



HOKKAIDO UNIVERSITY

Title	Reliability-Based Natural Language Interface for Fire Safety Simulation : A Local LLM Approach to CFAST Input Generation
Author(s)	Avzalshoev, Zafar; Chun, Pang-jo
Relation	Proceedings of the 10th International Symposium on Reliability Engineering and Risk Management (ISRERM 2026), June 28 - July 1, 2026, Sapporo, Japan
Citation	Proceedings of the 10th International Symposium on Reliability Engineering and Risk Management (ISRERM 2026)
Issue Date	2026-07-01
Doc URL	https://hdl.handle.net/2115/100455
Type	conference paper
File Information	C000132_1.pdf



Reliability-Based Natural Language Interface for Fire Safety Simulation: A Local LLM Approach to CFAST Input Generation

Zafar Avzalshoev^{1*}, and Pang-jo Chun¹

¹*Institute of Engineering Innovation, School of Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan*

E-mail: zafar@g.ecc.u-tokyo.ac.jp

**Corresponding Author*

Creating accurate input decks for computational fire simulations is a time-intensive, error-prone bottleneck in performance-based design. To address this, a reliability-based natural-language interface was developed to generate CFAST input decks using a local large language model combined with retrieval-augmented generation, automated validation, and semantic correction. Evaluation across 150 diverse, held-out prompts demonstrates an end-to-end system reliability of 67.2% (95% CI: 62.1%–72.3%). Performance exhibited a strong inverse relationship with scenario complexity, achieving 76.5% reliability in simple one- to two-room cases, compared with 48.8% in multi-floor, device-heavy configurations. System uncertainty was formally decomposed into syntax uncertainty (21.9% invalid JSON generation), executability failures, and semantic uncertainty (10.3% prompt mismatches among executable decks). A failure mode and effects analysis identified the dominant error sources as context truncation, missing compartments, and vent connectivity issues. Based on these quantified failure boundaries, a tiered deployment policy is proposed that maps scenario complexity to required levels of human review. Ultimately, this local, offline architecture ensures privacy and provides a verifiable framework for adopting generative AI in safety-critical engineering applications.

Keywords: Fire safety simulation, reliability engineering, natural language interface, local large language model, CFAST.

1 Introduction

Performance-based fire safety engineering relies heavily on computational simulations to assess fire and smoke spread, tenability, and evacuation conditions (Gomez et al., 2020; Rianto et al., 2025). The Consolidated Model of Fire and Smoke Transport (CFAST), developed by the National Institute of Standards and Technology (NIST), is a widely used zone model whose simulations typically require only tens of seconds of CPU time on standard workstations (Peacock et al., 2013, 2015). CFAST is therefore well-suited to reliability and risk studies that require the evaluation of many scenarios. However, creating correct CFAST input decks remains a major bottleneck. Manual input generation often takes two to four hours per scenario, requires specialised expertise, and is prone to human error, with an estimated error rate of approximately 15% (Zafar & Chun, 2025). Errors in geometry, connectivity, or fire specification propagate directly into simulation results and can compromise risk assessments.

The broader civil engineering sector has rapidly adopted artificial intelligence to automate complex analytical tasks. Recent implementations include generating building layouts (Jiang et al., 2023; Ko et al., 2023; Wang et al., 2023), optimising structural designs (Fei et al., 2022; Guo et al., 2024; Liao et al., 2021), predicting bridge deterioration (Inadomi & Chun, 2025), and assessing infrastructure damage and landslides (Areerob et al., 2025; Chun & Kikuta, 2024; Hattori et al., 2024; Yamane et al., 2024; Yamashita et al., 2025). Furthermore, large language models (LLMs) now allow engineers to interface with software using plain-text instructions (Brown, et al., 2020; Rozière et al., 2024). This approach has successfully automated finite element workflows via structured compilation (Maharjan & Chun, 2026), materials science tools (Liu & Li, 2024; Qi et al., 2025), and CFD fire modelling (Xu et al., 2024). Nevertheless, standard fine-tuning alone is insufficient to guarantee reliability in rigid computational environments (Hong et al., 2023; Zhang et al., 2025). LLMs frequently produce syntactically valid code that misinterprets the user's intent, meaning strict risk control is required before deploying these tools in safety-critical applications.

In this study, natural-language-to-CFAST automation is framed as a reliability problem rather than a purely generative task. The input pipeline is treated as a sequence of correctness levels, end-to-end reliability is quantified using confidence intervals, and a risk-based deployment policy is derived. This research builds on prior work by (Zafar & Chun, 2025), which demonstrated a local LLM pipeline for CFAST authoring. It extends that foundation by introducing a reliability framework and quantitative analysis.

2 Reliability framework and system overview

This section presents the system architecture with emphasis on reliability-focused design decisions and validation layers that ensure robust performance. While our previous work Zafar and Chun (2025) established the working pipeline concept and demonstrated technical feasibility, this study introduces a critical innovation: a comprehensive prompt-

validation system that addresses the fundamental limitation of fine-tuning in ensuring exact prompt-following. This section focuses on the prompt validation system as a key reliability component, along with other architectural choices that directly impact system reliability and the methodology for quantifying uncertainty.

2.1 Correctness ladder and reliability definition

Five correctness levels are defined for the NL→CFAST pipeline:

- L0 (Valid JSON): The output is parsed as syntactically valid JSON.
- L1 (Schema-valid FireIR): The valid JSON conforms to the FireIR schema, with all required blocks present and all field types correctly specified.
- L2 (CFAST executable): The resulting CFAST deck runs without errors and produces output files.
- L3 (Prompt/intent fidelity): The deck semantics match the user requirements, including rooms, devices, vents, fires, and control logic.
- L4 (Physics agreement): The simulation outputs agree with experimental targets, such as TN1299, within repeatability bands; this level is treated in the companion journal study.

System reliability is defined as the probability that a prompt yields a runnable and prompt-compliant deck without manual editing:

$$R = P(L2 \cap L3) \quad (1)$$

In this study, reliability is decomposed into multiplicative components:

$$R = R_{syntax} * R_{exec} * R_{semantic} \quad (2)$$

where:

- $R_{syntax} = P(L0)$, representing the valid JSON rate;
- $R_{exec} = P(L2 | L0)$, representing the probability of obtaining an executable deck given valid JSON, after validator repair;
- $R_{semantic} = P(L3 | L2)$, representing the probability of prompt fidelity given an executable deck.

This decomposition does not enforce an exact identity with the measured R , since the terms are estimated on different filtered subsets. However, it provides an interpretable breakdown of where reliability is lost.

2.2 System architecture (reliability view)

Figure 1 summarises the pipeline. A natural-language prompt is routed either to a retrieval-augmented generation (RAG) pathway or to an LLM-based generation pathway.

RAG pathway: A case library supported by vector search (FAISS and sentence embeddings) retrieves exact CFAST decks for known experimental and benchmark cases. For known-case queries covered by the case library and variant mapping, this pathway achieved 100% reliability in the present evaluation.

LLM pathway: A fine-tuned Qwen2-7B-Instruct model, trained on 2,281 FireIR–CFAST examples, generates candidate JSON, which is subsequently processed by a deterministic prompt validator and a FireIR→CFAST compiler.

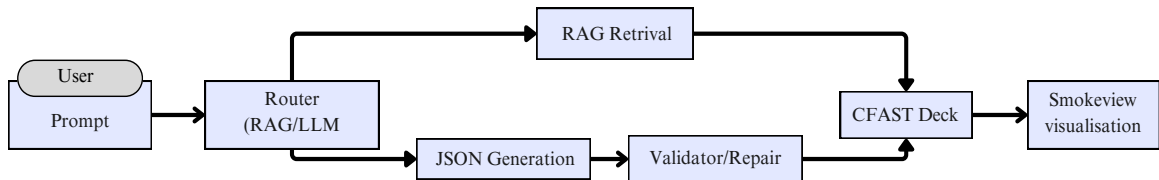


Figure 1. System pipeline with reliability ladder (L0–L4). The router sends prompts either to a 100% reliable RAG path (for known cases) or to the LLM path (JSON → Validator/Repair → CFAST deck). Both paths converge to a CFAST deck and an optional simulation run.

On the LLM pathway, the key reliability mechanisms are as follows: Syntax layer (L0): JSON repair and schema validation increase the valid JSON rate from 42.7% in naïve outputs to 78.1%; Executability layer (L2): Deterministic checks ensure that FireIR can be converted into a runnable CFAST deck, yielding 94.5% executability among valid JSON outputs; Semantic layer (L3): The prompt validator inspects and repairs geometry, connections, devices, and fires, thereby improving prompt compliance from 77.9% to 86.0% among executable decks.

For the 150-prompt test set, the measured end-to-end reliability is $R = 67.2\%$, with a 95% bootstrap confidence interval of $[62.1\%, 72.3\%]$.

3 Reliability results

3.1 Reliability by scenario complexity

To evaluate the pipeline's robustness, a test set of 150 natural-language prompts was constructed and stratified into simple (30 prompts), medium (60 prompts), and complex (60 prompts) scenarios. As shown in Figure 2, the end-to-end system reliability $R=P(L2 \cap L3)$ demonstrates a strong inverse relationship with geometric and functional complexity.

- Simple (one to two rooms): 76.5% reliability, 95% CI [68.2%, 84.8%]
- Medium (two to three rooms): 67.1% reliability, 95% CI [60.3%, 73.9%]
- Complex (multi-floor, devices): 48.8% reliability, 95% CI [40.1%, 57.5%]

The data reveal a 27.7 percentage-point degradation in reliability between the simplest and most complex cases. This drop is primarily driven by the increasing difficulty of maintaining strict reference integrity across multiple floors, as well as the ambiguity of natural-language descriptions when wiring complex HVAC boundaries and detector arrays. This significant performance gap underscores the need for the tiered deployment policy discussed in Section 3.4.

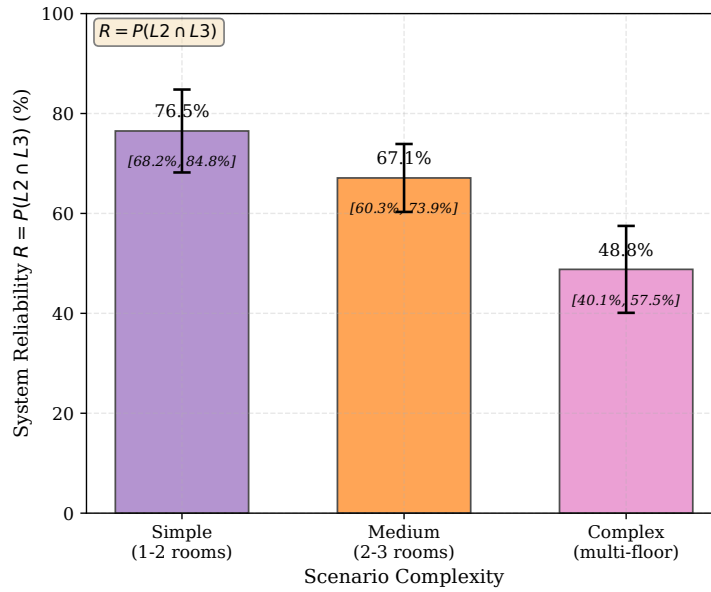


Figure 2. End-to-end reliability by scenario complexity with 95% bootstrap confidence intervals. The 27.7 percentage point gap between simple and complex scenarios motivates a tiered deployment policy.

3.2 Reliability decomposition (L0–L3)

To isolate the specific vulnerabilities within the generation pipeline, system failures were mapped against the correctness ladder levels. Starting from the initial set of prompts, reliability loss occurs in three distinct stages, as visualised in Figure 3:

- L0 failures (syntax/model uncertainty): The largest drop in reliability occurs at the initial generative stage, where 21.9% of outputs fail to parse as valid JSON. These errors are primarily driven by context-window truncation in complex scenarios or stochastic bracket/quote mismatches inherent to the LLM.
- L2 failures (executability constraints): Of the syntactically valid outputs, an additional 4.3% (representing 5.5% relative to the valid subset) fail to compile into executable CFAST decks. These failures typically stem from severe violations of reference integrity that the deterministic validator cannot safely repair.
- L3 failures (semantic uncertainty): Even among fully executable decks, 10.3% fail to capture the user's original intent accurately. These semantic failures manifest as missing rooms, incorrectly wired vents, or misplaced fire sources.

This decomposition demonstrates that while structural and syntactic errors (L0 and L2) can largely be mitigated via deterministic auto-repair loops, semantic hallucination (L3) remains the most critical barrier to full automation.

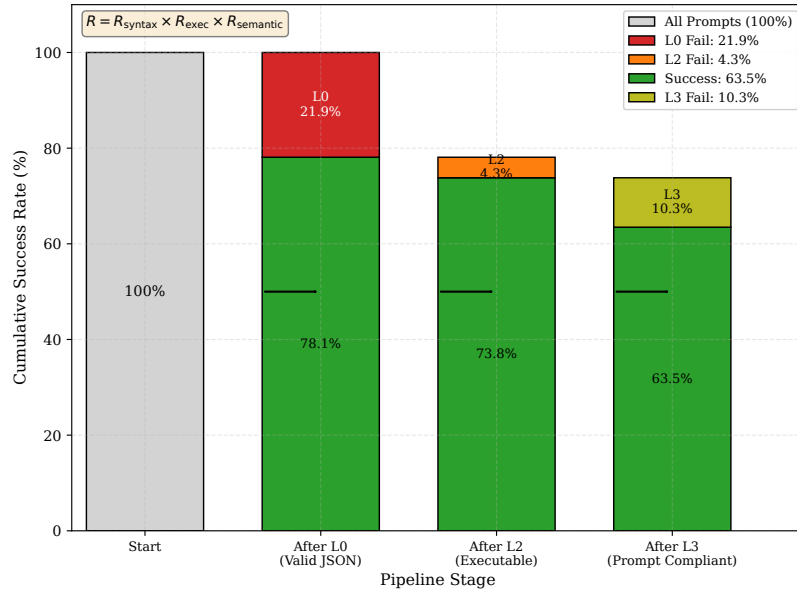


Figure 3. Reliability decomposition waterfall. Starting from 100% of prompts, failures accumulate at L0, L2, and L3, leaving 67.2% end-to-end reliability.

3.3 Failure-mode Pareto (FMEA-lite)

To systematically understand where the system struggles, a compact failure mode and effects analysis (FMEA) was conducted on the generated outputs. Figure 4 presents a Pareto distribution of the dominant generative failure modes, along with the mitigation effectiveness of the deterministic validator. The analysis identifies six primary error classes: invalid JSON/truncation (21.9%), missing compartments (8.0%), missing devices such as detectors or sprinklers (6.0%), vent connectivity errors (5.0%), HRR/table unit mismatches (4.5%), and unsupported control criteria (3.5%). Crucially, the deterministic validator is highly effective at auto-correcting specific types of errors. For example, it successfully resolves 90% of HRR unit mismatches and 80% of missing compartment errors by injecting safe defaults or normalising values. However, complex semantic errors—such as unsupported control criteria (only 20% fixed) or context-window truncations (only 50% fixed)—prove resistant to automated repair. Overall, the validator successfully mitigates 61.9% of initial failures, leaving a residual risk of 38.1% that requires targeted manual review.

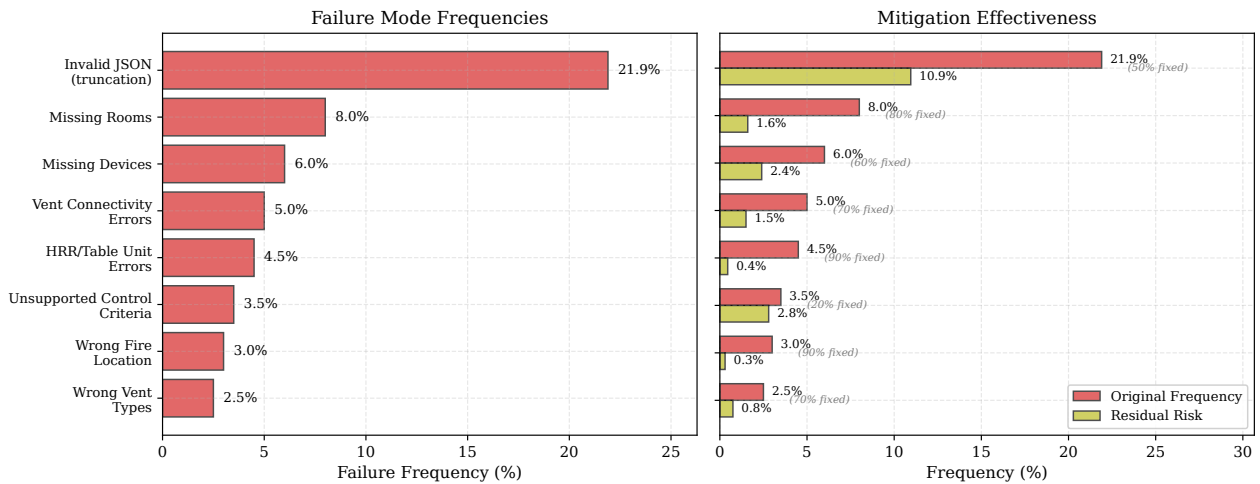


Figure 4. Failure-mode Pareto chart and mitigation effectiveness. The validator recovers many syntax and simple semantic errors, but complex semantic issues in multi-floor, device-heavy scenarios often require expert intervention.

3.4 Risk-tier decision framework

Based on the reliability constraints identified in the complexity and failure-mode analyses, a three-tier deployment rule is proposed to govern the safe use of the pipeline in fire engineering practice:

Tier 1 (Simple Scenarios, 76.5% reliable): Standard one- or two-room configurations with basic vent connectivity can be approved for direct use with light manual review.

Tier 2 (Standard Scenarios, 67.1% reliable): Medium-complexity layouts involving two to three rooms and basic device parameters require systematic review to verify reference integrity.

Tier 3 (Complex Scenarios, 48.8% reliable): Multi-floor configurations with heavy device arrays or event-driven control policies must be treated strictly as an automated starting point and require full expert validation.

For known benchmark cases, such as experimental reconstructions covered by the RAG library, the generation uncertainty is bypassed entirely, achieving 100% reliability. This tiered approach ensures that LLM-assisted authoring accelerates workflow efficiency without compromising the rigorous standards required in safety-critical risk assessments.

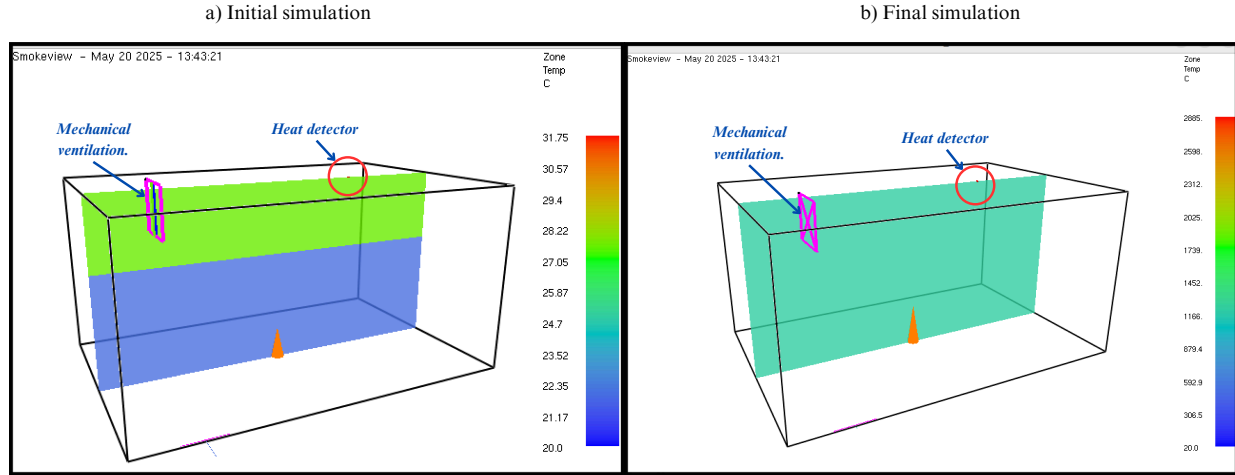


Figure 5. Risk-based decision framework mapping scenario complexity to recommended review level. Known-case queries covered by the RAG library bypass generation uncertainty and achieve 100% reliability for that subset.

4 Discussion and risk management implications

The integration of a local LLM with deterministic validation achieves an end-to-end reliability of 67.2% for NL→CFAST authoring, reducing scenario-generation time by orders of magnitude. However, this reliability is highly scenario-dependent. While simple geometries are handled with high fidelity, complex cases featuring multiple devices and floors remain challenging for the current architecture.

Consequently, the deployment of LLMs in fire engineering must be governed by strict risk management protocols. The key implications for practice are:

Quantified uncertainty: The use of bootstrap confidence intervals provides explicit, defensible reliability bounds, moving beyond anecdotal evaluations of generative AI.

Failure-mode awareness: Because the dominant generative failures are structural (e.g., missing rooms or broken vent connectivity), they are easily identifiable and can be systematically checked by trained engineering personnel.

Tiered deployment: Mapping simple, standard, and complex scenarios to appropriate levels of human oversight ensures safety standards are maintained.

For regulatory or safety-critical applications, this system should function as a reliability-enhancing assistant rather than a replacement for expert judgment. Routine benchmark cases should be served directly via RAG to ensure zero-defect generation. Conversely, novel and highly complex cases must undergo rigorous expert review, particularly when device logic or multi-compartment smoke transport drives life-safety outcomes.

5 Conclusions

This study reframes natural-language CFAST input generation as a structured reliability-engineering problem rather than a purely generative-AI task. By establishing a discrete correctness ladder (L0–L4) and formally defining end-to-end system reliability ($R=P(L2 \cap L3)$), the pipeline's performance was rigorously quantified across 150 diverse scenarios. The resulting uncertainty decomposition separated syntax, executability, and semantic errors, demonstrating that while deterministic validation effectively neutralises structural coding failures, complex semantic interpretation remains the primary constraint. To address this, a risk-tier deployment policy was formulated to provide a practical framework for integrating AI into safety-critical workflows based on scenario complexity. Ultimately, this framework supports the safe and effective adoption of LLMs in fire safety engineering by balancing automation efficiency with necessary human oversight. Broader physics validation of the generated CFAST outputs against experimental datasets (L4) and full architectural details of the FireIR compiler are deferred to the companion journal manuscript. Together, these parallel studies provide both the technical implementation and the reliability boundaries required for modern, AI-assisted fire zone modelling.

References

- Areerob, K., Nguyen, V., Li, X., Inadomi, S., Shimada, T., Kanasaki, H., Wang, Z., Suganuma, M., Nagatani, K., Chun, P., and Okatani, T., Multimodal Artificial Intelligence Approaches Using Large Language Models for Expert-Level Landslide Image Analysis, *Comput.-Aided Civ. Infrastruct. Eng.*, 40(19), 2900-2921, 2025.
- Brown, T., Mann, B., Ryder, N., Subbiah, M. M., Kaplan, J. D., Dhariwal, P., and Amodei, D., Language Models are Few-Shot Learners, *Adv. Neural Inf. Process. Syst.*, 33, 1877-1901, 2020.
- Chun, P. and Kikuta, T., Self-Training with Bayesian Neural Networks and Spatial Priors for Unsupervised Domain Adaptation in Crack Segmentation, *Comput.-Aided Civ. Infrastruct. Eng.*, 39(17), 2642-2661, 2024.
- Fei, Y., Liao, W., Zhang, S., Yin, P., Han, B., Zhao, P., Chen, X., and Lu, X., Integrated Schematic Design Method for Shear Wall Structures: A Practical Application of Generative Adversarial Networks, *Buildings*, 12(9), 1295, 2022.
- Gomez, R. S., Porto, T. R. N., Magalhães, H. L. F., Santos, A. C. Q., Viana, V. H. V., Gomes, K. C., and Lima, A. G. B., Thermo-Fluid Dynamics Analysis of Fire Smoke Dispersion and Control Strategy in Buildings, *Energies*, 13(22), 6000, 2020.
- Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., Wiest, O., and Zhang, X., Large Language Model Based Multi-Agents: A Survey of Progress and Challenges, arXiv preprint arXiv:2402.01680, 2024.
- Hattori, K., Oki, K., Sugita, A., Sugiyama, T., and Chun, P., Deep Learning-Based Corrosion Inspection of Long-Span Bridges with BIM Integration, *Heliyon*, 10(15), e35308, 2024.
- Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Zhang, C., Wang, J., Wang, Z., Yau, S. K. S., Lin, Z., Zhou, L., Ran, C., Xiao, L., and Wu, C., MetaGPT: Meta Programming for a Multi-Agent Collaborative Framework, arXiv preprint arXiv:2308.00352, 2023.
- Inadomi, S. and Chun, P., Spatially Aware Markov Chain-Based Deterioration Prediction of Bridge Components Using a Graph Transformer, *Comput.-Aided Civ. Infrastruct. Eng.*, 40(14), 1932-1955, 2025.
- Jiang, F., Ma, J., Webster, C. J., Li, X., and Gan, V. J. L., Building Layout Generation Using Site-Embedded GAN Model, *Autom. Constr.*, 151, 104888, 2023.
- Ko, J., Ennemoser, B., Yoo, W., Yan, W., and Clayton, M. J., Architectural Spatial Layout Planning Using Artificial Intelligence, *Autom. Constr.*, 154, 105019, 2023.
- Liao, W., Lu, X., Huang, Y., Zheng, Z., and Lin, Y., Automated Structural Design of Shear Wall Residential Buildings Using Generative Adversarial Networks, *Autom. Constr.*, 132, 103931, 2021.
- Liu, H. and Li, L., On Languageing a Simulation Engine: Rapid Modeling of Nanoporous Media Sorption by Hierarchical Language Model, *Mater. Today Commun.*, 40, 109809, 2024.
- Maharjan, S. and Chun, P., A Large Language Model-Driven Framework for Automated Bridge Specification Generation and Simulation Validation, *Comput.-Aided Civ. Infrastruct. Eng.*, 100014, 2026.
- Peacock, R. D., Forney, G. P., and Reneke, P. A., CFAST – Consolidated Model of Fire Growth and Smoke Transport (Version 6) Technical Reference Guide, NIST SP 1026r1, National Institute of Standards and Technology, Gaithersburg, MD, 2013.
- Peacock, R. D., Reneke, P. A., and Forney, G. P., CFAST – Consolidated Model of Fire Growth and Smoke Transport (Version 7) Volume 2: User's Guide, NIST TN 1889v2, National Institute of Standards and Technology, Gaithersburg, MD, 2015.
- Qi, Y., Xu, R., and Chu, X., FeaGPT: An End-to-End Agentic-AI for Finite Element Analysis, arXiv preprint arXiv:2510.21993, 2025.
- Rianto, S., Zeng, Y., Huang, X., and Lu, X., Generative Artificial Intelligence for Fire Scenario Analysis in Complex Building Design Layouts, *Fire Saf. J.*, 155, 104427, 2025.
- Rozière, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., Tan, X. E., Adi, Y., Liu, J., Sauvestre, R., Remez, T., Rapin, J., Kozhevnikov, A., Evtimov, I., Bitton, J., Bhatt, M., Ferrer, C. C., Grattafiori, A., Xiong, W., Défossez, A., and Synnaeve, G., Code Llama: Open Foundation Models for Code, arXiv preprint arXiv:2308.12950, 2024.
- Wang, L., Liu, J., Zeng, Y., Cheng, G., Hu, H., Hu, J., and Huang, X., Automated Building Layout Generation Using Deep Learning and Graph Algorithms, *Autom. Constr.*, 154, 105036, 2023.
- Xu, L., Mohaddes, D., and Wang, Y., LLM Agent for Fire Dynamics Simulations, arXiv preprint arXiv:2404.00132, 2024.
- Yamane, T., Chun, P., Dang, J., and Okatani, T., Deep Learning-Based Bridge Damage Cause Estimation from Multiple Images Using Visual Question Answering, *Struct. Infrastruct. Eng.*, 1-14, 2024.
- Yamashita, M., Kawanishi, K., Hashizume, K., and Chun, P., Infrared Thermography and 3D Pavement Surface Unevenness Measurement Algorithm for Damage Assessment of Concrete Bridge Decks, *Comput.-Aided Civ. Infrastruct. Eng.*, 40(19), 2772-2792, 2025.
- Zafar, A. and Chun, P., An Offline Large-Language-Model Pipeline for Rapid Zone-Fire Simulation Using CFAST, Manuscript submitted for publication, 2025.
- Zhang, Y., Lin, S., Xiong, Y., Li, N., Zhong, L., Ding, L., and Hu, Q., Fine-Tuning Large Language Models for Interdisciplinary Environmental Challenges, *Environ. Sci. Ecotechnol.*, 27, 100608, 2025.