Causal-Constrained GraphRAG: Enhancing Explainability and Reliability in Safety-Critical Parameter Extraction

Anonymous ACL submission

Abstract

Causal-Constrained GraphRAG is a neurosymbolic framework designed to improve the explainability and reliability of parameter extraction from long, structured, and causally rich regulatory corpora used in safety-critical engineering domains such as fire safety and evacuation modeling. The framework parses heterogeneous technical manuals into a topologically grounded knowledge graph, extracts explicit cause–effect–action chains through weak supervision combined with an active-learning annotation loop, and injects validated causal structures into a GraphRAG reasoning stage to constrain large language model inference. Instantiation on a curated Evacuation and Fire Safety question–answer dataset (approximately 24k pairs) and a Neo4j-based causal knowledge graph enables evaluation across multi-hop regulatory reasoning, numerical and sign consistency, and end-to-end parameter generation tasks. Experimental results demonstrate substantial improvements in factual retrieval accuracy and causal coherence, including notable F1 gains in occupant-load interpretation and high mean opinion scores from domain experts, alongside improved response stability under repeated technical queries. The framework supports safe and auditable automation of engineering parameter extraction while explicitly preserving human-in-the-loop validation for cases involving ambiguous, incomplete, or conflicting regulatory guidance.

1 Introduction

The deployment of Large Language Models (LLMs) in safety-critical engineering domains, such as fire safety and evacuation modeling, is hindered by a persistent reliability gap. Standard Retrieval-Augmented Generation (RAG) frameworks have demonstrated proficiency in factual retrieval, but they frequently struggle with the implicit reasoning required to navigate complex regulatory landscapes (Gwynne et al., 2015). In these domains, practitioners must synthesize information from fragmented technical manuals, such as NFPA 101 (Tubbs et al., 2024), BS 9999 (British Standards Institution, 2017), and the SFPE Handbook (Hurley et al., 2016), to derive precise and mutually consistent simulation parameters.

A critical bottleneck arises when simulation-ready parameters are missing or underspecified within the source text (Dimyadi et al., 2008). Traditional LLMs, when faced with such gaps, are prone to hallucinations or rely on shallow statistical correlations rather than the domain-specific causal logic required for safety-critical validation (Durmus et al., 2024). For instance, a model may successfully retrieve a baseline walking speed but fail to adjust it for the interdependent causal effects of smoke density and visibility unless explicitly guided by a structured knowledge framework (Kuligowski and Gwynne, 2010). To address these challenges, we propose a *Neurosymbolic framework* that integrates the linguistic flexibility of LLMs with the structural rigor of Causal-Constrained GraphRAG. Our approach moves beyond simple vector-based retrieval by parsing unstructured regulations into a topologically grounded knowledge graph (Edge et al., 2024). Furthermore, we introduce an NLP pipeline that leverages weak supervision and active learning to extract explicit causal chains, such as $SmokeDensity \uparrow\rightarrow Visibility \downarrow\rightarrow WalkingSpeed \downarrow$, directly from technical corpora.

By injecting these causal structures into the reasoning process of LLMs and a fine-tuned Gemma 3 (4B) model, we ensure that inferred parameters are both explainable and compliant with established engineering standards. Our evaluation, including multi-hop reasoning tasks and Mean Opinion Scores (MOS) from domain experts, demonstrates that this framework significantly improves the causal coherence and technical accuracy of pa-

rameter extraction in high-stakes environments.

Although we instantiate and evaluate our framework in fire safety and evacuation modeling, the underlying methods generalize to other regulatory domains involving long, structured, and causally rich corpora (e.g., transportation safety guidelines, accessibility standards, and healthcare protocols).

Contributions

- We formulate causally-constrained regulatory parameter inference as an NLP task and construct a regulatory QA resource over fire and evacuation standards to support training and evaluation.
- We propose a Causal-Constrained GraphRAG architecture that combines graph-structured retrieval over regulatory codes with a neurosymbolic causal knowledge graph extracted via weak supervision and active learning.
- We conduct a comprehensive evaluation of multiple LLMs (including a fine-tuned 4B-parameter model) on multi-hop regulatory QA, causal reasoning, and end-to-end parameter generation, with expert MOS and stability analyses in a safety-critical setting.

2 Related Work

2.1 RAG in Technical and Safety-Critical Domains

Traditional RAG systems primarily utilize vector-only retrieval, which excels at surface-level factual recall but often fails to capture the structural logic embedded in regulatory codes (Gao et al., 2024). In safety-critical fields like fire engineering, parameters are rarely standalone; they are governed by interdependent constraints found in manuals such as NFPA 101 and BS 9999. Previous research indicates that manual extraction from these sources is highly prone to error and inconsistency (Gwynne et al., 2015; Rafe et al., 2025). Emerging AI-driven approaches have attempted to automate this via NLP parsing and LLM-based retrieval (Li et al., 2024; Madireddy et al., 2025), but they typically treat parameters in isolation and do not jointly reason over interacting quantities such as occupant load, exit capacity, and smoke conditions.

Moreover, most existing RAG pipelines for technical domains implement a flat, chunk-based retrieval mechanism that ignores the underlying document structure and the logical relationships among provisions (Sarthi et al., 2024; Huang et al., 2025; Lu et al., 2025). As a result, they can retrieve relevant fragments while still producing parameter sets that violate implicit constraints in the standards (e.g., consistency between egress width, occupant density, and allowable evacuation times). In contrast, GraphRAG-style methods that operate over knowledge graphs promise more structure-aware retrieval (Edge et al., 2024), yet they have rarely been specialized to safety codes or tightly integrated with explicit causal constraints. Our work builds on this line by constructing a graph representation of regulatory sources and augmenting it with causal relations tailored to evacuation parameters.

2.2 Causal NLP and Knowledge Extraction

A major limitation of current LLMs is their reliance on statistical correlation rather than causal logic, which can result in counterintuitive or unsafe parameter sets in safety-critical settings (Hashmi et al., 2024). Recent advancements in Causal NLP have focused on extracting causal graphs from unstructured text and using them to guide model behavior (Wang et al., 2025; Kiciman et al., 2023). For example, Kiciman et al. (2023) demonstrate that causal-aware LLMs significantly outperform correlation-based systems in pairwise inference tasks when provided with explicit causal structures. Complementary work shows that embedding knowledge graphs into the reasoning process can improve generalization under limited or noisy data (Kiciman et al., 2023; Luo et al., 2024; Wu and Tsioutsiouliklis, 2024).

However, most existing causal-NLP frameworks are evaluated on short sentences, news-style text, or synthetic benchmarks, and they rarely address long, highly structured regulatory documents with dense cross-references. Furthermore, prior work typically focuses on causal inference tasks in isolation rather than on end-to-end generation of domain-specific artifacts such as simulation parameter sets.

3 Methodology

We propose a modular Neurosymbolic pipeline designed to transform unstructured regulatory corpora into causally-coherent simulation parameters. The framework consists of three primary stages: (1) structured multi-modal parsing, (2) targeted causal relationship extraction, and (3) Causal-Constrained

GraphRAG inference.

3.1 Multi-Modal Document Parsing and Structuring

Technical manuals such as NFPA 101 and the SFPE Handbook contain dense tabular data and hierarchical sectioning often lost in standard PDF-to-text conversion. To preserve semantic integrity, we evaluated multiple parsers using the DP-Bench benchmark (Upstage, 2024). LlamaParse (Liu, 2022) was selected as the primary engine, achieving a Normalized Indel Distance (NID) of 98.54 and a Tree Edit Distance-based Similarity (TEDS) of 92.38 for table recognition.

3.2 Targeted Causal Relationship Extraction

To address missing parameters, we implement a seven-stage extraction pipeline to identify "cause-effect-action" structure:

1. **Data Collection and Preprocessing:** Textual content from authoritative technical manuals, including the SFPE Handbook and NFPA 101—is cleaned and segmented into a sentence-level dataset $S = \{s_1, s_2, ..., s_N\}$.

2. **Candidate Identification:** We utilize dependency parsing via spaCy (Honnibal et al., 2020) to identify marked causality (e.g., "results in") and unmarked causality (e.g., implied relations).

3. **Weak Supervision:** Using the Snorkel framework (Ratner et al., 2017), we apply labeling functions Λ to assign probabilistic labels. The confidence score is computed as:

$$P(\text{causal}|s_l, \Lambda) = \frac{\sum_{j=1}^{m} w_j P(\lambda_j(s_l)|\text{causal})}{\sum_{j=1}^{m} w_j} \tag{1}$$

   where w_j is a weight assigned to trigger-specific labeling functions (e.g., $w_j = 1.5$ for simulation variables).

4. **Manual Labeling:** Top candidates are manually annotated for cause/effect spans and direction $d \in \{+, -\}$ using domain expertise[cite: 310, 312].

5. **Active Learning Loop:** We employ entropy-based uncertainty sampling via modAL (Danka and Horvath, 2018). Uncertainty $\mathcal{H}(s_i)$ is calculated using the multiline approach to fit column width:

$$\mathcal{H}(s_i) = -P(\text{causal}|s_i) \log P(\text{causal}|s_i) \\ -(1-P(\text{causal}|s_i)) \log(1-P(\text{causal}|s_i)) \tag{2}$$

   where the causal probability follows a Logistic Regression model[cite: 320]:

$$P(\text{causal}|s_i) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_{i1}+\cdots+\beta_n x_{in})}} \tag{3}$$

6. **Structure Inference:** Extracted relations are categorized by a reasoning model into chains ($v_1 \to v_2 \to v_3$), forks ($v_1 \leftarrow v_2 \to v_3$), or colliders ($v_1 \to v_2 \leftarrow v_3$).

7. **Graph Database Storage:** Finally, the structured causal relationships are encoded into a Neo4j (Neo4j, Inc., 2025) graph database, where each node represents a simulation variable and directional edges express validated causal influences.

3.3 Causal-Constrained GraphRAG and Reasoning Pipeline

The final stage implements a GraphRAG architecture to bridge the gap between static regulatory knowledge and dynamic reasoning for parameter extraction. This system utilizes a Neo4j knowledge graph to provide a topologically grounded reference for the language models.

Graph Construction The Neo4j Knowledge Graph is constructed by encoding structured document segments and extracted causal structures into a property graph $G = (V, E)$. Nodes V represent simulation variables (e.g., *Smoke Density*, *Walking Speed*), while directed edges E represent validated causal influences (chains, forks, or colliders) or regulatory hierarchies. For instance, the graph for NFPA 101 encompasses over 33,000 nodes and 76,000 relationships.

Causal-Aware "Thinking" Prompts To fill parameter gaps where direct manual guidance is absent, we employ a "Thinking-Mode" reasoning framework. The process utilizes a structured four-step prompt to force the model to reason through the causal chain before delivering a final numerical estimate.

As defined in the prompt structure, the model is explicitly instructed to align its reasoning with retrieved causal relationships by injecting validated paths (e.g., $Smoke\ Density \uparrow \to Visibility \downarrow \to$

$Speed \downarrow$) into the context, thereby constraining inference within physical laws and domain logic. Reasoning is grounded in structured baseline values extracted from authoritative standards such as SFPE and BS 9999, ensuring regulatory consistency. When direct numerical guidance is unavailable, the model analyzes underlying causal mechanisms to derive reasonable estimates using logical constraints and appropriate statistical assumptions, such as log-normal distributions for delay processes. Each inferred output is accompanied by a structured justification to verify adherence to established domain dependencies and regulatory constraints.

4 Experimental Setup

4.1 Dataset and Corpus

The experimental corpus is built upon authoritative fire safety and evacuation manuals, including NFPA 101, BS 9999, and the SFPE Handbook. From these sources, we curated the *Evacuation and Fire Safety Q&A Dataset*, comprising approximately 24,000 question-answer pairs. The dataset was expanded using a three-stage process: manual extraction by domain experts (D_0), LLM-assisted expansion (D_{LLM}), and entropy-based active learning refinement (D_{AL}) to optimize for edge-case coverage and technical density.

4.2 Model Selection and Fine-Tuning

We evaluated four primary model configurations, including commercial reasoning models, GPT-5.2, Claude Opus 4.5, and Gemini 3 Pro, as well as a fine-tuned domain-specific model based on Gemma 3 (4B) trained on the curated Q&A dataset described earlier. Fine-tuning of Gemma 3 was conducted using the Unsloth framework on a single NVIDIA A100-SXM4 GPU with mixed-precision (bfloat16) training. Optimization minimized cross-entropy loss over the task-adapted parameters θ^*, converging to a stable minimum loss of 1.20 after approximately 400 training steps.

4.3 Evaluation Tasks and Metrics

Models were benchmarked across four distinct safety-critical task categories:

1. *Tabular Data Interpretation:* Extracting specific factors (e.g., occupant load factors) from structured regulatory tables.
2. *Textual Information Retrieval:* Accurate clause identification and retrieval from unstructured text.
3. *Numerical Sign/Units Handling:* Correct detection and interpretation of signs (e.g., negative tolerances) and unit conversions.
4. *Calculation Application:* Multi-step application of regulatory formulas (e.g., exit capacity based on area and density).

Performance was quantified using standard IR/QA metrics: Precision, Recall, and F1-score. For qualitative reasoning and causal consistency, we utilized the MOS) scale (1–5), where domain experts assessed technical accuracy, standards compliance, and causal coherence.

4.4 Case Study: Spatial Layout Injection

To evaluate the framework's end-to-end performance in a real-world setting, we utilized a four-story educational building as a case study. This multi-functional facility, comprising classrooms, laboratories, administrative offices, and a public café, provided a complex architectural environment to test spatial grounding and functional context.

We transformed building plans into a topological graph $G_{spatial} = (V, E)$. In this representation, nodes V denote individual rooms or circulation zones, while edges E correspond to horizontal or vertical connections such as corridors, doors, and stairwells. Each node was enriched with geometric and semantic attributes, including net area and occupancy classifications. This spatial graph was integrated into the RAG pipeline, allowing the LLMs to query and generate simulation-ready parameters (e.g., occupant load) tailored to the specific functional constraints of each individual room.

5 Results and Analysis

5.1 Retrieval Performance and Task Accuracy

The retrieval accuracy was assessed across four technical categories as illustrated in Figure 1. Claude Opus 4.5 consistently achieved the highest performance, with F1 scores ranging from 0.917 to 0.937. Specifically, it reached a peak F1-score of 0.937 in *Textual Information Retrieval*, demonstrating a superior capacity for long-context semantic parsing and clause resolution in regulatory corpora.

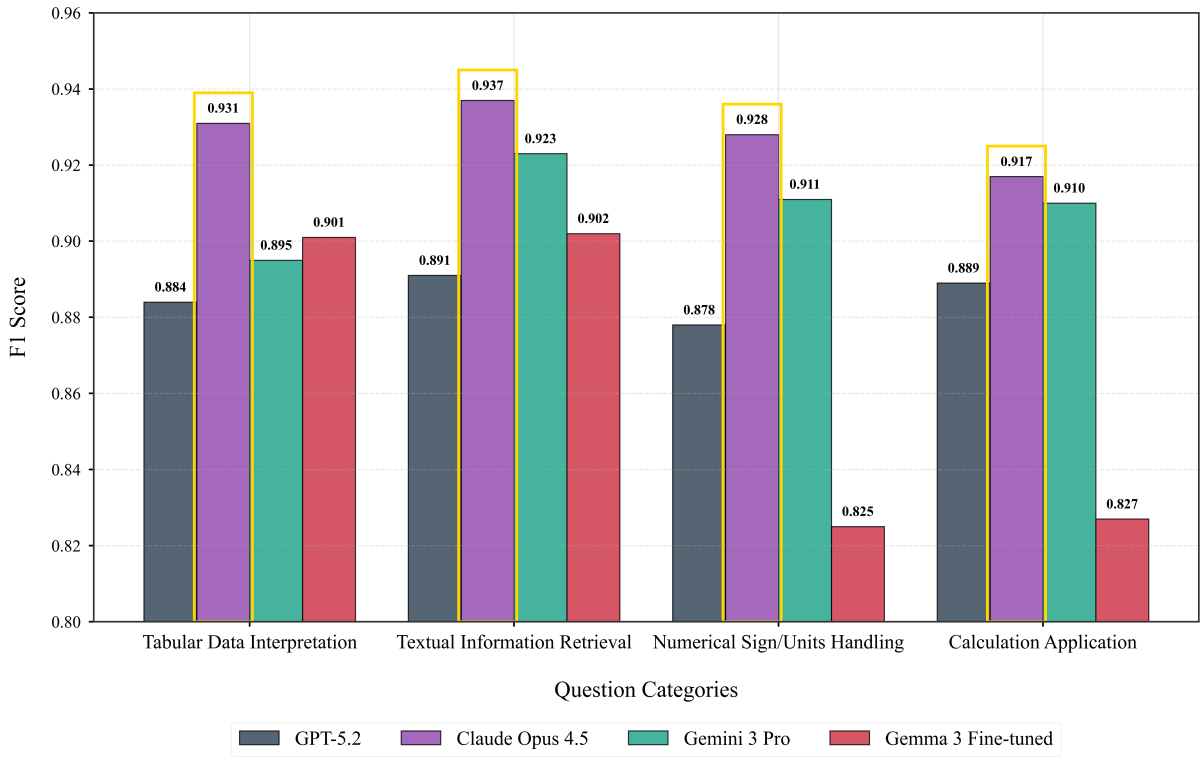

Figure 1: Performance of four LLMs across evacuation-related question categories.

The Specialization vs. Scale Trade-off The Gemma 3 Fine-tuned model (4B) exhibited significant adaptability, achieving an F1 score of 0.902 in *Textual Information Retrieval*. As illustrated in Figure 2, this reflects substantial domain-specific improvements across all regulatory categories compared to the base model.

The most pronounced advancement was observed in *Occupant Load* interpretation, which saw a $+140.2\%$ improvement in BERT-F1 score (increasing from 0.390 to 0.937). Similarly, *Fire Protection* and *Emergency Planning* tasks yielded gains of $+77.2\%$ and $+16.0\%$ respectively, validating the efficacy of the three-stage Q&A dataset generation process.

However, a performance ceiling was observed in computational tasks like *Numerical Sign/Units Handling* (0.825) and *Calculation Application* (0.827) as shown in Figure 1. From an NLP perspective, this suggests that while fine-tuning on domain-specific Q&A successfully encodes *lexical* and *terminological* knowledge, the *symbolic reasoning* required for mathematical manipulation remains highly dependent on the pre-training scale of larger models like Claude Opus 4.5. This validates the hybrid architecture of our framework: utilizing a fine-tuned small-language model (SLM) for efficient textual classification while delegating complex multi-step arithmetic to higher-capacity reasoning engines.

5.2 Reasoning Capabilities and Stability

A critical metric for safety-critical NLP is the ability to maintain consistency across repeated technical queries, a prerequisite for engineering reproducibility. In our stability tests (Table 1), Claude Opus 4.5 demonstrated high reliability, returning stable responses in approximately $75\% - 80\%$ of trials.

Impact of Document Heterogeneity A significant finding emerged regarding the linguistic structure of source documents. Gemma 3 Fine-tuned achieved perfect stability $(10/10)$ with the structured, rule-heavy content of *BS 9999*. However, its consistency plummeted to $5/10$ when faced with *NFPA 101*, a document characterized by greater heterogeneity, recursive cross-references, and implicit clause logic. This indicates that SLMs, even when fine-tuned, are more susceptible to "contextual noise" and "distractor clauses" within complex technical documents. In contrast, the "thinking mode" in Claude Opus 4.5 acts as a self-correction mechanism, allowing it to navigate heterogeneous layouts with consistent attention to the relevant

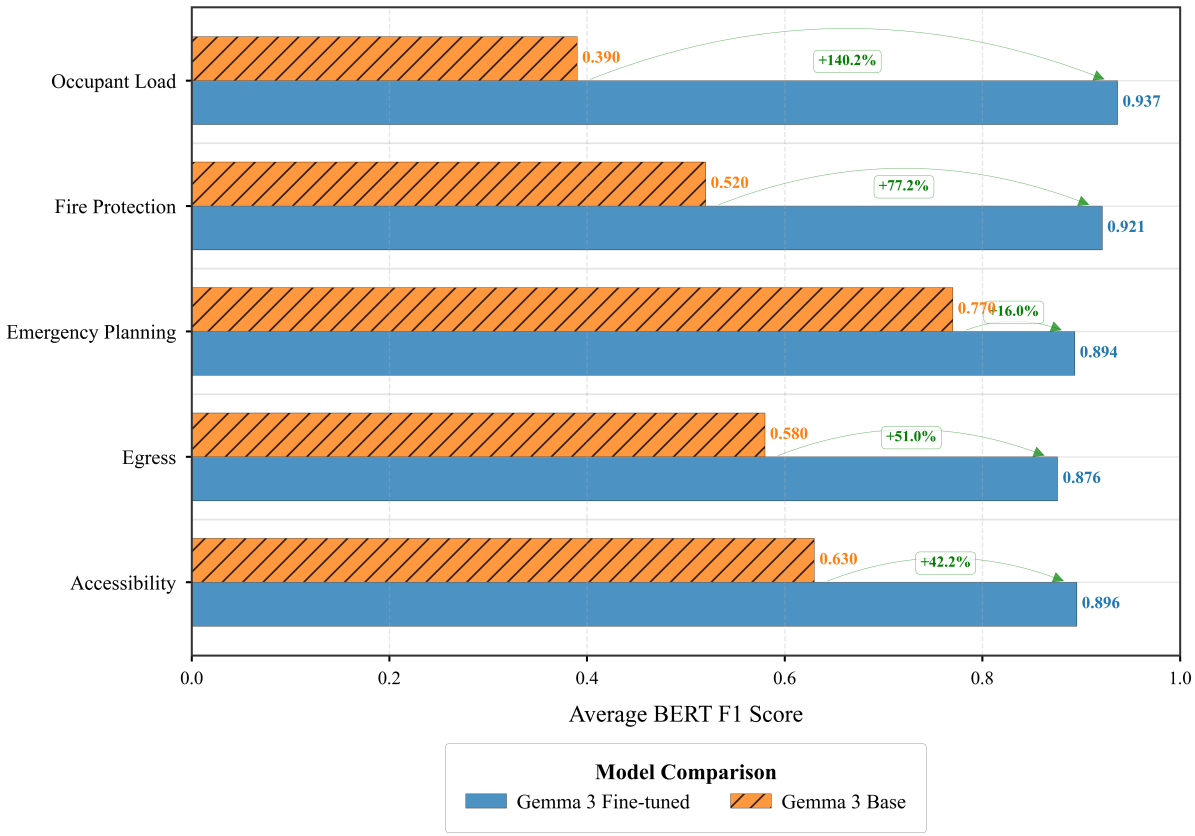

Figure 2: Comparative performance gains of Gemma 3 Fine-tuned over Gemma 3 Base across safety-critical domain categories.

LLM	SFPE Handbook	BS 9999	NFPA 101
Claude Opus 4.5	8/10	7/10	8/10
Gemma 3 (4B) Fine-tuned	7/10	10/10	5/10

Table 1: Stability Test Results: Number of stable, identical responses out of 10 queries per source.

topological links in the Neo4j graph.

5.3 Causal-Constrained Multi-hop Reasoning

Evaluation of multi-hop reasoning performance (Table 2) reveals the limitations of standard statistical retrieval. Claude Opus 4.5 successfully integrated complex, multi-variable causal chains (e.g., $smoke \rightarrow visibility \rightarrow speed$) in 9 out of 10 cases.

Neurosymbolic Injection vs. Stochastic Generation Most general-purpose models (GPT-5.2, Gemini 3 Pro) displayed a tendency to oversimplify or truncate causal logic when document segments were non-contiguous. This "shallow reasoning" is a byproduct of pure statistical prediction. By injecting the Causal Relationship Graph (Figure 3-10) directly into the prompt (Code 3-1), we force the model to adopt a *Neurosymbolic* reasoning path. Instead of predicting the next likely token based on correlation, the model must traverse the chain $v_1 \rightarrow v_2 \rightarrow v_3$ (e.g., $Smoke \uparrow \rightarrow Visibility \downarrow \rightarrow Speed \downarrow$) before arriving at a numerical estimate. This ensures that "missing" parameters are not just guessed, but inferred via physical and regulatory logic.

5.4 Explainability and Expert Evaluation (MOS)

The core contribution of this work is "Mechanistic Interpretability"—the ability to show *why* a parameter was chosen. This was validated through expert MOS in Table 3.

Experts rated Claude Opus 4.5 with a near-perfect *Causal Coherence* score of 4.9. The qualitative feedback indicated that while models like GPT-5.2 were efficient, their lack of explicit causal

Model	Accuracy	Linguistic & Reasoning Performance Notes
Claude Opus 4.5	9/10	Successfully integrated multi-variable causal chains with interpretable justifications and realistic statistical distributions.
GPT-5.2	8/10	Provided high numerical precision but occasionally exhibited "hallucinated simplifications" where complex causal pathways were truncated.
Gemini 3 Pro	7/10	Strong mathematical derivations; however, the model often prioritized numeric modeling over behavioral logic.
DeepSeek R1	7/10	Precise outputs for specific variables, but exhibited "weak contextual transition" when reasoning across three or more interdependent steps.
Sonar Reasoning Pro	6/10	Heavily grounded in citations, but failed to synthesize multi-hop logic when parameters were spread across non-contiguous document sections.

Table 2: Multi-hop reasoning accuracy and performance notes for each LLM configuration.

Model	Tech. Accuracy	Std. Compliance	Causal Coherence	Avg. MOS
Claude Opus 4.5	4.8	4.5	4.9	4.7
GPT-5.2	4.5	4.5	4.4	4.5
Gemini 3 Pro	4.6	4.2	3.7	4.2
Sonar Reasoning Pro	4.2	4.4	3.8	4.1
DeepSeek R1	3.5	4.4	4.0	4.0

Table 3: Mean Opinion Score (MOS) for LLM outputs by domain experts.

constraints led to parameter sets that were occasionally technically "correct" but logically disjointed. Our framework's output, by contrast, was deemed "trustworthy for engineering use" because the justification explicitly mirrors the causal logic found in authoritative manuals like the *SFPE Handbook*.

5.5 Generalization to Safety-Critical Domains

The primary implication of these results is that parameter extraction in safety-critical domains cannot rely solely on larger datasets. The $+140.2\%$ improvement in the fine-tuned Gemma model proves that domain-specific alignment is essential, but the stability and multi-hop reasoning results prove that an external "causal guardrail" (GraphRAG) is necessary to ensure those models operate within the bounds of physical and regulatory laws. This framework provides a reproducible blueprint for other high-stakes domains, such as medical protocol extraction or legal compliance, where "shallow retrieval" is insufficient.

6 Conclusion

This study introduced a *Neurosymbolic framework* to address the fundamental reliability gap in standard RAG pipelines for safety-critical engineering domains. By integrating a Causal-Constrained GraphRAG architecture with a 7-stage causal extraction pipeline, we demonstrated a robust methodology for transforming unstructured regulatory documents into verifiable simulation parameters.

Our findings establish that injecting explicit causal structures, such as chains, forks, and colliders, significantly bounds the "stochastic generation" of LLMs within established physical and regulatory logic. The empirical evaluation showed that high-capacity reasoning models like Claude Opus 4.5 achieved superior accuracy in multi-hop reasoning tasks, while our domain-specialized Gemma 3 Fine-tuned model demonstrated a improvement in interpreting specialized occupant load provisions. The successful generation of a simulation-ready XML file for a complex educational building case study validates the framework's readiness for real-world engineering workflows.

The primary implication of this work is a major shift from subjective, manual parameter extraction toward an objective, transparent, and reproducible process. This approach not only enhances the safety of performance-based design but also

provides a generalizable blueprint for other high-stakes domains (e.g., medical and legal compliance) where hallucinatory outputs are unacceptable.

Limitations

Despite its performance, the framework's output quality remains inherently dependent on the clarity and comprehensiveness of the source regulatory documents. While the causal-constrained reasoning mitigates hallucinations, the system is not infallible and still requires human-in-the-loop expert oversight, particularly in complex edge cases where regulatory guidance is silent or contradictory. Furthermore, the causal knowledge graph, while extensive, relies on an initial "seed" of expert-labeled relationships, and expanding this pool to a broader range of practicing fire engineers would be necessary for wider industry validation.

Ethical Considerations

In safety-critical domains such as fire safety and evacuation planning, the automation of engineering judgment carries significant ethical weight. Our framework is designed as a decision-support tool to enhance the reliability of simulation inputs, not as a replacement for certified professional engineers. The risk of over-reliance on AI-generated parameters must be mitigated through clear transparency of sources, provided by our GraphRAG's justification mechanism, ensuring that all derived values are traceable to authoritative engineering standards. Additionally, the use of open-weight models like Gemma 3 ensures that these tools can be deployed in secure, local environments, protecting sensitive building data from third-party exposure.

References

British Standards Institution. 2017. Bs 9999:2017 fire safety in the design, management and use of buildings. code of practice. British Standard BS 9999:2017, BSI, London, UK.

Tivadar Danka and Peter Horvath. 2018. modal: A modular active learning framework for python. *arXiv preprint arXiv:1805.00979*.

J. Dimyadi, M. Spearpoint, and R. Amor. 2008. Sharing Building Information using the IFC Data Model for FDS Fire Simulation. *Fire Safety Science*, 9:1329–1340.

Dilan Durmus, Alberto Giretti, Ori Ashkenazi, Alessandro Carbonari, and Shabtai Isaac. 2024. The Role of Large Language Models for Decision Support in Fire Safety Planning. Lille, France.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint*. ArXiv:2312.10997 [cs].

Steven MV Gwynne, Erica Kuligowski, Michael Spearpoint, and Enrico Ronchi. 2015. Bounding defaults in egress models. *Fire and Materials*, 39(4):335–352.

Syed Muhammad Ali Hashmi, Tamas Fekete, and Hendro Wicaksono. 2024. Causal AI in the Automotive Industry: Impact Analysis Through Carbon Emission Case Study. In *2024 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pages 1267–1271, Bangkok, Thailand. IEEE.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, and 1 others. 2020. spacy: Industrial-strength natural language processing in python.

Haoyu Huang, Yongfeng Huang, Junjie Yang, Zhenyu Pan, Yongqiang Chen, Kaili Ma, Hongzhi Chen, and James Cheng. 2025. Retrieval-augmented generation with hierarchical knowledge. *arXiv preprint arXiv:2503.10150*.

Morgan J. Hurley, Daniel Gottuk, John R. Hall, Kazunori Harada, Erica Kuligowski, Milosh Puchovsky, José Torero, John M. Watts, and Christopher Wieczorek, editors. 2016. *SFPE Handbook of Fire Protection Engineering*. Springer, New York, NY.

Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*.

Erica D. Kuligowski and Steve M. V. Gwynne. 2010. The Need for Behavioral Theory in Evacuation Modeling. In *Pedestrian and Evacuation Dynamics 2008*, pages 721–732, Berlin, Heidelberg. Springer.

Sihao Li, Jiali Wang, and Zhao Xu. 2024. Automated compliance checking for BIM models based on Chinese-NLP and knowledge graph: an integrative conceptual framework. *Engineering, Construction and Architectural Management*, 32(6):3832–3856. Publisher: Emerald Publishing Limited.

Jerry Liu. 2022. LlamaIndex.

Wensheng Lu, Keyu Chen, Ruizhi Qiao, and Xing Sun. 2025. Hichunk: Evaluating and enhancing retrieval-augmented generation with hierarchical chunking. *arXiv preprint arXiv:2509.11552*.

Linhao Luo, Zicheng Zhao, Gholamreza Haffari, Yuan-Fang Li, Chen Gong, and Shirui Pan. 2024. Graph-constrained reasoning: Faithful reasoning on knowledge graphs with large language models. *arXiv preprint arXiv:2410.13080*.

Soumya Madireddy, Lu Gao, Zia Ud Din, Kinam Kim, Ahmed Senouci, Zhe Han, and Yunpeng Zhang. 2025. Large Language Model-Driven Code Compliance Checking in Building Information Modeling. *Electronics*, 14(11):2146. Number: 11 Publisher: Multidisciplinary Digital Publishing Institute.

Neo4j, Inc. 2025. Neo4j graph database.

Amir Rafe, Peter J. Lawrence, Ruggiero Lovreglio, Michael Spearpoint, and Patrick A. Singleton. 2025. Enhancing Occupant Evacuation Simulation Using LLMs and Retrieval-Augmented Generation. pages 389–400. Publisher: American Society of Civil Engineers.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: rapid training data creation with weak supervision. *Proc. VLDB Endow.*, 11(3):269–282.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. 2024. Raptor: Recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations*.

Jeffrey Tubbs, Gregory Harrington, and Yee Joe, editors. 2024. *NFPA 101: life safety code*. National Fire Protection Association.

Upstage. 2024. Dp-bench: Document parsing benchmark. https://huggingface.co/datasets/upstage/dp-bench. Accessed: 2025-01-05.

Nengbo Wang, Xiaotian Han, Jagdip Singh, Jing Ma, and Vipin Chaudhary. 2025. CausalRAG: Integrating Causal Graphs into Retrieval-Augmented Generation.

Xue Wu and Kostas Tsioutsiouliklis. 2024. Thinking with knowledge graphs: Enhancing llm reasoning through structured data. *arXiv preprint arXiv:2412.10654*.

Appendices

A Prompt Engineering for Causal Inference

To reduce hallucinations and ensure physical consistency, we use a structured prompt that forces the LLM to reason over retrieved causal graphs before generating parameters.

```
Prompt Structure for Reasoning LLMs.

## **LLM Query: Structured Reasoning-Based
    Extraction for Walking Speed Distributions**

**[QUESTION]**

### **Step 1: Consider These Causal
    Relationships**
Before proceeding, ensure your reasoning aligns
    with the following **causal relationships**:

[CAUSAL RELATIONSHIPS LIST]

### **Step 2: Utilize Structured Knowledge
    Extracted from Manuals**
The following **structured data** has been
    extracted from fire safety manuals (NFPA,
    SFPE, BS-9999, CIBS Guide E, and other
    regulations):

[BASELINE VALUES]

### **Step 3: Reason & Generate a Structured
    Answer**
Using the provided **causal relationships** and
    **structured extracted data**, determine the
     most appropriate response.

- If numerical data is directly available,
    provide **exact probability distributions**
    (e.g., normal, log-normal, Weibull,
    triangular, uniform, discrete) along with **
    mean, standard deviation, and range**.
- If no direct numerical data is available,
    analyze the causal mechanisms and **derive
    reasonable estimates based on logical
    constraints**.
- If conflicting values exist, provide the **
    range of valid values** along with
    justifications.
- If necessary, supplement reasoning with **peer-
    reviewed studies**, but prioritize
    structured data and causal reasoning first.
- If conflicting values exist between **manuals
    & peer-reviewed studies**, explain the
    differences and justify the most appropriate
     choice.
- Do **not fabricate studies**--only reference
    those explicitly available.
- Clearly indicate **assumptions made** if gaps
    in the extracted knowledge exist.

### **Step 4: Format Output as Follows**
1. **Final Answer** (Provide the most
    appropriate statistical distribution &
    values for walking speeds in different
    environments)
2. **Justification & Reasoning** (Explain using
    causal mechanisms & extracted knowledge)
3. **Comparison With Literature (If Available)**
     (Compare with known fire safety studies)
4. **Causal Relationship Validation** (Ensure
    output aligns with known causal dependencies
    )
```

Listing 1: Structured Reasoning-Based Extraction Prompt

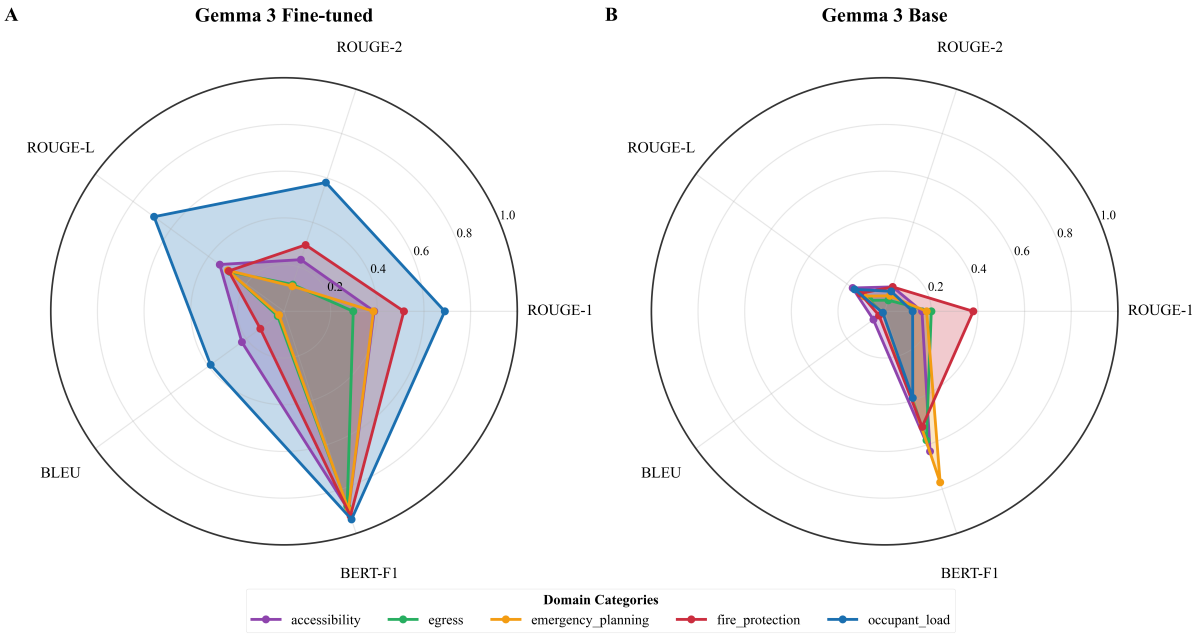

Figure 3: Radar charts comparing BERT-F1, ROUGE, and BLEU scores. (A) Gemma 3 Fine-Tuned shows balanced high performance across all metrics. (B) Gemma 3 Base exhibits significant degradation in specialized tasks like Occupant Load.

B Fine-Tuning Performance Analysis

Figure 3 presents a radar-based comparison of the fine-tuned Gemma 3 (4B) model (Panel A) against its base counterpart (Panel B) across five safety-critical domain categories: accessibility, egress, emergency planning, fire protection, and occupant load. Each radar axis corresponds to a complementary evaluation metric (ROUGE-1, ROUGE-2, ROUGE-L, BLEU, and BERT-F1), capturing both surface-level lexical overlap and deeper semantic alignment.

The fine-tuned model demonstrates substantial and consistent gains across all domains and metrics, with the most pronounced improvements observed in BERT-F1 and ROUGE-L, indicating enhanced semantic fidelity and clause-level understanding of regulatory and safety text. In particular, the occupant-load and fire-protection domains exhibit the largest performance envelopes, reflecting the effectiveness of domain-specific fine-tuning on numerically grounded and rule-intensive content. BLEU scores, while lower in absolute magnitude relative to semantic metrics, also increase uniformly after fine-tuning, suggesting improved local sequence consistency without overfitting to surface n-gram patterns.

In contrast, the base Gemma 3 model (Panel B) exhibits compressed radar profiles with limited separation across domains, especially for ROUGE-2, ROUGE-L, and BLEU. While moderate BERT-F1 scores indicate some baseline semantic capability, the overall narrow metric spread highlights the model's limited ability to robustly generalize across heterogeneous safety standards without targeted adaptation.

C Network Visualizations

Figure 4 presents the Neo4j-based regulatory knowledge graph derived from NFPA standards, illustrating a representative subset comprising 500 nodes out of a total of 33,686 and 974 relationships out of 76,377. Figure 5 depicts the abstract causal dependency network used to constrain and guide reasoning about evacuation performance. Nodes correspond to interpretable simulation and behavioral variables (e.g., *Pre-Evacuation Time*, *Walking Speed*, *Visibility*, *Occupant Compliance*), while directed edges encode hypothesized or empirically supported causal influences. Central outcome variables such as *Evacuation Time* emerge as hubs, integrating effects from perceptual factors (smoke density, lighting), social dynamics (group cohesion, staff credibility), environmental cues, and individual characteristics (physical limitations, prior experience).

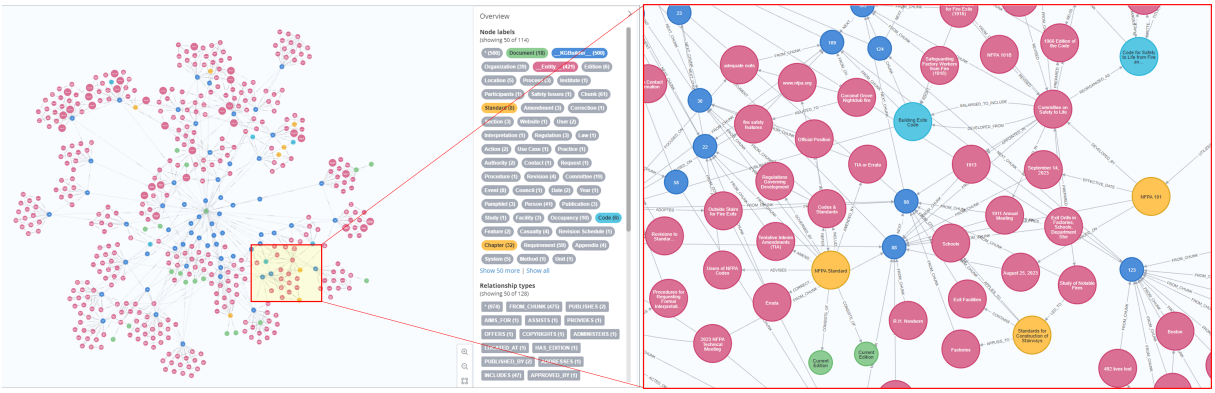

Figure 4: Sample Visualization of NFPA 101 Knowledge Graph in Neo4j.

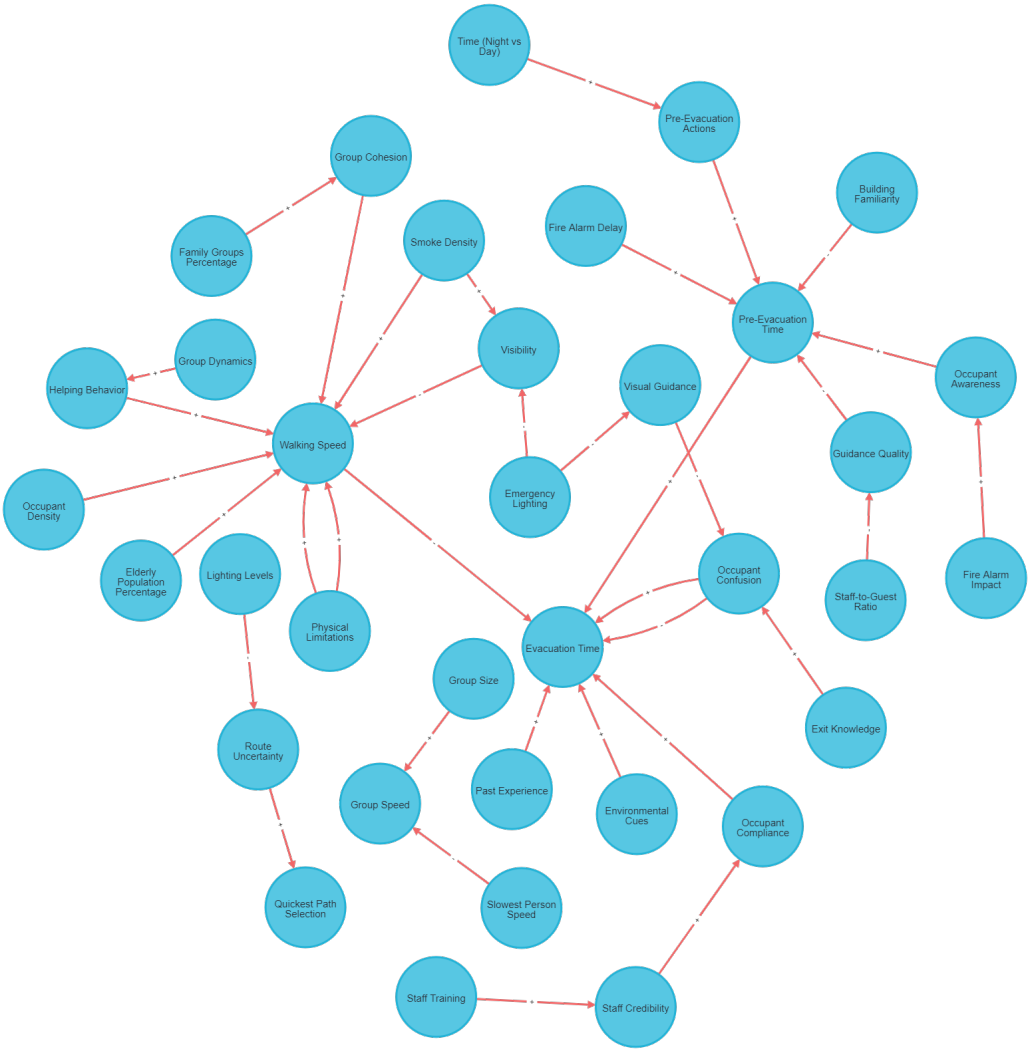

Figure 5: Causal Relationship Network Visualization in Neo4j.