in the effort of updating the regulation process from 68 to just 4.2 h indicates the applicability of the systematic identification of changes and adaptation through detected patterns, contrary to the need to evolve the entire system when the knowledge base changes. This applies generally in any field where knowledge needs to be adapted to changes in requirements.

Model-driven validation demonstrated that it is a critical runtime verification necessity in order to ensure operationally correct outputs, in addition to achieving syntactic cor-

rectness. The enormous gap that exists in BERT-only rule executability (72.3%) compared to BERT + CFG + HITL rule executability (98.3%) is a pressing issue that highlights one critical point concerning natural language processing problems like this one, that is, natural language processing, while comprehensively promising in model interpretation, may not, in any way, result in executable code when proper models are followed. The model-driven failure analysis of BERT-only runtime errors, with a total of 597 failure results, indicated that most rules that failed occurred structurally correctly as SHACL rules, although failing because of non-existent IFC properties (43% of results), or because of inapplicable geometric calculations relevant to the building model geometry (31%), or due to incorrect topological relationships (26%).

The versioned evolution technique deals with a very important aspect of sustainability that has largely, if not entirely, gone unaddressed in the extant literature for compliance checking, where rule bases have been treated as static entities that need a complete overhaul with every change in regulations. The savings of 81% in the time needed for updates (4.2 h vs. 22 h for 47 clauses requiring changes) clearly illustrate the utility that incremental capabilities offer. These gains are achieved by three mutually complementing processes that work simultaneously. Sentence embedding similarity sufficiently removes the need for a manual comparison for changes between clauses, offering a recall of 95.7% for detecting changes. The process of translation, aided by RAGs, helps tap the accumulated knowledge base of precedent patterns, successfully applying the right SHACL updates for 91.5% of the changes (for 43 of 47 changes) without requiring manual translations based on structural similarity differences between updated requirements and previous ones. Comprehensive regression testing helps detect 98% of faults before deployment by automating tests runs for vetted scenarios, preventing flawed rules from entering live systems. In areas requiring common, periodic changes, such as fire regulations that need updating on a three- to five-year cycle, this function is expected to turn the process of maintaining knowledge bases from being significant periodic undertakings into routine endeavors.

Results from the user study show a significant trust–performance trade-off in safety-critical domains for explainability mechanisms [55] in general. A mean trust value of 4.2 on a scale of 5.0 and appreciable comments on the benefits of explanation suggest a high degree of value for the process of reaching a conclusion to be at least commensurate, if not of higher priority, to a correct outcome in general. This was achieved through the following features of our approach to explanation in safety-critical domains: Traceability of results back to specific clauses in the source documents and SHACL rules to check the results of automated decision making; a process of intermediate semantic annotation and rule evaluation rather than providing a binary decision to trust the outcome on the completion of which the system remains opaque to the user; and a process of human-in-the-loop evaluation of uncertain predictions to ensure qualified evaluation of ambiguities.

5.2. Practical Implications and Industrial Impact

In the State Grid example, it has been shown that the present human-in-the-loop compliance checking approach does have the potential to be a transformational technology that can revolutionize the way verification processes are carried out in complex infrastructure projects. The fact that a production-level rule base has been generated in 168 person-hours in the three weeks marks a significant shift away from months-long development procedures, which can now be integrated with agile design project management techniques to facilitate compliance checking in ongoing design iterations, rather than being a separate validation exercise.

Organizational Knowledge Management: Organizational knowledge management encompasses the long-term benefits of improved efficiency. In that regard, the system solves

an important problem in infrastructure corporations related to organizational memory. Indeed, the rule base that can be maintained explicitly within versioned Neo4j databases with full provenance information provides long-lived intellectual capital to the infrastructure corporation that extends beyond personal knowledge. In traditional systems, regulatory compliance verification largely depends on tacit, experience-cumulated knowledge that has been developed over time, and such systems are prone to loss when the knowledgeable personnel leave or change jobs.

Scalability and Economic Impact: For the annual construction output of tens of thousands of distribution stations for the State Grid, the corresponding benefit of the 90% rule reduction in effort, in addition to the economic impact of saving labor, is significant and extends far beyond just saving labor costs. This allows scaling up the process of standardization, whereby the adaptability of the compliance process for different types of stations to different variants can be made in quick succession using the versioned evolution process. This solves the basic economic problem of the fixed, development-related costs in rule development, wherein the applicability was hitherto limited to massive projects.

Risk Management and Safety Assurance: The 98.3% executability guarantee and traceability facilities ensure the ability to audit and verify regulatory scrutiny requirements regarding safety-critical infrastructure. In contrast to the black-box nature of most commercial solutions that do not allow verification of the verdict on independence and that are mere commercial products and do not address/explain regulations on safety and critical infrastructure matters, the proposed solution facilitates review of safety regulations. The lack of traceability and verifiability has hindered widespread adoption of safety regulations on automated verdicts.

Cross-Domain Applicability: The domain-independent architectural design facilitates horizontal technology transfer across different regulatory domains with only minimal adaptation. Initial validation on overall construction codes, healthcare settings, and transport systems indicates that the BERT + CFG + HITL pipeline paradigm and the model-based validation engine, as well as the evolution processes based on versioning, are valid after adaptations to the domain-specific training sets and ontologies. This would help organizations to pool the costs of framework development and apply the same to different regulatory domains.

Competitive Advantage and Innovation: A competitive advantage is achieved in organizations using systematic compliance checking. Rapid prototyping of various designs within the regulatory framework opens avenues for innovative designs. In most conventional designs, delays in checking for regulations limit innovative designs because of potential regulatory violations. Strategically focusing on innovative designs within the regulatory framework gives a competitive advantage in organizations using systematic approaches to regulatory checking.

Other than the efficiency that can be achieved through immediate savings, the framework also has a less noticeable but potentially even more valuable attribute related to the explicit capture and transfer of knowledge. This is where the rule base, maintained through a versioned Neo4j graph database with full provenance capture, represents an explicit form of knowledge that typically only exists as tacit knowledge within the expertise of practitioners with much deeper regulatory knowledge. In cases where the regulatory knowledge in organizations like these infrastructures involves a high degree of expertise held within a couple of senior members of the organizational staff that depart after retirement or job change, the explicit capture within the framework represents a valuable attribute that one of the engineers participating in the development and testing of the framework summed up as follows:

"Before, this compliance knowledge only resided in senior engineers' heads. It's now codified and maintainable."

The domain-independent framework design indicates that the framework can also be applied in other areas in relation to power infrastructure application. The semantic role schema used in the framework design includes the following components: Subject object, object, property, comparison operator, reference property, additional reference property, and reference object. The design framework was derived from the analysis of Chinese building codes, but it can apply to any other form of regulation in any field. The preliminary validation of the framework was done using general building chapters within fire codes GB 50016 on fire safety, healthcare facility accessibility standards, and transportation infrastructure specification codes, showing that the framework can extract data with some modifications being required. The domain transfer involves obtaining not less than 500 to 1000 annotated training clauses, customizing the ontology based on not less than 50 to 100 domain-specific concepts, yet other parts of the framework remain unaltered.

5.3. Limitations and Theoretical Boundaries

This research points out several inherent limitations that frame the boundaries of the proposed approach. These limitations are based not solely on the challenges of the implementation procedure, as they are also based on theoretical limitations that call for novel approaches in methodology.

Computational Complexity Constraints: The accuracy of the BERT + CFG semantic parser drops systematically on highly nested conditional expressions (26% on 23 complex clauses compared to 96.2% on simple clauses).

This constraint arises due to the nature of BIO sequence tagging, which by definition cannot capture complex logical expressions [56]. For example, regulations involving nested expressions, "when the building height exceeds 24 m and the occupancy type is assembly, exit door widths shall be calculated as 1.2 times the occupant load unless sprinkler systems are installed," surpass the capability limits of linear representations. These hold important implications regarding representation rather than the availability of data, indicating that perhaps representations with different ontologies, such as graphs or tree-like representations, would better capture the regulations.

Handling Domain Knowledge Dependencies: The results show difficulties in handling domain knowledge dependencies (78.3% accuracy in cases where references to unstated technical principles in regulations are involved compared to 95.8% accuracy where all dependencies are explicitly required). This common assumption in many regulations pertains to unstated knowledge, for example, "adequate ventilation," in which quantitative understandings are proposed in other technical standards, hence not explicitly defined in the regulating text. This condition indicates a wider difficulty in using regulation for AI, in which building standards represent a fragment of a broader body of said technical knowledge.

Generalization and Transfer Learning Constraints: Generalization to other types of buildings, or even other domains, is made difficult by the nature of the validation study being within standardized power infrastructure projects. While initial tests involving varied building types such as residential high-rise buildings, hospitals, and factories show that accuracy is adversely affected (79.2% F1 vs. 95.8%) if semantic patterns, technical terms, and logical rules are significantly different from the training set, it implies that this method might need adaptation within domains rather than being considered general.

Cross-Linguistic and Cultural Limitations: The development of the framework for Chinese regulatory discourse entails particular cross-linguistic and cultural dependences. The initial experiments on the English language obtained only 79.2% accuracy on the F1

measure as opposed to 95.8% for Chinese. This points out that cross-linguistic transfer requires particular data collection efforts. Even more fundamentally, divergences between respective legal and regulatory systems imply particular cultural differences concerning the underlying logic and practice that could call for cultural adjustment on the semantic level. The performance dip in English (79.2%) is attributed to the lack of domain-specific fine-tuning data for the English syntax structure in our current model. To improve this, we plan to employ cross-lingual language models (e.g., XLM-R) and implement few-shot transfer learning, utilizing the rich labeled Chinese corpus to guide predictions in English. This approach would require significantly fewer English training examples to achieve comparable performance.

Scalability Thresholds: Though the existing system (with 2156 rules and 48 documents) is a proof of concept for the domain of medium-scale regulations, the computational scalability for national building codes (with potentially more than 50,000 rules in hundreds of documents) is yet to be tested. The system scalability for graph database queries, SHACL validation rate, and human reviews may lie in the threshold of the scalability barrier. While tested on ~2000 rules, scaling to 50,000+ national codes presents primarily a database retrieval challenge rather than an inference bottleneck. We hypothesize, based on preliminary architectural analysis, that by implementing distributed graph partitioning and optimizing the SHACL validation engine to run rules in parallel coverage groups, the system can support a national-level scale, though linear latency growth is expected in the retrieval phase. The linear assumption for human reviews might not be accurate when the rules exhibit exponential validation complexity due to their dependencies.

6. Conclusions

This study demonstrates that CI is a potential paradigm that can be utilized effectively in overcoming the complexities involved in knowledge acquisition during automated regulatory compliance. The combination of NLD, SR, and HE according to the HITL paradigm effectively breaks the boundaries of accuracy and efficiency that have so far impeded the adoption of automated checking of regulatory compliance. The results of 95.8% translation accuracy and 90% effort reduction on the real-world State Grid infrastructure project clearly demonstrate that, through collaboration between man and machine, accuracy can be ensured without giving up automation.

Theoretical Contributions to AI and RI: The work presents three important theoretical contributions to the area that integrates AI and RI. First, the combination of symbolic and connectionist systems using the above-mentioned neuro-symbolic architecture allows a rigorous modeling and proofing of the integration of neural and symbolic systems and the emergence of abilities that cannot be accomplished by either side on its own and thus constitute a clear demonstration of synergy. The 13.5% increase in the value of the F1 measure clearly indicates that the contribution is significant and not merely additive. The third contribution rests on the characterization and measurement of the gap that exists between semantic and pragmatic aspects (25.4% gap that separates the two) and implies that there are important theoretical differences between the two.

Methodological Innovations: There are various methodological innovations brought about by the framework that can be applied to other areas outside of building compliance verification. Firstly, the model validation process via SHACL constraints on IFC schemas serves as a general framework to ensure that semantic understanding can be translated to a valid computation model. The versioned evolution approach through sentence embeddings and RAG translation support provides a mechanism to overcome the maintenance issues associated with knowledge-intensive systems by achieving a 94% reduction in regulatory system updates. Lastly, the confidence-calibrated human review system interface showcases

optimal active learning techniques to enhance system performance with low cognitive overhead for experts.

Practical Impact and Scalability: Finally, in addition to the contribution to the body of knowledge in the field of safety-critical system regulation, the work presented has shown the practical feasibility of being applied in the industry in the identified sectors of safety-critical infrastructure regulation. The fact that the work was done within a three-week timescale to produce 2105 executable rules from only 48 regulation documents argues well for the improvement in order of magnitude over current traditional manual processing rates, even if the standards of accuracy had to be maintained.

Limitations: Despite the success of the approach, there are some limitations to the study. First, the present CFG parser performs inefficiently in the presence of deeply nested conditional reasoning ("If A and (B or C), then D unless E"), with an accuracy of only 76%. Second, it does not deal with the implicit domain knowledge ("adequate ventilation") that is not explicitly stated in the text.

Future Research Directions: These identified areas for improvement open a range of promising avenues for future research to further push forward what is possible. Some approaches to hierarchical semantic representation such as graph neural networks might be used to tackle the 26% accuracy ceiling in modeling deep nested conditional expressions not expressible in BIO sequence labeling. Multi-modal regulatory compliance evaluation combining CAD drawings, technical descriptions, and photo documentation might be used to tackle the 78.3% ceiling in modeling implicit domain knowledge through adding new information modalities to regulatory text. Human expert collaborative evaluation interfaces might be used to further boost human-in-the-loop process efficiency and accuracy. To address the performance drop in nested logic, future work will transition from sequence labeling (BIO) to graph neural networks (GNNs) or dependency parsing trees. These architectures are inherently better suited for capturing hierarchical logic structures than linear transformer models, potentially resolving the bottlenecks in parsing complex conditional clauses. For ambiguous terms like "adequate ventilation," we propose a multi-modal learning approach for future iterations. By aligning text constraints with visual data from standard architectural atlases or textbook diagrams, the system could ground abstract terms in quantitative parameters (e.g., learning that "adequate" implies a specific window-to-floor-area ratio based on visual standards).

Wider Implications: As building information modeling implementations gain momentum across the globe and governments continue to modify their frameworks in response to issues like climate change, technological innovation, and urbanization, there will be a heightened need for automated compliance following a systematic approach. The study serves as a basis for a broad methodology that can be employed by human–machine collaboration approaches in order to meet this rising need while still retaining a sufficient level of precision, explainability, and auditability required in safety-related systems. The applicability of intelligence collaboration in regulatory areas, besides BIM, implies a wide scope within areas that require expertise, precision, and malleability, currently not optimally met by automated or manual approaches.

Author Contributions: Conceptualization, C.Z. and D.H.; methodology, D.L. and Z.Z.; software, Z.Z. and P.L.; validation, Z.Z., formal analysis, C.Z. and Z.Z.; investigation, Y.C.; resources, Y.C. and D.H.; data curation, H.Z. writing—original draft preparation, H.Z. and Z.Z.; writing—review and editing, D.H.; visualization, Z.Z.; supervision, D.H.; project administration, Y.C. and D.H.; funding acquisition, D.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by State Grid Corporation of China Science and Technology Project (Grant Number: 5400-202218186A-2-0-XG).

Data Availability Statement: The regulatory documents (Chinese building standards) can be retrieved free of charge from the Ministry of Housing and Urban–Rural Development of China (https://www.mohurd.gov.cn). The IFC model rule base in the State Grid project has confidential information which cannot be shared publicly. Sample data with anonymized information (20 example clauses, 10 example rules, 3 IFC elements) will be available at https://github.com/knowledgepuppy/Data-Availability-Statement, 11 November 2025.

Acknowledgments: The valuable domain expertise contributed by the electrical engineers and BIM managers from State Grid Corporation during the validation and deployment test phases is gratefully acknowledged. During the preparation of this manuscript, the authors used ChatGPT-4 (OpenAI) for the purposes of language translation and polishing, as the authors are non-native English speakers. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: Zhuoqun Zhang, Chunhui Zhao, Danli Li and Peijie Li were employed by the company State Grid Corporation of China. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Ministry of Housing and Urban-Rural Development. *China Building Standards System*; MOHURD: Beijing, China, 2023.
2. *GB 50016-2014*; Code for Fire Protection Design of Buildings. 2018 Edition; China Planning Press: Beijing, China, 2018.
3. Eastman, C.; Lee, J.; Jeong, Y.; Lee, J. Automatic rule-based checking of building designs. *Autom. Constr.* **2009**, *18*, 1011–1033. [CrossRef]
4. Solihin, W.; Eastman, C. Classification of rules for automated BIM rule checking development. *Autom. Constr.* **2015**, *53*, 69–82. [CrossRef]
5. Zhang, J.; El-Gohary, N.M. Automated information transformation for automated regulatory compliance checking in construction. *J. Comput. Civ. Eng.* **2015**, *29*, B4015001. [CrossRef]
6. State Grid Corporation. *Small Infrastructure Project Management Guidelines*; China Electric Power Press: Beijing, China, 2022.
7. Jotne, A.S. EDM Model Server. Available online: https://www.jotneit.no (accessed on 27 October 2024).
8. Beach, T.H.; Rezgui, Y.; Li, H.; Kasim, T. A rule-based semantic approach for automated regulatory compliance in the construction sector. *Expert Syst. Appl.* **2015**, *42*, 5219–5231. [CrossRef]
9. Zhang, R.; El-Gohary, N. A clustering approach for analyzing the computability of building code requirements. In Proceedings of the ASCE International Conference on Computing in Civil Engineering, Atlanta, GA, USA, 17–19 June 2019.
10. Chen, K.; Chen, W.; Li, C.T.; Cheng, J.C.P. Automated compliance checking in construction using large language models. *Autom. Constr.* **2024**, *158*, 105215.
11. Nazarenko, A.; Lévy, F.; Wyner, A. Towards a methodology for formalizing legal texts in LegalRuleML. In *Proceedings of the 29th Internasional Conference Legal Knowledge and Information Systems (JURIX)*; IOS Press: Amsterdam, The Netherlands, 2016; pp. 149–152.
12. Arora, C.; Sabetzadeh, M.; Briand, L.C.; Zimmer, F. Extracting domain models from natural-language requirements: Approach and industrial evaluation. In *Proceedings of the 19th Internasional Conference Model Driven Engineering Languages and Systems (MODELS)*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 250–260.
13. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [CrossRef]
14. Du, C.; Nousias, S.; Borrmann, A. Towards a copilot in BIM authoring tool using a large language model-based agent for intelligent human-machine interaction. *arXiv* **2024**, arXiv:2406.16903.
15. Amershi, S.; Cakmak, M.; Knox, W.B.; Kulesza, T. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Mag.* **2014**, *35*, 105–120. [CrossRef]
16. Holzinger, A. Interactive Machine Learning for Health Informatics: When Do We Need the Human-in-the-Loop? *Brain Inform.* **2016**, *3*, 119–131. [CrossRef] [PubMed]
17. Mosqueira-Rey, E.; Hernández-Pereira, E.; Hernández-Castiñeiras, M.; Alonso-Betanzos, A. Human-in-the-loop Machine Learning: A Survey. *Neurocomputing* **2023**, *523*, 83–105.
18. Kim, I.; Choi, J.; Cho, G.; Koo, B. Automated compliance checking system for Korean building codes. In Proceedings of the 34th International Symposium on Automation and Robotics in Construction, Taipei, Taiwan, 28 June–1 July 2017; pp. 495–502.

19. Choi, J.O.; Kim, I. BIM-Based Approach for Automated Code Compliance Checking for Building Elements. *KSCE J. Civ. Eng.* **2017**, *21*, 1013–1025.
20. Preidel, C.; Borrmann, A. Automated Code Compliance Checking Based on BIM: A Review. *Adv. Eng. Inform.* **2016**, *30*, 383–399.
21. Przybyla, J.; Nguyen, N.; Eastman, C. Tracking Code Revisions and Rule Evolution for BIM-Based Compliance Checking. *J. Comput. Civ. Eng.* **2021**, *35*, 04021007.
22. Zhang, Y.; Song, Z.; Xu, W.; El Saddik, A. Human-in-the-Loop Artificial Intelligence for High-Stakes Decision Support. *IEEE Trans. Hum.-Mach. Syst.* **2022**, *52*, 776–789.
23. Solibri. Solibri Model Checker. Available online: https://www.solibri.com (accessed on 27 October 2024).
24. Nawari, N.O. The challenge of computerizing building codes in a BIM environment. In *Computing in Civil Engineering 2012*; ASCE: Reston, VA, USA, 2012; pp. 285–292.
25. Peng, Y.; Lin, J.R.; Zhang, J.P.; Hu, Z.Z. A hybrid framework for automated compliance checking of building designs. *J. Constr. Eng. Manag.* **2023**, *149*, 04023016.
26. Zhou, P.; El-Gohary, N. Ontology-based multilabel text classification of construction regulatory documents. *J. Comput. Civ. Eng.* **2016**, *30*, 04015058. [CrossRef]
27. Madireddy, S.; Wan, L.; Brilakis, I. Automated generation of building SMART Data Dictionary definitions using large language models. In Proceedings of the European Conference on Computing in Construction, Crete, Greece, 10–12 July 2024.
28. Pauwels, P.; Terkaj, W. EXPRESS to OWL for construction industry: Towards a recommendable and usable ifcOWL ontology. *Autom. Constr.* **2016**, *63*, 100–133. [CrossRef]
29. Ren, X.; Li, S.; Wang, F. Active Learning with Expert-in-the-Loop for Safety-Critical Recognition. *Pattern Recognit.* **2021**, *118*, 108034.
30. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On Calibration of Modern Neural Networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 1321–1330.
31. Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30, pp. 6402–6413.
32. Kendall, A.; Gal, Y. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30, pp. 5574–5584.
33. Kuleshov, V.; Fenner, N.; Ermon, S. Accurate Uncertainties for Deep Learning Using Calibrated Regression. In Proceedings of the ICML, Stockholm, Sweden, 10–15 July 2018; pp. 2796–2804.
34. Knublauch, H.; Kontokostas, D. *Shapes Constraint Language (SHACL), W3C Recommendation, Version 1.0*; W3C: Cambridge, MA, USA, 2017. Available online: https://www.w3.org/TR/shacl/ (accessed on 11 November 2025).
35. Corman, J.; Gayo, J.; Prud'hommeaux, E.; Solbrig, H. Validating SHACL Constraints over a SPARQL Endpoint. In Proceedings of the International Semantic Web Conference, Monterey, CA, USA, 8–12 October 2018; pp. 1–16.
36. Labra Gayo, J.E.; Prud'hommeaux, E.; Boneva, I.; Kontokostas, D. *Validating RDF Data; Synthesis Lectures on the Semantic Web: Theory and Technology*; Morgan & Claypool Publishers: San Rafael, CA, USA, 2017; Volume 7, pp. 1–328.
37. González, D.R.; Hogan, A.; Rivero, C. SHACL4P: Scalable Validation of RDF Graphs against SHACL Constraints. *J. Web Semant.* **2022**, *75*, 100764.
38. Xu, Y.; Deng, Y.; Li, Z.; Lu, Y. Semantic Approach to Compliance Checking of Chinese Building Fire Codes. *Autom. Constr.* **2021**, *130*, 103851.
39. Liu, X.; Zhang, S.; Lu, M. Automated Rule-Based Checking for Chinese Fire Codes Using BIM. *Eng. Appl. Artif. Intell.* **2020**, *87*, 103281.
40. Li, H.; Hou, L.; Wang, X. BIM-Based Compliance Checking for Chinese Energy Codes. *Sustain. Cities Soc.* **2019**, *44*, 740–750.
41. Malsane, S.; Matthews, J.; Lockley, S.; Love, P.E.D.; Greenwood, D. Development of an Object Model for Automated Compliance Checking. *Autom. Constr.* **2015**, *49*, 191–206. [CrossRef]
42. Platt, J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classif.* **1999**, *10*, 61–74.
43. Jiang, S.; Cheng, J.C.P.; Wang, Q. Ontology-Based Representation of Building Codes for Automated Compliance Checking. *Autom. Constr.* **2022**, *135*, 104110.
44. Farias, T.M.B.; Farias, R.P.; Amorim, S.R.L. IFCWebServer: A REST Platform for IFC-Based Semantics and Linked Building Data. *J. Inf. Technol. Constr.* **2015**, *20*, 410–432.
45. Zhong, B.; Ding, L.; Li, H.; Luo, H.; Zhou, W. Ontology-Based Semantic Modeling of Regulation Constraints for Automated Compliance Checking. *Adv. Eng. Inform.* **2019**, *40*, 117–133.
46. Pauwels, P.; Daniele, L.; Törmä, S. Linked Data in Architecture and Construction: A Review. *Autom. Constr.* **2017**, *73*, 174–194.
47. Zhao, X.; Zhou, Y.; Yang, Y. Digitalization of GB Codes: A Knowledge Graph Approach. *Buildings* **2023**, *13*, 560.
48. *GB 50053-2013*; Design Code for 20kV and Below Substation. Ministry of Housing and Urban-Rural Development of the People's Republic of China: Beijing, China, 2013.

49. *GB 50054-2011*; Code for Design of Low Voltage Electrical Installations. Ministry of Housing and Urban-Rural Development of the People's Republic of China: Beijing, China, 2011.
50. *DGJ 08-2048-2016*; Code for Electrical Design of Civil Buildings. Shanghai Municipal Commission of Housing and Urban-Rural Development: Shanghai, China, 2016.
51. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
52. Fleiss, J.L. Measuring Nominal Scale Agreement among Many Raters. *Psychol. Bull.* **1971**, *76*, 378–382. [CrossRef]
53. Brooke, J. SUS-A Quick and Dirty Usability Scale. In *Usability Evaluation in Industry*; Taylor & Francis: London, UK, 1996; pp. 189–194.
54. Lifschitz, V. *Answer Set Programming*; Springer: Berlin/Heidelberg, Germany, 2019.
55. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
56. Banarescu, L.; Bonial, C.; Cai, S.; Georgescu, M.; Griffitt, K.; Hermjakob, U.; Knight, K.; Koehn, P.; Palmer, M.; Schneider, N. Abstract Meaning Representation for sembanking. In Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, Sofia, Bulgaria, 8–9 August 2013; pp. 178–186.