BuildSafe: A Generative AI Framework for Joint Construction Safety, Permit, and Community-Impact Reasoning from Linked NYC DOB-OSHA-311 Data

Shahid Ali[1*], Ammar Alzarrad[2], Hwapyeong Song[3], Husnu S. Narman[4]

[1]Department of Computer Science, Marshall University, One John Marshall Drive, Huntington, WV 25755; e-mail: shahidali@marshall.edu

[2]Department of Civil Engineering, Marshall University, One John Marshall Drive, Huntington, WV 25755; E-mail: alzarrad@marshall.edu

[3]Department of Computer Science, Marshall University, One John Marshall Drive, Huntington, WV 25755; e-mail: song24@marshall.edu

[4]Department of Computer Science, Marshall University, One John Marshall Drive, Huntington, WV 25755; e-mail: narman@marshall.edu

*** Correspondence:**
Shahid Ali
shahidali@marshall.edu

Keywords: Generative Artificial Intelligence; Large Language Models; Construction Safety; Urban Governance; Building Permits; Risk Prediction; New York City.

Abstract

BuildSafe investigates whether domain-adapted generative models can turn routine administrative text permit filings, incident narratives, and citizen complaints into early-warning signals for construction risk in New York City. The study links Department of Buildings (DOB) records, OSHA enforcement cases, and 311 complaints into a shared schema. It uses Gemini 1.5 Flash to generate three-section annotations covering job-site hazards, permit/code requirements, and community impacts. Three open models (Gemma-3-1B, Llama-3.2-3B, and Mistral-7B-Instruct-v0.3, 4-bit) are then fine-tuned with LoRA on this corpus using a unified prompt template, and evaluated with BERTScore, ROUGE, BLEU, METEOR, and bootstrap confidence intervals on a 2,833-record held-out test set, and supplemented with a 150-record two-expert consensus validation subset used for targeted human evaluation of factual accuracy, completeness, and hallucinations. Across the short-run LoRA adaptation setting, all models showed small train-validation-test gaps, suggesting stable proof-of-concept adaptation without obvious overfitting rather than evidence of full convergence. Llama-3.2-3B is the clear leader on lexical overlap metrics, Mistral-7B-Instruct delivers the strongest semantic fidelity and most structured narrative outputs, and Gemma-3-1B offers competitive semantic performance at a fraction of the compute cost, suggesting potential value for edge or real-time triage workloads. A qualitative case study of Manhattan permit records illustrates how these models produce multi-section safety narratives and reveals trade-offs among completeness, specificity, and verbosity. The paper highlights key limitations, most notably the reliance on machine-generated reference labels, the limited scale of human expert evaluation, and the NYC-specific regulatory context, and proposes a roadmap for future operational validation rather than claiming demonstrated deployment impact.

Priorities include richer and better-linked data sources, retrieval-augmented and cascaded inference architectures, impact- and fairness-oriented evaluation in agency pilots, and human-in-the-loop governance suited to high-stakes public-sector AI systems.

1 Introduction

The construction sector remains one of the most hazardous industries worldwide, with persistent challenges around accident prevention, regulatory compliance, and community disruption (Tang, 2024; Usama et al., 2024). In dense urban environments such as New York City (NYC), these challenges intensify: projects must navigate complex building code and permitting requirements while minimizing impacts on surrounding neighborhoods (Awolusi et al., 2022; Ceccato, 2013). Empirical analyses of construction injuries and incident patterns show that reactive safety management based on post hoc investigation and fragmented data has not delivered sustained improvements in safety performance (Awolusi et al., 2022; Pilskog Orvik, 2024; Tixier and Hallowell, 2023).

Cities like NYC now publish rich, complementary data describing construction activity and its consequences. NYC Department of Buildings (DOB) housing and permitting datasets provide free-text job descriptions and structured attributes for new buildings, major alterations, and demolitions (New York City Department of City Planning, n.d.; Nycdb, n.d.). OSHA enforcement and injury-reporting data add incident narratives, cited standards, and employer-reported injuries and illnesses (Occupational Safety and Health Administration, 2025; Open Knowledge Foundation, n.d.). NYC 311 service requests capture resident complaints about noise, unsafe building conditions, street obstructions, and related quality-of-life issues in fine spatial and temporal detail (City of New York, 2026; Mulligan et al., 2019; Tussey and Yan, 2025). Individually, these datasets support focused analyses of incidents, delays, or community response; collectively, they describe a lifecycle from permit filing to realized safety outcomes and neighborhood complaints (Ceccato, 2013; Mulligan et al., 2019; Zou and Ergan, 2019). However, they are maintained in separate systems with heterogeneous schemas and are rarely linked at scale, limiting the use of historical patterns when triaging new permit applications (Gao et al., 2023; Kamil et al., 2024; Yang et al., 2017).

Recent AI research in construction safety has advanced along three main streams: vision and sensor-based monitoring of unsafe site conditions, statistical or machine-learning prediction of accidents and violations, and emerging generative AI systems that summarize or explain safety risks. These approaches have improved task-specific decision support, but they remain limited as foundations for pre-construction urban governance. Vision and sensor systems typically operate after work has begun and depend on observable site conditions. Traditional predictive models often estimate isolated outcomes such as injury likelihood, violation risk, or complaint frequency. Early LLM-based systems can generate safety guidance, but they rarely connect permit descriptions, enforcement histories, and neighborhood complaint signals within a unified governance workflow. Thus, the main limitation is not technical immaturity but misalignment between model design and the cross-agency information environment of public construction oversight. BuildSafe addresses this limitation by shifting the unit of analysis from isolated safety events to linked permit-enforcement-complaint records that support earlier and more explainable risk triage.

BuildSafe differs from prior AI-based construction safety frameworks in four specific ways. First, it treats construction safety as a linked urban-governance problem rather than a single-task hazard classification or accident prediction problem. Second, it integrates DOB permit records, OSHA enforcement and injury narratives, and 311 complaint histories into one administrative-civic corpus, whereas most prior systems analyze safety, permitting, or community disruption separately. Third, it

formulates the task as tri-section generative reasoning: a single prompt-output template jointly produces job-site hazard predictions, likely permit/code requirements, and potential community-impact signals. Fourth, it evaluates multiple open-weight LLMs under a common LoRA/QLoRA fine-tuning protocol with held-out testing, bootstrap confidence intervals, ablation analysis, zero-shot comparison, and targeted human validation. This positioning shifts the contribution from another isolated safety-prediction model toward an integrated, city-scale framework for explainable construction-risk governance.

This framing also connects BuildSafe to computational research on urban resilience. Maksoud et al. (2024a) show that computational design and multi-objective optimization can support flood-resilient urban habitats by combining performance simulation, environmental adaptation, and data-driven design evaluation. BuildSafe extends this broader logic from environmental design to construction-safety governance by using linked administrative and civic data to anticipate site-level hazards and neighborhood-scale impacts before risks fully materialize. This study therefore focuses on NYC construction job filings and asks:

- **RQ1.** Can a single generative model, fine-tuned on linked DOB-OSHA-311 data, generate useful joint predictions of hazards, permitting needs, and community impacts directly from DOB job descriptions?
- **RQ2.** Does a unified generative pipeline trained on tri-task prompts produce internally coherent outputs across the hazard, permit, and community-impact sections, and does the shared prompt-output template maintain semantic consistency across sections?

The working hypothesis is that a unified generative pipeline, trained on prompts that bundle DOB job attributes with spatio-temporal context from OSHA incidents and nearby 311 complaint histories, can learn cross-task regularities that are difficult to capture when each task is modeled separately (Gao et al., 2023; Tussey and Yan, 2025; Zou and Ergan, 2019). Examples include recurrent combinations of project scope and location that simultaneously elevate fall or struck-by risk, trigger specific NYC Building Code sections and permit types, and correlate with after-hours noise or dust complaints. By using a shared prompt-output template during fine-tuning, the model can exploit common lexical and semantic structure work types, construction methods, regulatory references, and neighborhood characteristics across the hazards, permits, and community-impacts sections, potentially improving both semantic fidelity and internal coherence (Mohamed et al., 2025; Rasheed et al., 2024; Yoo et al., 2024).

BuildSafe is positioned as an experimental framework and benchmark study rather than a deployed production system. The paper makes three contributions: (1) constructs a 90,000-prompt corpus by spatially and temporally linking NYC DOB housing jobs, OSHA enforcement and injury-reporting data, and NYC 311 service requests, and by generating structured three-section annotations (hazards, permits, community impacts) with a consistent template suitable for supervised fine-tuning (City of New York, 2026; New York City Department of City Planning, n.d.; Occupational Safety and Health Administration, 2025); (2) fine-tunes three open-weight LLMs Gemma-3-1B, Llama-3.2-3B, and Mistral-7B-Instruct (4-bit) using parameter-efficient strategies (LoRA/QLoRA) via Unsloth, creating three BuildSafe variants that differ in capacity and deployment profile (Farabet, 2025; Mistral AI, 2024; Unslothai, 2025b); and (3) evaluates these models on a held-out DOB test set using standard text-generation metrics (BERTScore, ROUGE, BLEU, METEOR), analyzing complementary strengths in semantic similarity, lexical precision, and computational efficiency in this tri-task generative setting (Jiao, 2024; Mohamed et al., 2025; Wan et al., 2024). Beyond traditional n-gram metrics such as BLEU and ROUGE, embedding-based measures like BERTScore have been shown to

correlate more strongly with human judgments on generative language tasks (Zhang et al., 2020). Finally, the paper discusses how this framework could support municipal workflows, for example by flagging high-risk projects for prioritized review, assisting permit desk staff with code citations, and highlighting projects likely to generate specific community complaints while drawing on recent work in responsible and trustworthy AI to identify key governance requirements, including safety-aware fine-tuning, bias and fairness auditing, and appropriate regulatory oversight (Agapiou, 2024; Behzadan, 2024; Hsu et al., 2024).

2 Literature Review

The literature review is organized around the progression from traditional construction-safety risk assessment to data-driven prediction, generative AI, geospatial urban analytics, and trustworthy AI governance. Rather than treating these areas as separate bodies of work, the review emphasizes how each contributes only part of the BuildSafe problem: early identification of hazards, interpretation of permit and code requirements, and anticipation of community impacts from linked urban data.

2.1 Construction Safety - Scale and Persistence

Construction remains one of the deadliest industries in the United States, accounting for roughly one in five workplace fatalities and over 1,075 deaths in 2023, with the "Fatal Four" hazards (falls, struck-by, electrocutions, caught-in/between) responsible for nearly 60% of these deaths (Occupational Safety and Health Administration, 2023). Despite mature regulations and long-recognized hazard categories, fatality rates have not declined proportionally, underscoring the limits of traditional, reactive safety management. Expanding digital reporting has created large administrative datasets. OSHA's Injury Tracking Application (ITA) mandates the electronic submission of Forms 300/300A/301 for high-hazard establishments, significantly increasing the structured nature of injury records, while sectoral efforts, such as the American Petroleum Institute's guidelines, standardize occupational injury data (American Petroleum Institute, 2019; Occupational Safety and Health Administration, 2025). However, studies highlight that these records are rarely exploited for predictive modelling, and underreporting, fragmented subcontractor records, and project-based work further complicate longitudinal analysis (Piri and Panthi, 2024; Workers' Compensation Trust, 2021). Evidence from incident analyses shows that leading indicators such as near-misses and violation patterns are present in existing data but not systematically leveraged (Awolusi et al., 2022), suggesting that the core problem is ineffective data utilization rather than data scarcity, an opportunity for AI and large language models (LLMs).

2.2 From Probability - Impact Models to Data-Driven Risk Assessment

Historically, construction risk assessment has been dominated by Probability-Impact (P-I) matrices and related expert-elicited methods, which yield ordinal risk scores, struggle with multiple objectives, and poorly capture interdependencies among risk factors (Taroun et al., 2011). Subsequent work introduced fuzzy logic, the Analytic Hierarchy Process, and Bayesian networks, yet systematic reviews still report methodological fragmentation and limited cross-study comparability (Kumi et al., 2024). Recent data-driven approaches address these limitations by fusing heterogeneous sources and learning latent relationships directly from data. Gao et al. (2023) showed that deep neural networks trained on multi-source quality data outperform traditional models, while Kamil et al. (2024) demonstrated analogous gains for heterogeneous safety data integration in the process industry. Mostofi and Toğan (2023) further increased representational power using multi-head attention graph neural networks, capturing relational accident structures but on relatively small datasets of uncertain generalizability. These limitations motivate BuildSafe's move away from static, single-score risk assessment toward a linked,

data-driven generative framework that can reason jointly over hazards, permit requirements, and community impacts.

2.3 AI and Machine Learning in Construction Safety Management

AI and ML have increasingly been used to improve hazard detection, risk prediction, and decision support in construction safety. Reviews show progress in computer vision, NLP, predictive analytics, wearables, IoT sensors, drones, VR-based safety training, and AI-enabled compliance monitoring, but they also identify persistent barriers related to data quality, interpretability, standardization, and organizational readiness (Tang, 2024; Parekh and Mitchell, 2024; Savaş, 2025; Chandu et al., 2024; Hossain et al., 2025; Raliile and Haupt, 2020; HSI, 2025). Many proposed systems remain conceptual or proof-of-concept rather than field-validated, particularly in AI-driven risk management frameworks that are not yet integrated into routine agency or contractor workflows (Rasheed et al., 2024; Usama et al., 2024; Pilskog Orvik, 2024). This literature establishes the technical promise of AI in construction safety, while also showing that many systems remain disconnected from the administrative and civic data streams used in public construction governance. The remaining challenge is therefore not simply model development, but aligning AI methods with the cross-agency information environment in which public construction oversight actually occurs.

2.4 Deep Learning and Predictive Modelling for Construction Accidents

Deep learning expands construction-safety modeling by handling unstructured text, images, and sensor streams. GPT-based accident-prediction models and saliency methods have improved interpretability, while multi-company datasets show that broader shared data can outperform firm-specific models (Yoo et al., 2024; Tixier and Hallowell, 2023). Spatial and temporal context also improve predictive performance, with multi-order spatio-temporal models outperforming approaches that rely only on spatial or temporal features alone (Jiao, 2024). However, these studies typically focus on accident prediction or site-level monitoring rather than jointly modelling permitting requirements, safety enforcement signals, and neighborhood complaint patterns. BuildSafe extends this predictive literature by shifting the modeling target from isolated accident forecasting to joint reasoning across pre-construction permit language, historical enforcement evidence, and neighborhood complaint signals.

2.5 Generative AI and Large Language Models in Construction

Generative AI and LLMs shift the focus from pattern recognition to knowledge representation and explanation, which is critical for safety applications that require contextualized, actionable guidance rather than raw scores. Mohamed et al. (2025) performed a bibliometric analysis of generative AI in construction risk management, finding rapid growth since 2020 and categorizing benefits into technical, operational, technological, and integration domains, alongside nine risk categories spanning data quality, liability, and social acceptance. Wan et al. (2024) surveyed broader generative AI applications in the building industry (design, document generation, scenario simulation), but mostly at a conceptual level. Complementing this broader view, Maksoud et al. (2024b) show that image-generative AI tools such as Midjourney are already influencing architectural concept development, creative brainstorming, and design reasoning. However, their study also cautions that speed and creativity do not necessarily guarantee technical accuracy or input fidelity. A few studies propose working systems that combine vision and language. Hussain et al. (2026) built an intelligent assistant that integrates computer vision-based hazard detection with LLM-based safety advisories, while Tang and Luo (2025) showed that multimodal architectures can produce more nuanced recommendations than either vision or language alone.

Debate persists on trustworthiness. Behzadan (2024) argues that transparent architectures, explainable outputs, and user-centered design are prerequisites for adoption, and Mohamed et al. (2025) identify stakeholder trust as a key barrier. Proponents counter that LLMs need only outperform current manual practices, which are themselves error-prone and inconsistent (Yoo et al., 2024). BuildSafe contributes empirical evidence by fine-tuning domain-specific LLMs that can be evaluated against structured safety benchmarks. This motivates BuildSafe's use of generative AI not as a generic advisory chatbot, but as a structured tri-section reasoning system whose outputs are constrained to hazards, permit/code requirements, and community impacts.

2.6 Spatio-Temporal and Geospatial Approaches to Urban Construction Risk

Urban construction risk is inherently spatial. Ceccato (2013) pioneered the use of GIS to uncover risk hotspots, corridors, and temporal clusters that non-spatial methods overlook. Dufitimana et al. (2025) combined ML with geospatial data to map socioeconomic vulnerability to natural hazards, and Yang et al. (2017) developed a georelational learning framework to integrate heterogeneous urban datasets, including permit and infrastructure records. Zou and Ergan (2019) showed that neighborhood-level features such as building density, complaint history, and land-use patterns predict construction-related disruptions, demonstrating that civic data infrastructures contain safety-relevant signals beyond site-level records. Together with Jiao's (2024) findings on multi-order spatio-temporal models, this work motivates BuildSafe's reliance on geographically resolved NYC data: spatial features are not ancillary metadata but core predictive variables.

2.7 NYC Urban Data Sources for Construction Safety Research

New York City offers an unusually rich open-data ecosystem for construction safety research, combining scale, transparency, and regulatory complexity. Key sources include NYC 311 service requests, DOB housing and permitting records, the NYCDB consolidated housing datasets, and U.S. Department of Labor enforcement data (Mulligan et al., 2019; City of New York, 2026; New York City Department of City Planning, n.d.; Nycdb, n.d.; Open Knowledge Foundation, n.d.). Zerkin (2006) documented the broader regulatory landscape of building performance in New York City, providing historical context for the permitting and inspection environment. 311 complaints provide neighborhood-level signals of disruption and potential unsafe conditions, DOB and housing records capture where and when construction occurs, and federal enforcement and ITA datasets provide standardized injury and violation data. However, these sources were not designed for joint analysis: they differ in geographic identifiers, temporal granularity, and categorical schemas. Prior work has begun to tackle geo-alignment and linkage challenges (Yang et al., 2017; Tussey and Yan, 2025), but a fully integrated, construction-safety-focused dataset has not been demonstrated. BuildSafe's NYC DOB-OSHA-311 corpus directly addresses this gap. The importance of these datasets is therefore not only their scale, but their complementarity. DOB records describe proposed construction activity before work begins, OSHA records describe realized safety and compliance failures, and 311 complaints capture neighborhood-level disruption that may not appear in formal safety records. Prior urban analytics studies have shown the value of geospatial and civic data, but they rarely convert these linked administrative traces into a structured generative-reasoning task. BuildSafe uses this complementarity to position construction safety as a city-scale information-integration problem rather than a site-level monitoring problem alone. Section 4.7 later revisits this city-scale framing in relation to urban informatics, computational urban resilience, and AI-supported architectural design workflows.

2.8 Transfer Learning and Cross-domain Adaptation

Labelled construction safety data are scarce relative to the complexity of the tasks, making transfer learning (TL) particularly attractive. Junjia et al. (2024) classify TL techniques used in construction risk management as instance-based, mapping-based, network-based, and adversarial and document applications across computer vision, NLP, and expert systems, while highlighting ongoing challenges such as domain shift and the lack of benchmark datasets. Tixier and Hallowell's (2023) multi-company findings support the broader TL insight that wider source knowledge improves target-domain performance. In LLMs, fine-tuning can be viewed as network-based transfer learning: a large model pre-trained on general text is adapted to construction safety language and reasoning using domain-specific data. BuildSafe operationalizes this by fine-tuning open-weight models on a curated corpus of NYC construction safety data. This directly supports BuildSafe's use of LoRA/QLoRA as a resource-efficient domain-adaptation strategy for translating general-purpose open-weight LLMs into construction-safety reasoning models.

2.9 Fine-Tuning Large Language Models for Domain-Specific Tasks

Fine-tuning pre-trained LLMs has become the dominant strategy for domain specialization. Reviews by Weng (2024) and Lu et al. (2025) survey full fine-tuning, adapters, prompt tuning, and Low-Rank Adaptation (LoRA), emphasizing trade-offs between computational cost, adaptation depth, and catastrophic forgetting. Sagar (2025) shows that parameter-efficient methods like LoRA can deliver strong domain performance under tight resource constraints.

LoRA and its quantized variant QLoRA are especially attractive for academic settings. Zhao et al.'s (2024) LoRA Land report demonstrates that LoRA-adapted models can rival GPT-4 on specialized benchmarks, and frameworks such as Unsloth provide substantial speed and memory savings (Unslothai, 2025b; Unsloth, 2026), enabling fine-tuning of multi-billion-parameter models on single GPUs. This ecosystem underpins BuildSafe's resource-efficient fine-tuning strategy. Safety preservation during fine-tuning is a critical concern in safety-critical domains. Choi et al. (2024) propose safety-aware fine-tuning methods that maintain alignment, and Hsu et al. (2024) introduce Safe LoRA, constraining weight updates to a "safe" subspace. While BuildSafe does not yet implement these methods, their principles inform conservative hyperparameter choices and structured prompt templates, and they define a clear path for future safety-aware domain adaptation.

2.10 Base Model Selection - Open-Weight Architectures

Among open-weight models, Mistral-7B and Google's Gemma families are particularly relevant for resource-constrained research. Mistral-7B combines strong instruction-following performance with architectural innovations such as grouped-query and sliding-window attention, and is widely available in 4-bit quantized form suitable for single-GPU fine-tuning (Mistral AI, 2024; Data Science Dojo, 2025; Hugging Face, 2025a; Miller, 2024; Ubiai, 2024). Gemma models offer competitive performance and tight integration with Google's tooling, with versions explicitly designed to run on single GPUs or TPUs (Team G. et al., 2024; Farabet, 2025). For BuildSafe's core task, structured extraction and generation of risk-relevant information from DOB job descriptions and related context, these architectures provide a practical balance between capacity, performance, and deployability, while illustrating that domain-adapted LLMs no longer require proprietary, closed-weight bases.

2.11 Regulatory Frameworks, Ethics, and Trustworthy AI in Construction

Deploying AI for safety-critical decisions raises socio-legal and ethical challenges. Agapiou (2024) reviews responsible AI in construction health and safety and concludes that existing legal frameworks do not yet adequately address algorithmic bias, opacity, and liability allocation when AI contributes to

safety failures. More broadly, Huang et al. (2024), Sargiotis (2024), and Wörsdörfer (2025) highlight the rapid expansion of AI applications in civil engineering and the uneven global trajectories of AI regulation, reinforcing the need for stronger governance in safety-critical built-environment systems. Yang et al. (2024) propose a lifecycle framework for trustworthy AI in construction, spanning planning, data collection, algorithm development, deployment, and maintenance, and mapping principles such as safety, explainability, transparency, and fairness to concrete actions. For fine-tuned LLMs like BuildSafe, this implies that predictive accuracy must be accompanied by transparency about training data and limitations, mechanisms for human oversight, and alignment with data privacy and liability norms. Accordingly, the present study treats ethics and trustworthiness not only as background concerns but as deployment requirements, with Section 4.11 translating these principles into measurable fairness audits, privacy and cybersecurity controls, audit logging, and accountability procedures.

2.12 Research Gaps and Rationale for the Present Study

Taken together, the literature shows that construction-safety AI has progressed from expert-driven risk matrices to machine-learning prediction, computer-vision monitoring, geospatial analysis, and early generative-AI decision support. However, these streams remain only partially connected. Sensor, wearable, drone, and vision-based systems are most useful after work has begun, while permit-review agencies require risk signals earlier, when administrative records, project descriptions, location attributes, enforcement histories, and complaint patterns are the main available evidence. Traditional ML and deep learning studies improve prediction, but they usually treat accident occurrence, violation risk, and complaint frequency as separate targets rather than as interrelated elements of urban construction governance. Generative-AI studies broaden the scope by producing explanatory outputs, but many remain conceptual, rely on generic prompting, or do not integrate permitting, enforcement, and civic-impact data into a reproducible training and evaluation workflow.

This synthesis leaves four focused gaps that motivate BuildSafe. First, few studies evaluate working, domain-adapted LLMs against a structured construction-safety benchmark. Second, DOB permits, OSHA records, 311 complaints, and geospatial indicators are usually analyzed separately rather than as a linked training corpus. Third, resource-efficient LLM adaptation methods such as LoRA and QLoRA remain underexplored for construction-safety reasoning, despite their relevance for academic and public-sector settings with limited computing resources. Fourth, safety-aware adaptation and governance-oriented evaluation remain insufficiently operationalized, even though unsupported AI outputs could affect permitting, inspection, and public-sector decision workflows. BuildSafe addresses these gaps by testing whether linked administrative, enforcement, and civic data can support a unified generative framework for construction-safety reasoning.

3 Methodology

3.1 Data Acquisition and Sources

3.1.1 NYC DOB Housing Database

The core inventory of construction activity is the New York City Department of City Planning's Housing Database, which consolidates Department of Buildings (DOB) job filings affecting residential unit counts across the five boroughs beginning 1 January 2010 (New York City Department of City Planning, n.d.). The database distinguishes three primary job types: new buildings, Alteration Type 1 (major alterations that change dwelling-unit counts), and demolitions, and provides project-level records with attributes such as borough, block and lot, address, job type, and pre- and post-development

unit counts, all geocoded to support spatial joins. The Housing Database is distributed via NYC Open Data and BYTES of the BIG APPLE as CSV files. It is mirrored in a version-controlled PostgreSQL schema by the open-source nycdb project, which simplifies automated ingestion (Nycdb, n.d.). In BuildSafe, this database serves as the authoritative list of construction projects against which other safety-relevant signals are spatially and temporally aligned.

3.1.2 OSHA Enforcement and Injury-Tracking Data

Workplace safety outcomes are characterized using enforcement and injury data released by the U.S. Department of Labor. The US DOL enforcement data portal exposes OSHA inspection, violation, and penalty records as bulk CSV exports and APIs, linking inspection identifiers, cited standards, employer information, and penalties (Open Knowledge Foundation, n.d.). OSHA's Injury Tracking Application (ITA) complements these tables with establishment-specific illness and injury records from Forms 300, 300A, and 301 (Occupational Safety and Health Administration, 2025). Narrative fields in these datasets, such as alleged violation descriptions and incident summaries, encode fine-grained information about hazard types and failure modes that structured variables alone cannot capture. These narratives are used to expose the model to authentic descriptions of construction-related incidents and regulatory framing.

3.1.3 NYC 311 Service Requests

To capture community-level signals of construction impact, the study incorporates NYC 311 service requests, the city's centralized non-emergency reporting system. The dataset contains millions of complaints annually about noise, dust, unsafe building conditions, and street obstructions, with each record including timestamps, complaint type, responsible agency, and geocoded location (Mulligan et al., 2019; City of New York, 2026). The data are regularly updated on NYC Open Data and support both bulk downloads and API access, enabling large-scale spatio-temporal analysis. Complaint categories related to buildings and noise are particularly relevant for understanding public perceptions of construction activity. By aggregating complaint counts within spatial buffers around DOB project locations, BuildSafe infers patterns of community disruption associated with different project types.

3.2 Data Preprocessing and Integrations

3.2.1 Ingestion and Encoding Normalization

Fifteen heterogeneous CSV files spanning NYC 311, DOB housing records, OSHA enforcement tables, and auxiliary datasets constitute the raw input corpus. Each file is ingested into pandas using a two-stage decoding strategy: the pipeline first attempts UTF-8 and, on failure, retries with Latin-1 while skipping corrupted lines. For each file, the script logs the encoding used, row counts, and any errors, providing traceability for later data-quality audits.

3.2.2 Schema Harmonization

Because source systems use divergent naming conventions, a predefined mapping aligns each file's columns to a common minimal schema. This mapping harmonizes semantically equivalent fields (for example, boro and borough, job_desc and job_description) and drops attributes not needed for risk reasoning. Records lacking critical fields such as geographic identifiers or core descriptive text are removed, and remaining nulls in retained columns are converted to the literal string "N/A" to ensure syntactic completeness when forming prompts. The result is a consistent schema across all input tables with sufficient expressive power for safety assessment.

3.2.3 Sampling Strategy and Full-Corpus Construction

An initial deterministic sampling step selects up to five records per file as a quality-assurance subset to validate the annotation prompt template and verify response quality from Gemini 1.5 Flash. Using fixed random seeds, this per-file cap yields a balanced sample across datasets and complaint or incident types. After confirming that the three-section annotation structure (hazards, permits, community impacts) was reliably produced, the pipeline was extended to all records across the 15 source files that passed schema harmonization and null-filtering. In the full annotation run, every eligible record was submitted to Gemini 1.5 Flash using a fixed prompt template and a rate-limiting protocol. The resulting corpus comprises approximately 90,000 prompt-output pairs, each derived from a single DOB, OSHA, or 311 record's descriptive fields. The data were partitioned into training, validation, and test sets using a stratified random split with a fixed seed (3407), preserving the relative proportions of DOB, OSHA, and 311 sources across the splits. The test set contains 2,833 prompt-output pairs; the remaining records are divided between training and validation at roughly 80/20. Records with malformed or incomplete annotations, for example, outputs missing one or more of the three required sections, were excluded, with an overall exclusion rate below 2% of submitted records.

3.3 Annotation Pipeline and Dataset Construction

3.3.1 Prompt Construction

For each record, a helper routine constructs a multi-line text block in which each line follows the pattern field name: field value. When a dedicated description field exists (for example, a DOB job description or OSHA narrative), it is prominently included; otherwise, the composite block serves as the primary contextual description. This representation preserves column semantics while presenting the model with a human-readable summary. The text block is embedded into a fixed prompt template instructing the model to: (1) assess the overall risk level associated with the described situation or project; (2) identify relevant permits, regulations, or standards likely to apply; (3) anticipate potential worker safety and community impacts; and (4) flag obvious compliance concerns or missing safeguards. This standardized design ensures that annotations across heterogeneous records are comparable and suitable as labels for supervised fine-tuning.

3.3.2 LLM Invocation and Rate Limiting

The annotation pipeline calls Google's Gemini 1.5-Flash model via a configured client, which sends one prompt per record. To comply with usage policies (e.g., a maximum of about 15 requests per minute), the script enforces a fixed delay of approximately 4 seconds between calls. Structured exception handling logs transport errors, rate-limit responses, and other API faults; failed records are retried with backoff or marked for exclusion depending on error type.

3.3.3 Result Aggregation and Final Dataset

Each successful annotation is stored with its provenance (source file, row index, and exact prompt text). After processing all records, the annotations are consolidated into an intermediate CSV file that includes both raw prompts and model-generated outputs. A final projection step extracts two canonical columns for fine-tuning: prompt and output. Where available, prompt is set to the original human-written description; otherwise, it defaults to the composite text block. The output column stores the unmodified annotation string returned by Gemini. The resulting two-column dataset is model-agnostic and suitable for supervised fine-tuning of any compatible LLM.

3.3.4 Annotation Reliability and Silver-Standard Supervision

The Gemini-generated annotations are treated in this study as silver-standard supervision rather than as definitive human ground truth. This distinction is central to the experimental design because the annotation task requires contextual safety reasoning, regulatory interpretation, and community-impact inference, all of which may be affected by model hallucinations, omissions, or overgeneralizations. Gemini 1.5 Flash was used because it provided a scalable, consistent mechanism for generating structured three-section annotations across heterogeneous DOB, OSHA, and 311 records using a fixed prompt template. The annotation task was constrained rather than open-ended: each administrative record was mapped into three predefined sections: job-site hazards, permit/code requirements, and community impacts, so the model functioned primarily as a structured extraction and safety-reasoning annotator. This use is consistent with recent construction and built-environment literature that treats generative AI as a tool for document interpretation, compliance-oriented reasoning, and risk-management support, while emphasizing the need for safeguards against hallucination, bias, and overgeneralization (Mohamed et al., 2025; Tan et al., 2024; Wan et al., 2024). However, the resulting labels should be interpreted as machine-assisted reference outputs suitable for proof-of-concept supervised fine-tuning, not as fully verified expert regulatory determinations.

The reliability of the silver-standard corpus was strengthened through four safeguards. First, all records were annotated using a fixed prompt template requiring the same three sections: hazards, permit/code requirements, and community impacts. This reduced variation in output structure and allowed the downstream models to learn a consistent tri-task response format. Second, each annotation was stored with source-file provenance, row index, and the exact prompt text, allowing every generated label to be traced back to the original administrative record. Third, malformed annotations were excluded when they lacked one or more required sections, resulting in an exclusion rate of less than 2% of submitted records. Fourth, the automated evaluation was supplemented by a stratified 150-record expert subset, independently annotated by two domain experts. For this subset, A.A., a civil engineering faculty member with construction safety and permitting expertise, and H.S.N., a computer science researcher with expertise in AI systems, independently authored three-section safety narratives covering hazards, permit/code requirements, and community impacts using the same output template as BuildSafe. Both experts were blinded to the Gemini-generated teacher labels and to the fine-tuned model outputs during annotation. Disagreements were resolved through discussion, and the resulting consensus narratives were treated as a human-reference benchmark for the 150-record validation subset.

Accordingly, the automated metrics reported on the full held-out test set should be interpreted as measuring agreement with a consistent teacher model across a large corpus, not as direct proof of absolute regulatory correctness. BERTScore, ROUGE, BLEU, and METEOR, therefore, quantify whether the fine-tuned models reproduce the structured safety-reasoning pattern encoded in the silver-standard references. They do not, by themselves, establish that every hazard, code requirement, or community-impact statement is legally or operationally correct. However, the 150-record two-expert consensus subset provides a separate human-reference benchmark, not derived from Gemini outputs, for evaluating whether model outputs remain factually accurate, complete, and free from unsupported regulatory or hazard claims. This design allows the study to combine scalable silver-standard supervision for model adaptation with a smaller but stronger human-grounded validation set. The purpose of the Gemini-generated labels is to test whether open-weight LLMs can be aligned with a consistent construction-safety reasoning format using linked DOB-OSHA-311 data. At the same time, the expert consensus subset provides a more rigorous check on domain plausibility and hallucination behavior. Any operational deployment would still require human review, retrieval from current regulatory sources, larger external expert validation, reporting of inter-rater agreement before

consensus resolution, and post-deployment monitoring before outputs could be used in safety-critical public-sector workflows.

3.4 Model Selection and Experimental Setup

3.4.1 Model Family Rationale

Three open-weight model families were selected for BuildSafe: Gemma, Llama, and Mistral, spanning a spectrum from lightweight to mid-sized architectures. Gemma models, derived from Gemini research, are designed for efficient deployment on single GPUs or TPUs while maintaining strong language understanding, making smaller variants attractive for low-latency triage (Team G. et al., 2024; Farabet, 2025). Llama 3.2 models provide mid-scale capacity (1B-3B parameters) and strong instruction-following performance suitable for more nuanced reasoning on modest hardware (Grattafiori et al., 2024).

Mistral-7B-Instruct-v0.3, a 7.3-billion-parameter decoder-only transformer, has been shown to perform strongly on a range of benchmarks and incorporates grouped-query attention and sliding-window attention for efficient long-sequence processing (Mistral AI, 2024; Data Science Dojo, 2025). The availability of a 4-bit quantized variant via bitsandbytes and Unsloth further reduces hardware requirements (Hugging Face, 2025a; Unslothai, 2025b). In the experiments, Gemma-3-1B serves as the most compact generative model for fast hazard and permit triage, Llama-3.2-3B as a mid-tier reasoning model, and Mistral-7B-Instruct-4bit as the primary candidate for high-fidelity risk assessment and narrative generation.

3.4.2 Parameter-Efficient Fine-Tuning with Unsloth

Adapting these base models to the construction safety domain relies on parameter-efficient fine-tuning (PEFT) via the Unsloth framework. Unsloth integrates Low-Rank Adaptation (LoRA) and quantized LoRA (QLoRA) to reduce VRAM usage and training time, with reported speedups and memory savings over naive fine-tuning (Hugging Face, 2025b; Unslothai, 2025a). This aligns with broader findings that LoRA-style adaptations can deliver domain-specific performance comparable to full fine-tuning at a fraction of the computational cost (Zhao et al., 2024; Sagar, 2025; Weng, 2024; Lu et al., 2025). In this study, LoRA modules are applied to the language, attention, and feed-forward layers while keeping the vast majority of base weights frozen, offering a practical compromise between adaptation depth and computational feasibility for academic-grade hardware.

3.4.3 Fine-Tuning Configuration

Supervised fine-tuning is conducted using the two-column prompt-output dataset, along with the SFTTrainer implementation provided by Unsloth. The configuration employs LoRA with a rank of 8 and alpha of 8, without dropout or additional bias parameters, and is applied exclusively to text-based modules. Training uses a per-device batch size of 2 with gradient accumulation set to 4, resulting in an effective batch size of 8. Optimization is performed using the 8-bit AdamW optimizer with a learning rate of 2×10^{-4} for short exploratory runs, while a reduced rate of 2×10^{-5} is recommended for longer training schedules; weight decay is set to 0.01. The learning schedule used 5 warmup steps followed by linear decay across a maximum of 60 optimization steps. This short schedule should be interpreted as a constrained-compute, parameter-efficient adaptation test rather than as evidence of full convergence or globally optimized model performance. The purpose was to determine whether small LoRA updates could align pretrained open-weight models with the BuildSafe tri-task output schema

under academic-grade hardware constraints. Because the base models already encode broad linguistic and instruction-following capabilities, the adapters primarily learned the structured "hazards-permits-community impacts" response format and domain-specific lexical patterns rather than construction-safety reasoning from scratch. Metrics were logged at every training step, and train, validation, and test performance were compared as diagnostic checks for obvious instability or overfitting. However, these stability checks do not rule out the possibility that longer training, additional random seeds, alternative learning rates, checkpoint-based early stopping, or larger hyperparameter sweeps could further improve performance. Therefore, the reported results should be interpreted as reproducible proof-of-concept adaptation results, while a full convergence study remains necessary before claiming production-level optimization. A fixed random seed of 3407 was maintained throughout data loading, sampling, and training to support reproducibility. Accordingly, the reported metrics should be interpreted as evidence of short-run adapter alignment and output-format stabilization, not as evidence that the models have reached their maximum attainable domain performance.

3.5 Human Evaluation Sub-Study

Because the main automated metrics compare model outputs against Gemini 1.5 Flash-generated references on the full test set, a targeted human evaluation study was conducted to provide a separate human-reference benchmark, not derived from Gemini outputs, for a smaller subset of records. A stratified random sample of 150 records was drawn from the 2,833-record held-out test set, preserving the proportional distribution of source types (DOB, OSHA, 311) and job types (new building, Alteration Type 1, demolition). For this subset, two domain experts independently authored human reference narratives. A.A., a civil engineering faculty member with expertise in construction safety and building permitting, and H.S.N., a computer science researcher with expertise in AI systems, each produced three-section safety narratives covering Hazards, Permit/Code Requirements, and Community Impacts using the same template as BuildSafe's outputs. Both experts were blinded to the Gemini-generated teacher labels and to the fine-tuned model outputs during this process. Disagreements between the two expert-authored narratives were resolved through discussion, and the final consensus narratives were treated as a human-reference benchmark for the 150-record subset. Because both experts are members of the author team, this consensus process strengthens reliability relative to a single-evaluator design but does not substitute for future external validation by independent agency reviewers, safety inspectors, or permitting professionals. Model outputs for this subset were assessed against the consensus human references using ordinal expert judgments for Factual Accuracy, Completeness, and Overall Usefulness, along with a hallucination assessment indicating whether the output referenced a non-existent or clearly inapplicable statute, code section, hazard category, or community-impact claim. Because this validation was designed as a targeted internal plausibility check rather than a fully powered external annotation study, the results are reported as directional expert-assessment trends rather than as definitive quantitative human-evaluation scores.

3.6 Zero-Shot, Ablation, and Task-Structure Baselines

To strengthen experimental rigor, three additional baseline and ablation analyses were included in the revised evaluation design. *First*, zero-shot baselines were evaluated by applying the original Gemma-3-1B, Llama-3.2-3B, and Mistral-7B-Instruct models to the held-out prompts without LoRA adaptation. These baselines test whether the observed BuildSafe performance reflects domain adaptation rather than only general instruction-following ability. *Second*, data-source ablations were conducted to examine the contribution of each linked urban data stream. Ablated variants were trained or evaluated using DOB-only, DOB+OSHA, DOB+311, and full DOB-OSHA-311 configurations. This design tests whether enforcement narratives and complaint histories provide distinct value beyond

permit descriptions alone. *Third*, task-structure ablations compared the unified tri-task model against single-task variants specialized for hazards, permit/code requirements, or community impacts. This comparison tests whether joint generation improves cross-section coherence or whether separate specialized models perform better on individual sections.

Traditional machine learning baselines such as logistic regression, support vector machines, random forests, and gradient-boosted trees were considered. However, they were not used as primary comparators because BuildSafe is not formulated as a single-label classification or scalar risk-scoring task. Its output is a structured, multi-section narrative that jointly explains hazard reasoning, permit/code interpretation, and community-impact assessment. A conventional ML baseline would therefore require converting the task into separate categorical or binary labels, such as a hazard category, a permit-escalation flag, or a high-complaint-risk indicator. Such a transformation would remove much of the generative reasoning and cross-sectional coherence that this study is designed to evaluate. Therefore, the most direct comparators for the present benchmark are zero-shot LLM outputs, data-source ablations, and task-structure ablations under the same prompt-output format. The study does not claim superiority over traditional ML methods on matched classification tasks; instead, future work should derive simplified classification or ranking targets from the BuildSafe corpus to enable fair comparison with logistic regression, support vector machines, random forests, gradient boosting, and retrieval-based non-generative systems.

3.6.1 Baseline Implementation Details

For the zero-shot comparison, the original unfine-tuned Gemma-3-1B, Llama-3.2-3B, and Mistral-7B-Instruct models were evaluated on the same held-out prompt set used for the fine-tuned BuildSafe models. The same input template, output-section requirements, evaluation scripts, and text-generation metrics were used so that differences reflected the effect of LoRA adaptation rather than differences in prompting or scoring. For data-source ablations, the DOB-only, DOB+OSHA, DOB+311, and full DOB-OSHA-311 configurations were compared using the same tri-task output structure to examine how each linked data stream affected hazards, permit/code reasoning, and community-impact narratives. For task-structure ablations, single-section variants were compared with the unified tri-task formulation to assess whether separating hazards, permits/code requirements, and community impacts improved section-specific accuracy or weakened cross-section coherence. These additional analyses test directional robustness and task-design sensitivity rather than final production optimization.

4 Results and Discussion

The results are interpreted in light of the annotation design. Because the full-scale automated metrics compare fine-tuned model outputs with Gemini-generated silver-standard references, they primarily measure consistency with a structured teacher model rather than verified regulatory correctness. The two-expert consensus subset provides a separate human-reference check on factual accuracy, completeness, usefulness, and hallucination behavior. Accordingly, the metric tables are treated as evidence of how well each model reproduces the BuildSafe tri-task reasoning format, not as proof that the outputs constitute legally validated regulatory determinations. Because BuildSafe generates open-ended three-section narratives, exact n-gram metrics such as BLEU, ROUGE, and METEOR are interpreted as indicators of lexical consistency. BERTScore is used as the primary semantic-similarity metric, while the remaining metrics help assess how closely each model follows the wording and structure of the silver-standard references. Detailed model-to-model interpretation is consolidated in Section 4.4, and qualitative output behavior is examined in Section 4.7.

4.1 Gemma-3-1B

Gemma-3-1B is the compact BuildSafe variant intended for rapid first-pass screening under limited compute conditions. Table 1 shows that its train, validation, and test scores remain tightly aligned, with test BERTScore F1 = 0.7703, indicating stable short-run adaptation without obvious overfitting. Figure 1 places Gemma within the broader model comparison, where its main value is low-compute semantic screening rather than detailed regulatory phrasing.

Table 1. Fine-Tuning Results of Gemma-3-1B

Metric	Train	Validation	Test
BLEU	0.002507	0.002551	0.002444
BERTScore Precision	0.770643	0.770674	0.769623
BERTScore Recall	0.771463	0.771450	0.771280
BERTScore F1	0.770924	0.770929	0.770319
ROUGE-1	0.116625	0.116443	0.115389
ROUGE-2	0.048707	0.048967	0.047900
ROUGE-L	0.078756	0.078706	0.078486
ROUGE-LSum	0.112053	0.111912	0.111150
METEOR	0.020600	0.020636	0.020598

Gemma-3-1B's main value is computational efficiency. Despite being the smallest model, it maintains stable semantic performance across the train, validation, and test sets. Its weaker lexical-overlap scores suggest that it is less appropriate for detailed regulatory phrasing or complete reviewer-facing narratives, but useful for rapid first-pass screening where low latency and modest hardware requirements are priorities.

4.2 Llama-3.2-3B

Llama-3.2-3B is the lexical precision leader among the BuildSafe variants. Table 2 shows that it achieves the strongest surface-form alignment, including test ROUGE-1 = 0.1470 and METEOR = 0.0379, while maintaining high semantic similarity with test BERTScore F1 = 0.7742. This profile makes Llama especially suitable for permit-review support, where consistent terminology and standardized reviewer summaries are important.

Table 2. Fine-Tuning Results of Llama-3.2-3B

Metric	Train	Validation	Test

BLEU	0.009324	0.009458	0.009143
BERTScore Precision	0.769804	0.769924	0.768811
BERTScore Recall	0.779874	0.780015	0.779851
BERTScore F1	0.774722	0.774850	0.774214
ROUGE-1	0.148643	0.149000	0.147028
ROUGE-2	0.054165	0.054723	0.053449
ROUGE-L	0.095700	0.095653	0.094863
ROUGE-LSum	0.139127	0.139381	0.137762
METEOR	0.037639	0.037821	0.037862

Llama-3.2-3B provides the strongest lexical consistency, with the highest BLEU, ROUGE-1, and METEOR scores among the three models. This profile makes it suitable for workflows where standardized wording, repeated permit-review language, and consistent reviewer summaries are important. Its strength is not that it produces the most elaborate explanations, but that it most closely follows the reference structure and terminology.

4.3 Mistral-7B-Instruct-v0.3 (4-bit)

Mistral-7B-Instruct provides the strongest semantic and narrative profile in the BuildSafe suite. Table 3 shows the highest test BERTScore F1 value, 0.7747, despite weaker lexical-overlap scores than Llama. This pattern suggests that Mistral preserves meaning while using less reference-matched wording, making it more appropriate for complex or escalated cases that require coherent explanatory narratives rather than short standardized summaries.

Table 3. Fine-Tuning Results of Mistral-7B-Instruct (4-bit)

Metric	Train	Validation	Test
BLEU	0.00244	0.00247	0.00237
BERTScore Precision	0.7750	0.7751	0.7746
BERTScore Recall	0.7754	0.7755	0.7748
BERTScore F1	0.7752	0.7753	0.7747
ROUGE-1	0.1169	0.1168	0.1156
ROUGE-2	0.0492	0.0494	0.0483

ROUGE-L	0.0791	0.0790	0.0787
ROUGE-LSum	0.1124	0.1122	0.1114
METEOR	0.02045	0.02050	0.02046

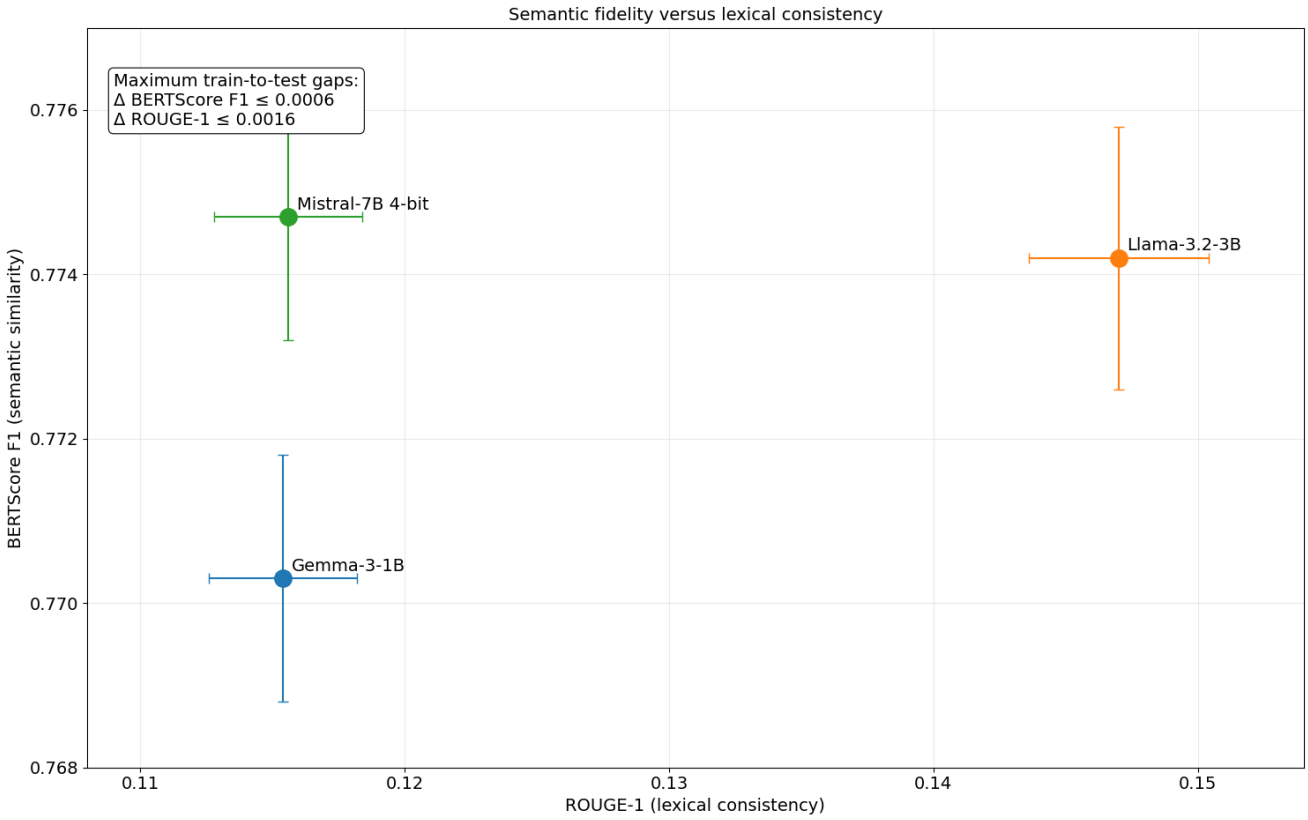
Mistral-7B-Instruct provides the strongest semantic profile, achieving the highest BERTScore F1 despite weaker lexical overlap than Llama. This pattern suggests that Mistral tends to preserve meaning while using more compressed or less reference-matched wording. It is therefore better suited to escalated cases where explanatory coherence matters more than exact reproduction of teacher-reference phrasing.

4.4 Comparative Analysis

Rather than treating the three models as independent case descriptions, this section compares them along four dimensions that matter for construction-governance use: semantic fidelity, lexical consistency, generalization stability, and computational feasibility. Table 4 summarizes the main test-set metrics for Gemma-3-1B, Llama-3.2-3B, and Mistral-7B-Instruct, highlighting the trade-off between surface-form reproduction and meaning preservation.

Table 4. Test-Set Metric Summary Across Models

Model	BLEU	BERTScore F1	ROUGE-1	METEOR
Gemma-3 1B	0.002444	0.7703	0.1154	0.0206
Llama-3.2 3B	0.00914	0.7742	0.1470	0.0379
Mistral-7B 4-bit	0.00237	0.7747	0.1156	0.0205



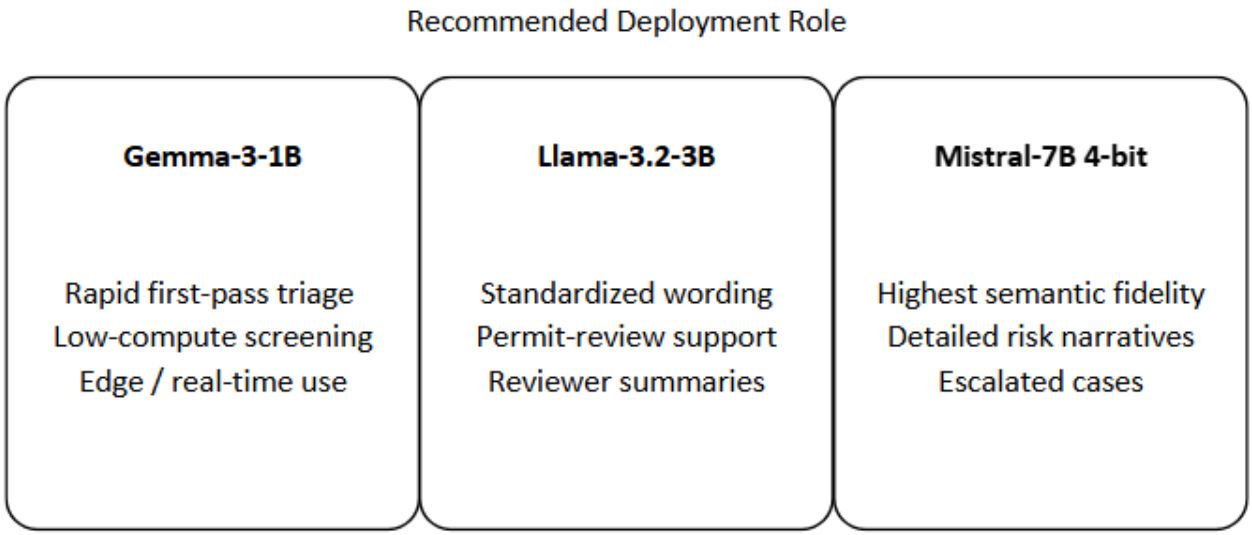


Figure 1. Comparative performance and deployment positioning of BuildSafe model variants

(A) BERTScore F1 plotted against ROUGE-1 for the three BuildSafe variants, with 95% bootstrap confidence intervals. The plot highlights the trade-off between semantic fidelity and lexical consistency: Llama-3.2-3B shows the strongest lexical overlap, while Mistral-7B-Instruct achieves marginally higher semantic similarity. Models positioned toward the upper-right region of the plot represent the strongest combined semantic and lexical balance; Llama-3.2-3B's placement indicates lexical dominance with strong semantic fidelity, whereas Mistral-7B-Instruct's placement reflects semantic depth with more compressed lexical overlap. (B) Recommended workflow roles implied by

the performance profiles, positioning Gemma-3-1B for rapid first-pass triage, Llama-3.2-3B for standardized permit-review support, and Mistral-7B-Instruct for detailed explanatory review in escalated cases.

Table 5 translates these metric patterns into practical model roles. Instead of ranking the models by a single aggregate score, the table links each model's empirical strength to the type of BuildSafe workflow for which it is most appropriate.

Table 5. Model-Level Interpretation and Recommended Deployment Role

Model	Main empirical strength	Main limitation	Recommended BuildSafe role
Gemma-3-1B	Stable semantic performance with lowest compute demand	Weaker lexical overlap and less complete narratives	Rapid triage and edge/low-latency screening
Llama-3.2-3B	Strongest lexical consistency across BLEU, ROUGE-1, and METEOR	Less semantically rich than Mistral for longer explanations	Permit-review support and standardized reviewer summaries
Mistral-7B-Instruct 4-bit	Highest BERTScore F1 and strongest long-form narrative coherence	Higher compute demand and weaker lexical overlap than Llama	Escalated cases requiring detailed risk narratives

Rather than identifying a single universally superior model, Table 5 frames the results as a tiered decision-support strategy. This role-based interpretation is more relevant for municipal deployment than ranking models by one aggregate metric because each model occupies a different point in the trade-off among compute demand, lexical consistency, and explanatory depth.

4.5 Statistical Reliability of Metric Estimates

To assess whether observed differences reflect genuine performance distinctions rather than sampling noise, 95% bootstrap confidence intervals were computed on the test set (n = 2,833). For each model and metric, 10,000 bootstrap resamples were drawn with replacement, and the 2.5th and 97.5th percentiles were recorded as interval bounds. Table 6 reports the resulting 95% bootstrap confidence intervals for all test-set metrics and models. The bootstrap analysis reveals two patterns.

First, Llama's advantage on lexical metrics is statistically robust: its ROUGE-1 interval [0.1436, 0.1504] does not overlap with those of Gemma [0.1126, 0.1182] or Mistral [0.1128, 0.1184], confirming that the ≈28% ROUGE-1 lead reflects a genuine performance difference. The same non-overlapping separation holds for BLEU and METEOR.

Second, the BERTScore F1 scores for Llama (0.7742) and Mistral (0.7747) differ marginally, and their confidence intervals substantially overlap ([0.7726, 0.7758] versus [0.7732, 0.7762]), indicating that these two models are statistically indistinguishable on semantic similarity. Gemma's BERTScore interval [0.7688, 0.7718] shows a slight but likely meaningful separation from the other two. In

summary: Llama is the clear leader on surface-form reproduction, Llama and Mistral are effectively tied on semantic fidelity, and Gemma trails on both dimensions but by a practically small margin on BERTScore (≈0.004 points). These confidence intervals help characterize the stability of the reported test-set estimates under the short-run 60-step adaptation setting, but they should not be interpreted as evidence that the models are fully converged or production-optimized.

Table 6. 95% Bootstrap Confidence Intervals on Test-Set Metrics

Metric	Gemma-3-1B	Llama-3.2-3B	Mistral-7B-4bit
BERTScore F1	0.7703 [0.7688, 0.7718]	0.7742 [0.7726, 0.7758]	0.7747 [0.7732, 0.7762]
ROUGE-1	0.1154 [0.1126, 0.1182]	0.1470 [0.1436, 0.1504]	0.1156 [0.1128, 0.1184]
BLEU	0.00244 [0.00208, 0.00280]	0.00914 [0.00832, 0.00996]	0.00237 [0.00202, 0.00272]
METEOR	0.0206 [0.0198, 0.0214]	0.0379 [0.0368, 0.0390]	0.0205 [0.0197, 0.0213]

4.6 Comparative Scientific Insights and Practical Implications

The main distinction among the BuildSafe variants is linguistic behavior rather than generalization stability. All three models show small train-test gaps, suggesting that the short LoRA adaptation mainly aligned pretrained models with the BuildSafe output schema rather than producing obvious memorization. The practical choice is therefore not simply "which model is best," but which trade-off is most appropriate. Llama-3.2-3B is preferable when reviewers need standardized terminology and repeatable permit-review phrasing. Mistral-7B-Instruct is stronger when longer explanatory narratives are needed for complex or escalated cases. Gemma-3-1B remains useful as a low-compute screening model because its semantic performance is only modestly lower than that of the larger models. These differences support a tiered decision-support design rather than a single-model ranking.

4.7 Interdisciplinary Positioning within Urban Informatics and Computational Built-Environment Research

Beyond model-level performance, BuildSafe contributes to the broader literature on urban informatics and computational built-environment intelligence by treating construction safety as a city-scale governance problem rather than a site-isolated prediction task. Prior computational urban-design research has shown that simulation, optimization, and AI-assisted design workflows can support resilient urban habitats and adaptive environmental decision-making (Maksoud et al., 2024a). Similarly, recent generative AI work in architectural design demonstrates that AI systems are increasingly used not only for prediction, but also for conceptual reasoning, design exploration, and decision support across the built environment (Maksoud et al., 2024b). BuildSafe extends these directions into construction governance by linking administrative permitting data, safety enforcement narratives, and civic complaint records into a unified tri-task reasoning framework. Unlike broader smart-city or multimodal urban-governance systems that often emphasize sensor streams, imagery,

mobility data, or environmental monitoring, BuildSafe focuses on administrative and civic text already embedded in permit review, enforcement, and complaint workflows. Its novelty therefore lies in converting existing DOB-OSHA-311 records into a tri-task generative reasoning layer for hazards, permit/code concerns, and neighborhood impacts rather than adding another sensor-centered monitoring pipeline (Jiao, 2024; Yang et al., 2017; Maksoud et al., 2024a). This positioning connects AI-enabled construction safety with smart-city analytics, urban risk prediction, and public-sector decision support, while preserving the need for human oversight, regulatory grounding, and operational accountability.

4.8 Baseline and Ablation Results

The zero-shot baselines consistently performed worse than the fine-tuned BuildSafe variants on BERTScore and ROUGE, indicating weaker alignment with the three-section reference narratives. Qualitatively, zero-shot outputs often appeared plausible but were less reliable in maintaining the required hazards, permit/code, and community-impact structure. They also showed higher hallucination rates, including incorrect code citations, implausible hazard categories, and spurious complaint types. Performance variance was larger across project types: zero-shot outputs were more acceptable for common jobs such as small interior renovations, but degraded more sharply for specialized or highly regulated work, including hoists, façade operations, and after-hours activities. These findings indicate that general instruction-following ability alone is insufficient for trustworthy high-stakes construction triage. Fine-tuning improved both semantic fidelity and structural coherence, particularly in the permit/code and community-impact sections. This supports the central BuildSafe premise that domain adaptation to linked DOB-OSHA-311 data improves reliability compared with generic LLM prompting. The improvement was not limited to surface wording; fine-tuned models more consistently preserved section boundaries, used NYC-relevant regulatory language, and connected project scope to plausible safety and neighborhood-impact concerns.

The data-source ablations showed that each linked data stream contributes differently. DOB-only models retained some basic hazard signal because permit descriptions encode job type, work scope, and location. However, they produced less nuanced permit and community-impact predictions when OSHA incident narratives and 311 complaint histories were removed. Adding OSHA data improved hazard-section alignment and reduced hallucinated hazard types, especially for fall and struck-by risks. Adding 311 data improved community-impact narratives by better aligning outputs with noise, dust, sidewalk obstruction, and neighborhood-disruption complaints. In some cases, 311 data also modestly improved hazard and permit sections, likely because complaint histories provide shared spatial and contextual signals about dense or disruption-prone construction environments. Task-structure ablations further clarified the value of the unified tri-task design. Single-task models achieved slightly higher scores in their specialized sections, as expected, but the tri-task model showed better internal coherence across sections. In particular, the tri-task outputs linked the same underlying construction activity to hazards, permit/code requirements, and community impacts more consistently. This advantage was most evident for rarer combinations of hazards and complaints, where shared context helped the model avoid treating safety and community disruption as unrelated. Thus, the tri-task formulation trades a small amount of section-specific optimization for stronger cross-section consistency, which is important for municipal review workflows.

Based on Table 7, the fine-tuned BuildSafe variants consistently outperformed their zero-shot counterparts, with the largest gains appearing in ROUGE-1 and section-structure consistency, indicating that LoRA adaptation improved alignment with the required three-section safety narrative

format. Table 7 reports zero-shot and fine-tuned results under the same held-out evaluation protocol, while Table 8 summarizes section-level patterns from the data-source ablations.

Table 7. Zero-Shot Versus Fine-Tuned BuildSafe Performance

Model	Setting	BERTScore F1	ROUGE-1	Hallucination trend	Section structure
Gemma-3-1B	Zero-shot	0.7350	0.0600	Higher	Often incomplete; sections sometimes merged or misordered
Gemma-3-1B	Fine-tuned	0.7703	0.1154	Lower	Consistent Hazards → Permits → Community Impacts
Llama-3.2-3B	Zero-shot	0.7450	0.0800	Higher	Inconsistent; missing or merged sections
Llama-3.2-3B	Fine-tuned	0.7742	0.1470	Lower	Stable three-section template
Mistral-7B 4-bit	Zero-shot	0.7400	0.0650	Higher	Occasionally deviates from template
Mistral-7B 4-bit	Fine-tuned	0.7747	0.1156	Lower	Consistent tri-section outputs

Table 8. Data-Source Ablation Results and Section-Level Interpretation

Configuration	Hazard Section	Permit/Code Section	Community-Impact Section	Main Interpretation
DOB only	Basic hazard signal retained; weaker incident specificity	Moderate; relies mainly on job type and permit description	Weakest; generic neighborhood-impact language	Permit descriptions alone are insufficient for full urban-risk reasoning
DOB + OSHA	Strongest or near-strongest hazard alignment	Moderate to strong	Limited; fewer complaint-specific details	OSHA narratives improve safety specificity and

				reduce hallucinated hazard types
DOB + 311	Moderate; some gains from spatial/contextual complaint patterns	Moderate	Stronger; better noise, dust, obstruction, and disruption language	311 histories improve community-impact reasoning
DOB + OSHA + 311	Most balanced across sections	Strong	Strong	Full linked corpus provides the best cross-domain coverage and tri-task balance

4.9 Practical Implications for Tiered Decision Support

The baseline and ablation findings should be interpreted as design evidence rather than deployment proof. Zero-shot prompting produced weaker semantic alignment, less stable section structure, and more hallucination-prone outputs, indicating that generic LLM prompting is insufficient for high-stakes construction-safety triage. Domain adaptation improves the logic for a tiered workflow, but this study does not test agency adoption, review-time reduction, inspection prioritization, injury reduction, complaint reduction, or governance impact. In a future pilot, Gemma-3-1B could be evaluated as a rapid screening model, Llama-3.2-3B as a standardized permit-review assistant, and Mistral-7B-Instruct as an escalation model for complex cases requiring richer explanation.

4.10 Conceptual Public-Sector Deployment Architecture

From a translational research perspective, BuildSafe can be conceptualized as a decision-support layer for municipal construction governance rather than as an autonomous permitting, inspection, or enforcement system. Figure 2 presents a conceptual deployment architecture in which BuildSafe connects three public-sector data streams: DOB permit and job-filing records, OSHA enforcement and injury records, and NYC 311 community complaint data. In the proposed workflow, these sources enter a secure data-ingestion layer where records are standardized, geocoded, temporally aligned, de-identified where appropriate, and screened for data-quality issues. The resulting linked record is passed to the BuildSafe AI engine, which generates three structured outputs: job-site hazard signals, likely permit or code-review concerns, and potential community-impact indicators. The BuildSafe engine should be deployed behind an agency-controlled application programming interface rather than as a public-facing chatbot. For routine filings, a compact model such as Gemma-3-1B could provide rapid first-pass screening. Cases with higher uncertainty, unusual work types, incomplete records, or elevated predicted risk could be escalated to Llama-3.2-3B or Mistral-7B-Instruct for a more detailed narrative review. A retrieval-augmented module should also be connected to current DOB code provisions, OSHA standards, agency bulletins, and inspection guidance so that generated outputs can reference up-to-date regulatory material rather than relying only on learned parameters. In this configuration, BuildSafe would provide structured advisory summaries rather than final regulatory determinations. In a reviewer-facing interface, these summaries would appear as editable advisory panels showing the generated hazard, permit/code, and community-impact sections alongside source-record provenance, uncertainty flags, and reviewer action options such as accept, revise, escalate, or

override. The architecture is intended as a conceptual implementation framework for future agency pilots, not as evidence that BuildSafe has already been integrated with live DOB, OSHA, or 311 operational systems.

Conceptually, the architecture is designed to support three agency-facing workflows in future pilot deployments. DOB reviewers could use BuildSafe risk summaries to prioritize permit applications for additional review, inspection scheduling, or targeted document checks. OSHA analysts could use aggregated hazard patterns to identify recurring work-type risks or potential violation themes, while recognizing that enforcement actions require independent investigation. NYC 311 and local agency staff could use community-impact summaries to anticipate noise, dust, sidewalk obstruction, after-hours work, or neighborhood-disruption concerns in dense urban areas. In all cases, the system should maintain a human-in-the-loop review structure, allowing authorized staff to accept, reject, edit, or override BuildSafe outputs, with each action logged for auditability. Because BuildSafe would operate in a high-stakes public-sector environment, deployment would require explicit safeguards for privacy, cybersecurity, and accountability. Data minimization should be applied so that only fields necessary for safety, permit, and community-impact reasoning are processed. Personally identifiable or sensitive information from complaint records, injury reports, and contractor records should be masked, aggregated, or access-controlled before model inference. Cybersecurity controls should include encrypted data transfer, role-based access control, secure containerized deployment, model version tracking, prompt/output logging, vulnerability testing, and incident response procedures. These controls are consistent with public-sector AI and cybersecurity guidance emphasizing AI risk governance, risk mapping, risk measurement, risk management, and cybersecurity governance, identification, protection, detection, response, and recovery (NIST, 2023; NIST, 2024). They also respond to LLM-specific risks such as prompt injection, insecure output handling, training-data poisoning, sensitive information disclosure, model supply-chain vulnerabilities, and excessive agency (OWASP, 2025). Legal accountability should remain with authorized public officials, not the AI system. BuildSafe outputs should therefore be accompanied by uncertainty flags, provenance information showing which source records contributed to the summary, version identifiers for the model and retrieval corpus, and clear documentation of model limitations. Reviewer overrides should be preserved in an audit log so that future evaluations can examine whether the system improves triage efficiency without producing disparate impacts across boroughs, neighborhoods, contractor types, or project categories. This approach is consistent with New York City's broader AI governance direction, which emphasizes responsible agency use, governance structures, public engagement, and guidance for emerging AI tools (NYC OTI, 2023). Under this architecture, BuildSafe becomes a transparent and reviewable decision-support system for construction governance rather than an opaque automated decision-maker.

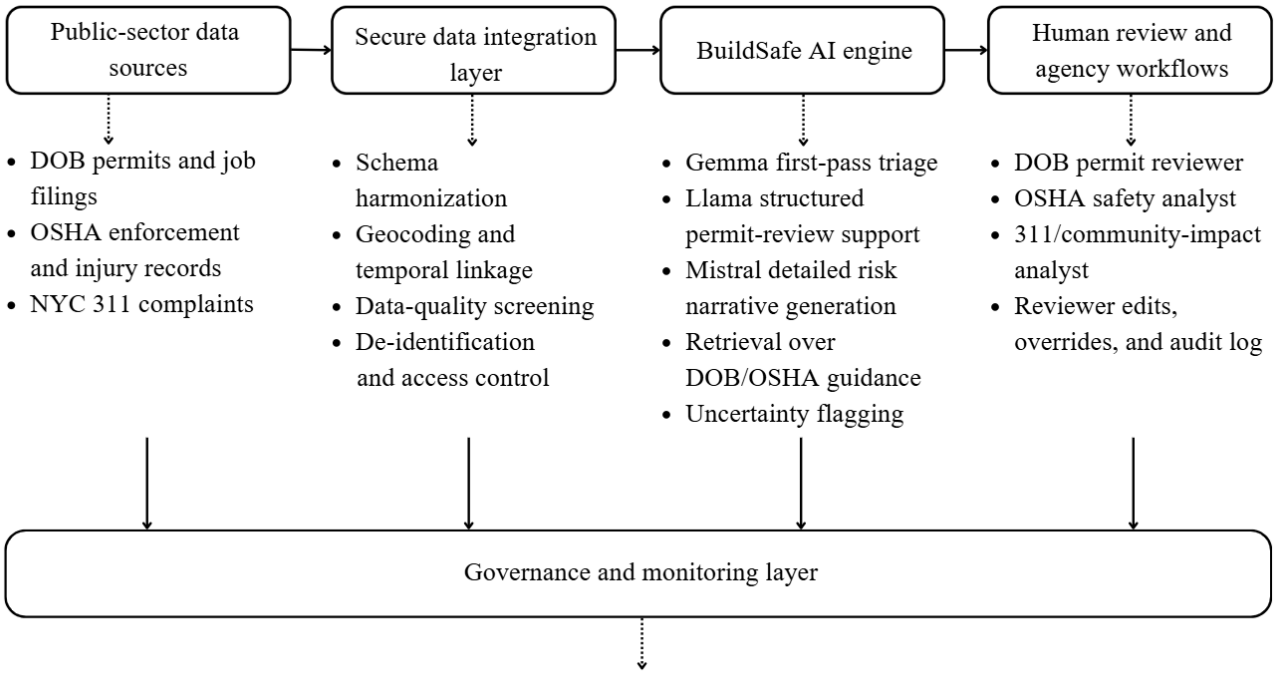


Figure 2. Conceptual public-sector deployment architecture for BuildSafe

The framework shows how DOB permit and job-filing records, OSHA enforcement and injury records, and NYC 311 complaint data are integrated into a secure data layer for schema harmonization, geocoding, temporal linkage, de-identification, and data-quality screening before being processed by the BuildSafe AI engine. The generated hazard, permit/code, and community-impact summaries would support human agency reviewers rather than automate permitting, inspection, or enforcement decisions. The governance layer emphasizes privacy protection, cybersecurity controls, fairness and bias auditing, model versioning, legal accountability, audit logging, and post-deployment performance monitoring. This architecture is intended for future agency pilots and does not imply that BuildSafe has already been integrated with live DOB, OSHA, or 311 operational systems.

4.11 Operational Fairness, Bias Auditing, and Regulatory Risk Management

A deployment-oriented BuildSafe system would require measurable fairness and bias-auditing procedures in addition to general statements about trustworthy AI. Bias may enter the pipeline through several mechanisms: uneven 311 complaint reporting across neighborhoods, historically unequal inspection or enforcement intensity, missing or inconsistent contractor records, and differences in how project descriptions are written across boroughs, job types, and applicant categories. If these patterns are learned uncritically, the system could over-flag projects in highly complaint-active neighborhoods or under-detect risks in areas where hazards are less frequently reported. Therefore, BuildSafe should not be evaluated only by aggregate BERTScore, ROUGE, or hallucination rates; it should also be audited for differential performance across geography, project type, work type, complaint category, and contractor history. In a pilot deployment, these risks should be audited through subgroup-level error analysis across boroughs, neighborhoods, project types, contractor categories where legally available, and complaint-density strata, using measurable indicators such as false-positive escalation rates, false-negative missed-risk rates, hallucination frequency, reviewer override rates, calibration error, and consistency of recommended review actions.

A practical fairness audit should report disaggregated metrics by borough, community district, job type, work type, and complaint-density category. For each subgroup, the audit should measure section-completion rate, factual-accuracy score, hallucination rate, false-positive risk flags, false-negative missed-hazard cases, reviewer override rate, and average reviewer confidence. These indicators would allow agency users to identify whether BuildSafe is systematically more likely to produce incomplete, exaggerated, or unsupported safety narratives for particular neighborhoods or project categories. In addition, calibration checks should compare model risk flags with later observed outcomes, such as inspections, violations, stop-work orders, injury records, or 311 complaint clusters, where such follow-up data are available. This would shift fairness evaluation from abstract ethical language toward measurable operational monitoring.

Regulatory risk management should also be built into the model lifecycle. Before deployment, agencies should document the intended use, prohibited use, training-data sources, known limitations, expected failure modes, and required human-review procedures. During deployment, outputs should be logged with model version, retrieval-corpus version, prompt context, uncertainty flags, reviewer actions, and override explanations. After deployment, periodic audits should examine whether system recommendations produce disparate impacts across neighborhoods, contractors, or project categories. These governance procedures are consistent with responsible AI lifecycle principles in construction and public-sector AI risk-management frameworks, which emphasize transparency, human oversight, accountability, and continuous monitoring (Agapiou, 2024; National Institute of Standards and Technology, 2023; New York City Office of Technology and Innovation, 2023; Yang et al., 2024). Importantly, BuildSafe should be restricted to advisory triage and reviewer support. It should not be used to automatically deny permits, trigger enforcement actions, assign penalties, or make final safety determinations without authorized human review. Affected applicants and agency staff should have access to explanations, source-record provenance, and mechanisms for contesting or correcting erroneous outputs. Under this governance model, fairness is not treated as a general aspiration but as a measurable set of auditing, documentation, monitoring, and accountability practices required for high-stakes construction-governance AI.

4.12 Evidence Boundaries for Operational Claims

The operational claims in this study should be interpreted as feasibility-oriented rather than as evidence of realized agency impact. The experiments show that open-weight LLMs can be adapted to generate structured hazard, permit/code, and community-impact narratives from linked DOB-OSHA-311 records, and that these outputs can be evaluated using automated metrics, bootstrap intervals, ablation comparisons, and a two-expert consensus validation subset. These findings support BuildSafe as a technical benchmark and decision-support prototype. They do not yet show that BuildSafe reduces inspection time, improves permit-review accuracy, lowers injury rates, reduces complaint volume, or improves regulatory outcomes in live agency settings. Such claims would require controlled deployment studies with agency reviewers, prospective outcome tracking, and comparison against existing review workflows. References to municipal deployment, governance support, and operational usefulness should therefore be understood as proposed use cases and design implications rather than demonstrated real-world impacts.

4.13 Sample Output Comparison

To illustrate how the quantitative differences translate into output behavior, the same Manhattan A3/EQ permit record was passed through all three fine-tuned BuildSafe models. The example is used as a qualitative illustration rather than as evidence of general performance. The input record contained borough = Manhattan, job type = A3, work type = EQ, permit status = issued, filing date = 05/10/2022,

issuance date = 05/10/2022, expiration date = 05/10/2023, job start date = 05/10/2022, and permittee = Frankie Colletta / Force Installations, LLC. In the NYC DOB context, the EQ work type denotes equipment installation, not earthquake-related work. The comparison therefore tests whether each model can preserve basic permit semantics while generating structured safety, permit/code, and community-impact reasoning.

Gemma-3-1B produced a structured three-section response covering Overall Risk (Moderate), Required Permits, and Potential Community Impact. It correctly identified that the same-day filing-to-issuance timeline suggests an expedited process and noted that Manhattan's density implies moderate noise and traffic disruption. However, the output was truncated mid-sentence at the community-impact section, and the model did not generate a dedicated Safety Issues section, reflecting the 1B model's more constrained generation capacity.

Llama-3.2-3B characterized the project as "relatively routine" and correctly interpreted the A3 job type as alterations or repairs. It offered a plausible (though imprecise) interpretation of the EQ work type as "earthquake-related work" and noted the swift permit process. The output was more concise than the other models, essentially a single analytical paragraph rather than a structured multi-section response, trading completeness for brevity.

Mistral-7B-Instruct (4-bit) produced the most comprehensive annotation, organized into four numbered sections: (1) Overall Risk with a reasoned "Moderate" assessment, (2) Required Permits including the suggestion of additional electrical permits, (3) Potential Community Impact broken into noise, traffic, and visual-impact sub-categories, and (4) Possible Safety Issues covering equipment hazards, working conditions, and environmental impact. The model also used the permittee's business name ('FORCE INSTALLATIONS, LLC') as a contextual cue, suggesting stronger contextual integration, although such inferences would still require human verification. The qualitative comparison reinforces the quantitative findings: Mistral excels at structured, comprehensive narrative generation; Llama produces accurate but condensed analysis; and Gemma provides adequate initial assessments that benefit from downstream refinement by larger models.

Comparative synthesis: The qualitative comparison is consistent with the metric results. Gemma produced a usable but shorter response, supporting its role as a lightweight screening model. Llama provided the most standardized and concise wording, aligning with its stronger lexical-overlap scores. Mistral produced the richest explanatory narrative, consistent with its stronger semantic profile. However, this additional detail also reinforces the need for human review and retrieval-grounded regulatory verification before operational use.

4.14 Preliminary Human Validation Results (n=150)

Table 9 presents human validation trends based on the stratified 150-record subset evaluated against two-expert consensus reference narratives. For this subset, A.A. and H.S.N. independently authored three-section human reference narratives while blinded to both Gemini teacher labels and fine-tuned model outputs, then resolved disagreements through discussion. The resulting consensus narratives were used as the human reference benchmark for assessing factual accuracy, completeness, overall usefulness, and hallucination behavior. Mistral-7B produced the most complete and structured outputs, with the highest perceived usefulness and lowest hallucination tendency. Llama-3.2-3B demonstrated strong factual accuracy but tended to give concise responses, which moderately limited completeness. Gemma-3-1B, while computationally efficient, produced shorter outputs that were occasionally truncated, reducing completeness scores relative to the larger models. Across all three models, these expert-assessed trends were broadly consistent with the automated BERTScore rankings, suggesting

that the metric ordering provides a useful preliminary proxy for domain output quality. However, because both experts are members of the author team and the subset contains only 150 records, these findings should be interpreted as internal validation rather than independent external validation.

Table 9. Preliminary Human Validation Results (n = 150)

Model	Factual Accuracy	Completeness	Overall Usefulness	Hallucination
Gemma-3-1B	Moderate	Moderate	Moderate	Moderate to Low
Llama-3.2-3B	High	High	High	Low
Mistral-7B-Instruct (4-bit)	High	High	Highest	Lowest

Future external validation should also report pre-consensus inter-rater agreement, such as weighted kappa, Krippendorff's alpha, or intraclass correlation, before adjudication to quantify expert agreement more transparently.

5 Limitations

First, the main fine-tuning supervision and full-scale automated evaluation references derive from Gemini 1.5 Flash rather than from a large, independently annotated expert dataset. As a result, BERTScore, ROUGE, BLEU, and METEOR on the full held-out test set primarily measure consistency with a machine-generated teacher reference, not absolute factual or regulatory correctness. To reduce this concern, the study includes a stratified 150-record human-reference subset for which two experts independently authored three-section safety narratives while blinded to both Gemini labels and fine-tuned model outputs. However, this subset remains limited in scale, and both experts are members of the author team. The human validation should therefore be interpreted as a stronger internal validation step rather than as fully independent external validation.

Second, the 60-step LoRA adaptation schedule should be interpreted as a reproducible short-run proof-of-concept rather than as a fully optimized convergence study. The stable train, validation, and test metrics suggest that the models learned the BuildSafe output structure without obvious overfitting. However, these results do not prove that longer training, different random seeds, alternative learning rates, or checkpoint-level early stopping would not improve performance.

Third, the baseline experiments strengthen the study but do not exhaust all possible comparisons. The revised analysis includes zero-shot baselines, data-source ablations, and task-structure ablations, which are appropriate for evaluating the tri-task generative format. However, traditional machine learning baselines were not directly comparable because BuildSafe generates structured hazard, permit/code, and community-impact narratives rather than single-label predictions. Future work should derive matched classification or ranking tasks, such as hazard category prediction, permit-review escalation, stop-work-risk flags, or high-complaint-risk indicators, to enable fair comparison with logistic regression, support vector machines, random forests, gradient boosting, and retrieval-based non-generative systems.

Fourth, BuildSafe has not yet been evaluated in live agency workflows and remains specific to NYC's data infrastructure, regulatory vocabulary, and urban conditions. The results support technical feasibility as a benchmark and decision-support prototype. However, they do not demonstrate reductions in inspection time, permit-review error, injury rates, complaint volume, or regulatory burden. Before operational deployment or transfer to other jurisdictions, BuildSafe would require prospective agency pilots, controlled reviewer studies, retrieval over current regulatory sources, privacy and cybersecurity safeguards, fairness and bias audits, audit logging, legal accountability procedures, local schema mapping, and independent expert validation.

6 Future Work

Integrate additional administrative sources (e.g., enforcement follow-ups, contractor histories, verified "clean" projects) and strengthen record linkage, with explicit reporting of linkage quality and data bias by geography and contractor type. Combine parameter-efficient fine-tuning with retrieval-augmented generation over up-to-date codes and regulations, and explore cascaded inference pipelines that route routine filings to compact models and escalate edge cases to larger ones. Evaluate BuildSafe in controlled agency pilots using impact-oriented KPIs (injury rates, review times, complaint trends), run disaggregated fairness audits, and formalize human-in-the-loop review, override, and logging workflows as prerequisites for any production deployment. These pilots should report subgroup-level performance across boroughs, community districts, job types, work types, complaint-density categories, and contractor-history profiles, including hallucination rates, false-positive and false-negative risk flags, reviewer override rates, and calibration against later inspection, violation, injury, or complaint outcomes. Subject to agency data-sharing agreements and cybersecurity review, develop a secure, containerized architecture that integrates with DOB and 311 systems, and in the longer term, explore multimodal extensions that fuse text with imagery, sensor streams, and BIM/CAD data for real-time early-warning use cases.

7 Conclusion

BuildSafe demonstrates that three open-weight LLMs Gemma-3-1B, Llama-3.2-3B, and Mistral-7B-Instruct can be adapted to jointly model construction hazards, permit/code requirements, and community impacts from linked administrative text in an approximately 90,000-record DOB-OSHA-311 corpus. On a 2,833-record held-out test set, the fine-tuned models achieved BERTScore F1 values of 0.7703, 0.7742, and 0.7747, respectively, compared with zero-shot baselines of 0.7350, 0.7450, and 0.7400, corresponding to adaptation gains of approximately 0.029-0.035 F1 points. The fine-tuned models also showed lower hallucination tendencies and more consistent three-section output structure than zero-shot prompting, with evaluation supplemented by a 150-record, two-expert consensus validation subset. The results show distinct model roles rather than a single universally superior model. Llama-3.2-3B produced the strongest lexical consistency, with ROUGE-1 = 0.1470, BLEU = 0.0091, and METEOR = 0.0379; Mistral-7B-Instruct produced the strongest semantic profile and most coherent long-form explanations; and Gemma-3-1B remained semantically competitive as a low-compute first-pass screening model. Scientifically, BuildSafe contributes a linked urban-governance formulation of construction risk, showing that administrative permit records, OSHA safety narratives, and 311 complaint histories can be organized into a tri-task reasoning corpus for hazards, permit/code concerns, and community impacts. The study remains bounded by its reliance on Gemini-generated silver-standard references, the limited scale of the two-expert validation subset, the short-run 60-step adaptation setting, and the NYC-specific regulatory context. BuildSafe should therefore be viewed as a benchmark and decision-support research framework rather than an autonomous permitting or enforcement system. The most actionable next step is a controlled DOB reviewer pilot that uses the

2,833-record benchmark as a technical baseline while measuring operational outcomes not tested here, including triage time-to-flag, reviewer override rate, false-positive escalation rate, false-negative missed-risk rate, and agreement with existing permit-review and inspection workflows.

8 Conflict of Interest

The 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.

9 Author Contributions

S.A.: Conceptualization, Methodology, Data curation, Software, Formal analysis, Visualization, Writing - original draft, Writing - review & editing. A.A.: Conceptualization, Writing - review & editing, Supervision (construction safety domain), Expert validation. H.S.: Software, Validation, Formal analysis, Writing - review & editing. H.S.N.: Supervision, Project administration, Expert validation, Writing - review & editing.

10 Funding

The authors declare that financial support was not received for the research, authorship, and/or publication of this article.

11 References

Agapiou, A. (2024). A systematic review of the socio-legal dimensions of responsible AI and its role in improving health and safety in construction. Buildings, 14(5), 1469. https://doi.org/10.3390/buildings14051469.

American Petroleum Institute. (2019). Survey of occupational injuries, illnesses, and fatalities in the petroleum industry: Guidelines and definitions. https://www.api.org/-/media/files/oil-and-natural-gas/pipeline/2019-awards/oii-guidelines-and-definitions.pdf.

Awolusi, I., Marks, E., Hainen, A., & Alzarrad, A. (2022). Incident analysis and prediction of safety performance on construction sites. Civileng, 3(3), 669-686. https://doi.org/10.3390/civileng3030039.

Behzadan, A. (2024). Formalizing trust in artificial intelligence for built environment decision-making. Human Factors in Design, Engineering, and Computing, 159(159). https://doi.org/10.54941/ahfe1005565.

Ceccato, V. (2013). Integrating geographical information into urban safety research and planning. Proceedings of the Institution of Civil Engineers-Urban Design and Planning, 166(1), 15-23. https://doi.org/10.1680/udap.11.00038.

Chandu, K. P., Raja, K. H., & Kumar, N. N. (2024). From reactive to proactive: The role of wearable technology, AI, and digital training in construction safety management. Libr. Prog. Int, 44, 22858-22864. https://bpasjournals.com/library-science/index.php/journal/article/view/2916.

Choi, H. K., Du, X., & Li, Y. (2024). Safety-aware fine-tuning of large language models. arXiv.Org. https://doi.org/10.48550/arxiv.2410.10014.

City of New York. (2026, February 27). 311 service requests from 2020 to present [Data set]. NYC Open Data. https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2020-to-Present/erm2-nwe9/about_data.

Data Science Dojo. (2025). Mistral 7B: A Revolutionary Breakthrough in LLMs. Retrieved April 17, 2025, from https://datasciencedojo.com/blog/mistral-7b-emergence-in-llm/.

Dufitimana, E., Gahungu, P., Uwayezu, E., Mugisha, E., & Bizimana, J. P. (2025). Integrating machine learning and geospatial data for mapping socioeconomic vulnerability to urban natural hazard. ISPRS International Journal of Geo-Information, 14(4), 161. https://doi.org/10.3390/ijgi14040161.

Farabet, C. (2025, March 13). Introducing Gemma 3: The most capable model you can run on a single GPU or TPU. Google. https://blog.google/technology/developers/gemma-3/.

Gao, B., Ma, Z., Gu, J., Han, X., Xiang, P., & Lv, X. (2023). Fusing multi-source quality statistical data for construction risk assessment and warning based on deep learning. Knowledge-Based Systems, 284, 111223. https://doi.org/10.1016/j.knosys.2023.111223.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., ... & Vasic, P. (2024). The llama 3 herd of models. https://doi.org/10.48550/arXiv.2407.21783.

Hossain, M. I., Hosen, M. M., Sunny, M. A. U., & Tarapder, S. A. (2025). Implementing advanced technologies for enhanced construction site safety. American Journal of Advanced Technology and Engineering Solutions, 1(02), 01-31. https://doi.org/10.63125/3v8rpr04.

HSI. (2025, March 18). Predict and prevent construction risks with AI-powered capabilities. https://hsi.com/blog/ai-construction-safety-risk-management.

Hsu, C. Y., Tsai, Y. L., Lin, C. H., Chen, P. Y., Yu, C. M., & Huang, C. Y. (2024). Safe lora: The silver lining of reducing safety risks when finetuning large language models. Advances in Neural Information Processing Systems, 37, 65072-65094. https://doi.org/10.48550/arxiv.2405.16833.

Huang, K., Joshi, A. V., Dun, S., & Hamilton, N. (2024). AI regulations. In A. V. Joshi, S. Dun, & N. Hamilton (Eds.), AI and the law (pp. 47–71). Springer. https://doi.org/10.1007/978-3-031-54252-7_3

Hugging Face. (2025a). Mistral-7B-Instruct-v0.3-bnb-4bit. Retrieved April 17, 2025, from https://huggingface.co/unsloth/mistral-7b-instruct-v0.3-bnb-4bit.

Hugging Face. (2025b). Unleashing the Power of Unsloth and QLora: Redefining Language Model Fine-Tuning. Retrieved April 17, 2025, from https://huggingface.co/blog/Andyrasika/finetune-unsloth-qlora

Hussain, R., Lee, D., Abbas, M. S., Zaidi, S. F. A., Pedro, A., and Park, C. (2026). Vision-language model-based intelligent assistant for onsite construction safety inspection. *Automation in Construction* 182, 106728. https://doi.org/10.1016/j.autcon.2025.106728.

Jiao, D. (2024). Safety Prediction Method of Urban Construction FP Based on Multi-Order Spatio-Temporal Features. IEEE Access, 12, 77009-77018. https://doi.org/10.1109/access.2024.3407160

Junjia, Y., Alias, A. H., Haron, N. A., & Bakar, N. A. (2024). Intelligent construction risk management through transfer learning: Trends, challenges, and future strategies. Artificial Intelligence Evolution. https://doi.org/10.37256/aie.6120255255.

Kamil, M. Z., Khan, F., Amyotte, P., & Ahmed, S. (2024). Multi-source heterogeneous data integration for incident likelihood analysis. Computers & Chemical Engineering, 185, 108677. https://doi.org/10.1016/j.compchemeng.2024.108677.

Kumi, L., Jeong, J., & Jeong, J. (2024). Systematic review of quantitative risk quantification methods in construction accidents. Buildings, 14(10), 3306. https://doi.org/10.3390/buildings14103306.

Lu, W., Luu, R. K., & Buehler, M. J. (2025). Fine-tuning large language models for domain adaptation: Exploration of training strategies, scaling, model merging and synergistic capabilities. npj Computational Materials, 11(1), 84. https://doi.org/10.48550/arxiv.2409.03444.

Maksoud, A., Alawneh, S.I.A.-R., Hussien, A., Abdeen, A., and Abdalla, S.B. (2024a). Computational design for multi-optimized geometry of sustainable flood-resilient urban design habitats in Indonesia. *Sustainability* 16(7), 2750. https://doi.org/10.3390/su16072750

Maksoud, A., Elshabshiri, A., Saeed Hilal Humaid Alzaabi, A., and Hussien, A. (2024b). Integrating an image-generative tool on creative design brainstorming process of a Safavid mosque architecture conceptual form. *Buildings* 14(3), 843. https://doi.org/10.3390/buildings14030843

Miller, D. (2024, July 3). An introduction to Mistral-7B. Future Skills Academy. Retrieved April 17, 2025, from https://futureskillsacademy.com/blog/mistral-7b/

Mistral AI. (2024). Mistral 7B. Retrieved April 17, 2025, from https://mistral.ai/news/announcing-mistral-7b.

Mohamed, M. A. H., Al-Mhdawi, M. K. S., Ojiako, U., Dacre, N., Qazi, A., & Rahimian, F. (2025). Generative AI in construction risk management: a bibliometric analysis of the associated benefits and risks. Urbanization, Sustainability and Society, 2(1), 196-228. https://doi.org/10.1108/USS-11-2024-0069.

Mostofi, F., & Toğan, V. (2023). Construction safety predictions with multi-head attention graph and sparse accident networks. Automation in Construction, 148, 105102. https://doi.org/10.1016/j.autcon.2023.105102.

Mulligan, K., Cuevas, C., Grimsley, E., Chauhan, P., & Bond, E. (2019). Justice data brief: Understanding New York City's 311 data. https://datacollaborativeforjustice.org/wp-content/uploads/2019/03/DCJ-Justice-Data-Brief-NYC-311-Calls.pdf.

National Institute of Standards and Technology (NIST). (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. Available at: https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf (Accessed May 13, 2026).

National Institute of Standards and Technology (NIST). (2024). The NIST Cybersecurity Framework (CSF) 2.0. NIST Cybersecurity White Paper 29. Available at: https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.29.pdf (Accessed May 13, 2026).

New York City Department of City Planning. (n.d.). Housing Database: Project-level and unit-change summary files [Data sets]. NYC Open Data & BYTES of the BIG APPLE. https://data.cityofnewyork.us/Housing-Development/Housing-Database/6umk-irkx.

New York City Office of Technology and Innovation (NYC OTI). (2023). The New York City Artificial Intelligence Action Plan. Available at: https://www.nyc.gov/assets/oti/downloads/pdf/reports/artificial-intelligence-action-plan.pdf (Accessed May 13, 2026).

Nycdb. (n.d.). GitHub - nycdb/nycdb: Database of NYC Housing Data. GitHub. https://github.com/nycdb/nycdb.

Occupational Safety and Health Administration. (2023). Construction injury statistics. https://www.osha.gov/data/commonstats.

Occupational Safety and Health Administration. (2025). Injury Tracking Application (ITA). U.S. Department of Labor. https://www.osha.gov/injuryreporting.

Open Knowledge Foundation. (n.d.). US DOL enforcement data. DataPortals.org. https://enforcedata.dol.gov.

OWASP. (2025). OWASP Top 10 for Large Language Model Applications. Available at: https://owasp.org/www-project-top-10-for-large-language-model-applications/ (Accessed May 13, 2026).

Parekh, R., & Mitchell, O. (2024). Progress and obstacles in the use of artificial intelligence in civil engineering: An in-depth review. *International Journal of Science and Research Archive*, *13*(1), 1059–1080. https://doi.org/10.30574/ijsra.2024.13.1.1777.

Pilskog Orvik, C. (2024, August). Improving Safety through Leveraging Machine Learning and Safety-Related Data in the Construction Industry. In IOP Conference Series: Earth and Environmental Science (Vol. 1389, No. 1, p. 012012). IOP Publishing. https://doi.org/10.1088/1755-1315/1389/1/012012.

Piri, S., & Panthi, K. (2024). Unravelling Safety Concerns in Construction: A Comprehensive Data Analysis. Proceedings of 60th Annual Associated Schools, 5, 912-920. https://doi.org/10.29007/snfn.

Raliile, M. T., & Haupt, T. C. (2020). Machine learning applications for monitoring construction health and safety legislation and compliance. Proc. Int. Struct. Eng. Constr, 7, 231-6. https://doi.org/10.14455/ISEC.2020.7(2).CON-23.

Rasheed, O. A., Baalah, M. P. G., & Segun, O. J. (2024). AI-driven risk mitigation: Transforming project management in construction and infrastructure development. WORLD, 13(2), 611-623. https://doi.org/10.30574/wjaets.2024.13.2.0628.

Sagar, S. (2025). Low-Resource Fine-Tuning of LLMs for Domain-Specific Tasks. Universal Research Reports. https://doi.org/10.36676/urr.v12.i4.1621.

Sargiotis, D. (2024). Transforming civil engineering with AI and machine learning: innovations, applications, and future directions. Applications, and Future Directions (November 15, 2024). https://doi.org/10.2139/ssrn.5021723.

Savaş, S. (2025). Artificial intelligence in construction project management: Trends, challenges and future directions. Journal of Design for Resilience in Architecture and Planning, 6(2), 221-238. https://doi.org/10.47818/DRArch.2025.v6i2165.

Tan, Z., Li, D., Wang, S., Beigi, A., Jiang, B., Bhattacharjee, A., Karami, M., Li, J., Cheng, L., & Liu, H. (2024). Large language models for data annotation: A survey. *arXiv*. https://doi.org/10.48550/arXiv.2402.13446.

Tang, B., & Luo, H. (2025). Computer vision and large language model-based safety management for construction project sites. In Proceedings of the 14th Creative Construction Conference (CCC 2025), Zadar, Croatia (pp. 249–257). IAARC. https://doi.org/10.22260/CCC2025/0031.

Tang, K. H. D. (2024). Artificial intelligence in occupational health and safety risk management of construction, mining, and oil and gas sectors: advances and prospects. Journal of Engineering Research and Reports, 26(6), 241-253. https://doi.org/10.9734/jerr/2024/v26i61177.

Taroun, A., Yang, J.-B., & Lowe, D. (2011). Construction risk modelling and assessment: insights from a literature review. The Built & Human Environment Review, 4(Special Issue 1), 87–97.

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., Tafti, P., Hussenot, L., Sessa, P. G., Chowdhery, A., Roberts, A., Barua, A., Botev, A., Castro-Ros, A., Slone, A., . . . Kenealy, K. (2024). GemMa: Open models based on Gemini research and technology. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2403.08295.

Tixier, A. J.-P., & Hallowell, M. R. (2023). Safer together: Machine learning models trained on shared accident datasets predict construction injuries better than company-specific models. arXiv.Org. https://doi.org/10.48550/arXiv.2301.03567.

Tussey, D., & Yan, J. (2025). Principles for Open Data Curation: A Case Study with the New York City 311 Service Request Data. *ArXiv.org*. https://doi.org/10.48550/arxiv.2502.08649.

Ubiai. (2024). Fine-Tuning Mistral 7B for Named Entity Recognition. Retrieved April 17, 2025, from https://ubiai.tools/fine-tuning-mistral-for-ner/.

Unsloth. (2026). Fine-tuning LLMs guide. Unsloth Documentation. https://unsloth.ai/docs/get-started/fine-tuning-llms-guide.

Unslothai. (2025a). FAQ + Is Fine-tuning Right For Me? Retrieved April 17, 2025, from https://docs.unsloth.ai/get-started/beginner-start-here/faq-+-is-fine-tuning-right-for-me.

Unslothai. (2025b). GitHub - unslothai/unsloth: Fine-tuning & Reinforcement Learning for LLMs. Train OpenAI gpt-oss, DeepSeek, Qwen, Llama, Gemma, TTS 2x faster with 70% less VRAM. GitHub. https://github.com/unslothai/unsloth.

Usama, M., Ullah, U., Muhammad, Z., Islam, T., & Hashmi, S. S. (2024). AI-enabled risk assessment and safety management in construction. 9781032648323-8. https://doi.org/10.1201/9781032648323-8.

Wan, H., Zhang, J., Chen, Y., Xu, W., & Feng, F. (2024). Exploring Gen-AI applications in building research and industry: A review. arXiv.Org. https://doi.org/10.48550/arxiv.2410.01098.

Weng, B. (2024). Navigating the landscape of large language models: A comprehensive review and analysis of paradigms and fine-tuning strategies. arXiv.Org. https://doi.org/10.48550/arxiv.2404.09022.

Workers' Compensation Trust. (2021, December 2). Guide to OSHA's electronic illness and injury reporting requirements. https://www.wctrust.com/Customer-Content/www/CMS/files/Loss%20Control%20Uploads/Final_OSHA_RK_Electronic_Reporting_Kit_12_02_21.pdf.

Wörsdörfer, M. (2025). Ten reasons why—the case for more and better AI regulation. AI and Ethics, 6(1), 62. https://doi.org/10.1007/s43681-025-00865-8.

Yang, L., Allen, G., Zhang, Z., & Zhao, Y. (2024). Achieving on-site trustworthy AI implementation in the construction industry: A framework across the AI lifecycle. Buildings, 15(1), 21. https://doi.org/10.3390/buildings15010021.

Yang, Z., Gupta, K., Gupta, A., & Jain, R. K. (2017). A data integration framework for urban systems analysis based on geo-relationship learning. In Computing in Civil Engineering 2017 (pp. 467-474). https://doi.org/10.1061/9780784480823.056.

Yoo, B., Kim, J., Park, S., Ahn, C. R., & Oh, T. (2024). Harnessing generative pre-trained transformers for construction accident prediction with saliency visualization. Applied Sciences, 14(2), 664. https://doi.org/10.3390/app14020664.

Zerkin, A. J. (2006). Mainstreaming high performance building in New York City: A comprehensive roadmap for removing the barriers. Technology in Society, 28(1-2), 137-155. https://doi.org/10.1016/j.techsoc.2005.10.017.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. ICLR 2020. https://doi.org/10.48550/arXiv.1904.09675.

Zhao, J., Wang, Y., Abid, W., Angus, G., Garg, A., Kinnison, J., Sherstinsky, A., Molino, P., Addair, T., & Rishi, D. (2024). LoRA Land: 310 fine-tuned LLMs that rival GPT-4, a technical report. arXiv.Org. https://doi.org/10.48550/arxiv.2405.00732.

Zou, Z., & Ergan, S. (2019). Leveraging data driven approaches to quantify the impact of construction projects on urban quality of life. *arXiv*. https://doi.org/10.48550/arXiv.1901.09084.

12 Data Availability Statement

Some or all data, models, or code that support the findings of this study are available from the corresponding author upon reasonable request. The DOB permit-issuance records and 311 service-request data were obtained from the NYC Open Data portal (https://opendata.cityofnewyork.us/).

OSHA enforcement and inspection records were obtained from the U.S. Department of Labor's public data catalog (https://enforcedata.dol.gov/views/data_catalogs.php). Fine-tuning code and LoRA adapter weights are available from the corresponding author upon reasonable request.