Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.081652

ARTICLE

# Knowledge Graph-Driven Training Data Construction for Urban Flood-Traffic Scenario Generation Using Small Language Models

**Geunhwi Park[1], Juneyoung Park[2,*], Chunjoo Yoon[3] and Jaehong Park[3]**

[1]Department of Smart City Engineering, Hanyang University, Ansan-si, Republic of Korea
[2]Department of Transportation and Logistics Engineering, Hanyang University, Ansan-si, Republic of Korea
[3]Department of Highway & Transportation Research, Korea Institute of Civil Engineering and Building Technology, Goyang-si, Republic of Korea

*Corresponding Author: Juneyoung Park. Email: juneyoung@hanyang.ac.kr

Received: 06 March 2026; Accepted: 14 May 2026; Published: 15 June 2026

**ABSTRACT:** Urban flooding caused by extreme rainfall events disrupts transportation systems, yet generating realistic flood-traffic scenarios for disaster preparedness remains a labor-intensive manual process. This study proposes a Knowledge Graph (KG)-driven pipeline that automatically generates domain-specific training data for fine-tuning small language models (sLLMs) to synthesize urban flood-traffic scenarios. A domain KG comprising 58 entities and 285 relationships was constructed for Jinju City, South Korea, integrating empirical flood data from 112 local documents with quantitative rainfall-traffic impact values from 14 international studies. Nine domain constraint rules, including a novel spatial consistency rule, ensure the physical plausibility of generated scenarios. Through constrained weighted graph walks, 800 semi-structured English narrative scenarios were automatically generated in approximately 5 min, substantially reducing the labor required compared to manual creation. Three sLLMs spanning different architectures and parameter scales—Flan-T5-Large (770M), Qwen2.5-3B-Instruct (3B), and Qwen2.5-7B-Instruct (7B)—were fine-tuned using QLoRA on a single GPU with 16 GB VRAM. Evaluation on 78 test samples demonstrated consistent performance improvements with increasing model scale: Qwen2.5-7B achieved BLEU-4 of 0.5524, ROUGE-L of 0.6883, BERTScore F1 of 0.9662, and KG Fact Consistency of 1.0000, representing a 33.8% BLEU-4 improvement over Flan-T5-Large. Both Qwen models achieved KG Fact Consistency of 1.0000. The 3B model achieved 98.6% of the 7B model's BLEU-4 at 53% of the VRAM cost with identical factual consistency, representing the most cost-effective configuration. All models were trained for 10 epochs on the same GPU, demonstrating practical feasibility for municipal disaster response deployment.

**KEYWORDS:** Knowledge graph; training data generation; urban flood; traffic scenario; small language model; fine-tuning; text generation; QLoRA

## 1 Introduction

Urban flooding, intensified by climate change and rapid urbanization, poses significant threats to transportation infrastructure and public safety worldwide [1,2]. When heavy rainfall overwhelms urban drainage systems, the resulting road inundation disrupts traffic through reduced vehicle speeds, diminished capacity, and road closures, often triggering cascading congestion that hinders emergency response [3,4]. These impacts highlight the need for systematic approaches to anticipate flood-traffic interactions for effective disaster management.

Scenario-based planning is a critical tool for disaster preparedness, enabling agencies to simulate potential flood events and evaluate response strategies. However, generating high-quality flood-traffic scenarios remains predominantly manual and labor-intensive—domain experts must synthesize weather conditions, hydrological dynamics, geographic vulnerabilities, traffic operations, and emergency protocols into coherent narratives. This manual approach suffers from limited scalability, potential inconsistency, and poor adaptability to new cities.

While large language models (LLMs) offer promising avenues for text generation automation, deploying cloud-based LLMs for disaster scenarios encounters several limitations: connectivity may be unavailable during extreme events, sensitive infrastructure data must be transmitted externally, and lack of domain grounding can produce factually inaccurate outputs [5]. These constraints motivate small language models (sLLMs) that can be deployed locally on commodity hardware. However, fine-tuning sLLMs requires sufficient domain-specific training data, which is scarce and expensive to create manually for location-dependent flood-traffic scenarios. Knowledge Graphs (KGs) offer a principled solution by encoding domain entities, relationships, and constraints in a structured representation that can serve as a generative framework for producing diverse, factually consistent training scenarios through graph traversal.

This study proposes a KG-driven pipeline that automatically generates domain-specific training data for urban flood-traffic scenario synthesis, making two contributions:

Contribution 1: KG-Driven Training Data Generation Pipeline. This study constructs a domain KG comprising 58 entities and 285 relationships encoding causal, spatial, and quantitative relationships for Jinju City, South Korea, integrating empirical flood data from 112 local documents with traffic impact values from 14 international studies. Through constrained weighted graph walks with nine domain constraint rules—including a novel FLOOD_CLOSES spatial consistency rule—the pipeline generates 800 semi-structured English narrative scenarios in approximately 5 min, eliminating the bottleneck of manual scenario authoring.

Contribution 2: Comparative Evaluation of Three sLLMs across Architectures and Scales. The work compares three sLLMs spanning architecture and parameter scale: Flan-T5-Large (770M, encoder-decoder), Qwen2.5-3B-Instruct (3B, decoder-only), and Qwen2.5-7B-Instruct (7B, decoder-only). All models are fine-tuned using QLoRA on a single GPU (RTX 4060 Ti, 16 GB VRAM). Evaluation across BLEU-4, ROUGE-L, BERTScore F1, and a novel KG Fact Consistency Score reveals consistent performance improvements with increasing model scale, with the 7B model achieving KG Fact Consistency of 1.0000.

The remainder of this paper is organized as follows. Section 2 reviews related work on rainfall-traffic impacts, LLMs for text generation, and KG-based data augmentation. Section 3 details the methodology. Section 4 presents the experimental results. Section 5 discusses implications, limitations, and concludes the study.

## 2 Literature Review

### *2.1 Rainfall and Flood Impacts on Urban Traffic*

Extensive empirical research has quantified the effects of rainfall on traffic operations. Tsapakis et al. [6] demonstrated that the impact of rainfall on macroscopic urban travel times is intensity-dependent, with empirical and meta-review studies confirming speed reductions of 2%–17% and capacity reductions of 4%–32% under varying rainfall intensities on freeways and urban roads [7–9]. Pregnolato et al. [3] developed a depth-disruption function relating flood depth to vehicle speed, and Ni et al. [4] demonstrated that waterlogging amplifies traffic impacts by 1.3–2.0 times compared to rainfall alone. Despite this extensive body of quantitative knowledge, these findings have rarely been systematically integrated into automated scenario generation pipelines for disaster preparedness. Moreover, data scarcity remains a pervasive challenge across

traffic and disaster-response applications, particularly when conventional data collection becomes difficult or unsafe under extreme weather conditions.

### *2.2 Language Models for Domain-Specific Text Generation*

Large language models (LLMs) such as GPT-3 [10] and Llama 2 [11] have demonstrated strong text generation capabilities. However, as comprehensively reviewed by Moradi et al. [12], LLM-generated outputs in specialized domains frequently suffer from hallucination (producing fluent but factually inaccurate content [5]), and current alignment techniques remain insufficient for domain-critical applications such as disaster management. Small language models (sLLMs), typically under 10 billion parameters, offer a practical alternative through domain-specific fine-tuning, enabling local deployment without cloud dependency. Recent work on instruction-tuned models such as Flan-T5 [13] and Qwen2.5 [14] has shown that efficient fine-tuning techniques like QLoRA [15] can achieve competitive performance at substantially reduced computational costs. Shen et al. [16] further demonstrated that layer type is a more significant determinant than layer depth when optimizing fine-tuning in quantized LLMs, providing practical guidance for parameter-efficient strategies in resource-constrained settings. However, fine-tuning requires sufficient domain-specific training data, which remains a limiting factor for specialized applications.

### *2.3 Knowledge Graphs for Training Data Generation*

Knowledge Graphs have emerged as a promising approach for structured data augmentation and grounded text generation. Lewis et al. [17] proposed Retrieval-Augmented Generation (RAG), an approach that integrates parametric language model memory with non-parametric retrieved knowledge to ground language model outputs and reduce factual errors in knowledge-intensive tasks. Unlike RAG, which retrieves external knowledge at inference time and requires runtime access to a graph store, the present approach distills KG knowledge into model parameters via fine-tuning. This design enables offline operation during disaster events when network connectivity may be compromised. In domain-specific applications, Liu et al. [18] demonstrated that combining template-generated synthetic data with fine-tuned LLMs significantly improves domain-specific information extraction under data scarcity, and Chen et al. [19] applied knowledge graph-guided LLM reasoning to emergency decision-making, demonstrating its applicability in disaster-response contexts. While these studies demonstrate the potential of KG-driven text generation, none have applied this approach to urban flood-traffic scenario synthesis or systematically evaluated its effectiveness for fine-tuning sLLMs across different model scales.

## 3 Methodology

Fig. 1 illustrates the overall framework of the proposed KG-driven pipeline, which consists of six stages.

### *3.1 Data Collection and Preprocessing*

The study area is Jinju City, South Gyeongsang Province, South Korea, which experiences recurring urban flooding driven by monsoon rainfall and the Nam River (Namgang) system (Fig. 2). The city's topography (characterized by low-lying areas along the Namgang and Yeongcheongang rivers, densely developed commercial districts, and multiple bridge crossings) creates spatially concentrated flood vulnerabilities that frequently disrupt road networks.

Two types of flood-related textual data were collected: (1) 89 news articles from BIG KINDS, a Korean news archive platform, covering January 2001 to December 2024; and (2) 165 disaster alert messages from the Korean Disaster and Safety Data Sharing Platform, covering September 2023 to December 2024.

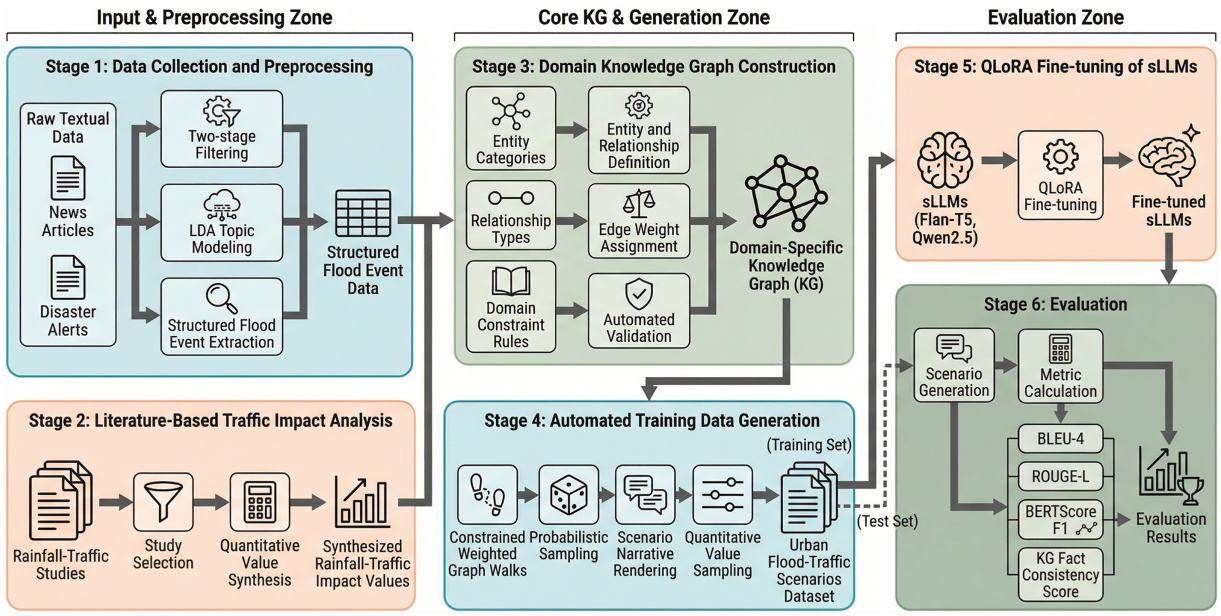


**Figure 1:** Overall framework of the proposed KG-driven pipeline for urban flood-traffic scenario generation.

**Figure 2:** Study area map of Jinju City.

A two-stage filtering process was applied in Stage 1, a hierarchical keyword filter selected documents containing strong flood-related terms (flood, inundation, heavy rain, torrential rain, overflow, water level, dam discharge), reducing the corpus from 254 to 138 documents. Strong flood keywords overrode exclusion keywords (collision, festival, election, COVID-19, earthquake), while documents matching only supporting keywords (road closure, detour, river, levee, drainage) were subject to exclusion filtering. In Stage 2, a

re-validation step removed 26 documents that passed through supporting keywords only but described non-flood events (sinkholes, road construction, festivals), yielding 112 final documents (82 news articles, 30 disaster alerts).

To verify thematic coherence, Latent Dirichlet Allocation (LDA) topic modeling was applied using Gensim ($k = 15$). Four dominant thematic clusters were identified (heavy rainfall alerts, flood-induced traffic congestion, infrastructure damage, and weather forecasting), with zero off-topic documents detected. Structured flood event extraction yielded 67 flood-affected locations, 14 road closure records, and 27 spatial co-occurrence relationships, serving as the empirical foundation for KG construction.

### *3.2 Literature-Based Traffic Impact Analysis*

A literature pool of relevant rainfall- and flood-traffic studies [3,4,6–9] was consolidated based on the criterion of providing explicit quantitative values for traffic flow changes as a function of rainfall intensity or flood depth.

Table 1 summarizes the synthesized values across five rainfall intensity levels: light (<2.5 mm/h; speed reduction 2%–5%, capacity reduction 4%–10%), moderate (2.5–7.6 mm/h; 4%–9%, 7%–11%), heavy (7.6–15 mm/h; 6%–17%, 10%–30%), very heavy (15–50 mm/h; 12%–25%+, 15%–32%+), and extreme (>50 mm/h; 21%–40%+, road closure). Ni et al. [4] reported that waterlogging amplifies traffic impacts by 1.3–2.0 times compared to rainfall alone, and Pregnolato et al. [3] developed a depth-disruption function relating flood depth to vehicle speed. These values were directly mapped to the TrafficImpact entities in the KG.

**Table 1:** Synthesized rainfall-traffic impact values across five intensity levels.

| Intensity Level | Rainfall (mm/h) | Speed Reduction | Capacity Reduction | Road Status |
|---|---|---|---|---|
| Light | <2.5 | 2%–5% | 4%–10% | Slow traffic |
| Moderate | 2.5–7.6 | 4%–9% | 7%–11% | Slow traffic |
| Heavy | 7.6–15 | 6%–17% | 10%–30% | Partial closure possible |
| Very Heavy | 15–50 | 12%–25% | 15%–32% | Partial/Full closure |
| Extreme (+Flooding) | >50 | 21%–40% | Road closure | Full closure |

*Source studies [3,4,6–9].

### *3.3 Domain Knowledge Graph Construction*

#### *3.3.1 Entity and Relationship Extraction Pipeline*

Entity and relationship construction proceeded in three deterministic stages rather than via a learned named entity recognition (NER) or relation extraction (RE) model.

(i) Location entity extraction: a hand-curated gazetteer of approximately 81 Jinju place-name candidates was compiled, covering rivers, administrative districts (myeon/dong), urban neighborhoods, bridges, roads, intersections, and landmarks. This gazetteer was applied as a rule-based substring matcher against the 112 filtered documents. From the resulting candidate set, 26 final Location entities were retained in the KG based on corpus appearance frequency; candidates with negligible corpus evidence were excluded.

(ii) Non-spatial entity definition: the 32 remaining entities (5 WeatherCondition, 5 FloodState, 5 TimeContext, 10 TrafficImpact, 7 ResponseAction) were not extracted from text but defined a priori. WeatherCondition levels follow the rainfall-intensity classes derived from the traffic-engineering

literature summarized in Table 1. FloodState, TrafficImpact, and ResponseAction entities consolidate the corresponding domain categories observed in the corpus and literature.

(iii) Relationship extraction: edges were populated from (a) document co-occurrence counts for corpus-grounded relations (SUSCEPTIBLE_TO, AFFECTS, NEAR, FLOOD_CLOSES; 76.5%), (b) literature-derived quantitative values for WORSENS/RESULTS_IN/AMPLIFIES (8.4%), and (c) domain rules for CAUSES/TRIGGERS/ASSOCIATED_WITH (15.1%).

This hybrid rule-based plus expert-curation strategy is an established pattern in domain-specific knowledge graph construction [20]. The pipeline does not employ neural NER or RE; the resulting trade-off (higher reproducibility at the cost of manual labor for new domains) is addressed in Section 5.5.

### *3.3.2 Graph Structure and Constraints*

A domain-specific KG was constructed as a directed weighted graph using NetworkX in Python, comprising 58 nodes across six entity categories and 285 edges representing ten rela-tionship types (Table 2, Fig. 3). The entity categories are: Location (26 nodes—rivers, districts, bridges, and landmarks extracted from the 112-document corpus), WeatherCondition (5 intensity levels), FloodState (5 severity levels), TimeContext (5 seasonal periods), TrafficImpact (10 speed/capacity reduction levels from the literature review), and ResponseAction (7 emergency measures).

**Table 2:** KG entity categories and relationship types.

| Category | Type | Count | Weight Source | Description |
|---|---|---|---|---|
| Entity | Location | 26 | – | ✓ Rivers (4)<br>✓ Districts (8)<br>✓ Neighborhoods (5)<br>✓ Bridges (6)<br>✓ Landmarks (3) |
| | TrafficImpact | 10 | – | ✓ Speed/capacity reduction levels derived from literature review |
| | ResponseAction | 7 | – | ✓ Light<br>✓ Moderate<br>✓ Heavy<br>✓ Very Heavy<br>✓ Extreme rainfall |
| | FloodState | 5 | – | ✓ Water level rise<br>✓ Minor/Major flooding<br>✓ Dam discharge<br>✓ Severe inundation |
| | SUSCEPTIBLE_TO | 81 | Document frequency | Location susceptibility to flood states |
| | AFFECTS | 81 | Document frequency | Flood state impact on locations |

(Continued)

**Table 2 (continued)**

| Category | Type | Count | Weight Source | Description |
|---|---|---|---|---|
| Relationship | NEAR | 42 | Domain rules | Weather condition triggers flood states |
| | FLOOD_CLOSES | 14 | Document frequency | Observed flood-to-road-closure pairings |
| | CAUSES | 14 | Domain rules | Causal flood-traffic relationships |
| | ASSOCIATED_WITH | 13 | Domain rules | Temporal associations |
| | WORSENS | 10 | Literature values | Traffic impact severity escalation |
| | RESULTS_IN | 8 | Literature values | Flood-to-traffic-impact quantitative links |
| | AMPLIFIES | 6 | Literature values | Waterlogging amplification effects |

Edge weights were derived from three sources: document frequency from the corpus (76.5% of edges, including SUSCEPTIBLE_TO, AFFECTS, NEAR, and FLOOD_CLOSES relationships), literature-derived quantitative values (8.4%, including WORSENS, RESULTS_IN, and AMPLIFIES), and domain expert rules (15.1%, including CAUSES, TRIGGERS, and ASSOCIATED_WITH). The FLOOD_CLOSES relationship (14 edges) encodes observed flood-location to road-closure spatial pairings from the corpus (e.g., flooding at Namgang River → closure of Jinyang Bridge, observed in 4 events).

Nine domain constraint rules ensure scenario plausibility; their full specification is given in Table 3. Four allowance rules (A1–A4) define valid combinations; for example, A1 restricts light rainfall to minor speed reductions (2%–5%), while A2 requires that extreme rainfall with severe inundation be paired with a full road closure directive. Five prohibition rules (P1–P5) veto physically implausible scenarios, including P1 (light rainfall cannot cause severe inundation) and P5 (road closures must have an observed FLOOD_CLOSES edge with the flood Location).

FLOOD_CLOSES is a directed binary relation derived from Step-1 event extraction: for every ordered pair (f, r), FLOOD_CLOSES(f, r) is instantiated when the filtered corpus contains one or more documents in which a flood event at Location f co-occurs with a closure of road segment r, with the edge weight set to the normalized co-occurrence count.

The 14 observed FLOOD_CLOSES edges form three spatial clusters (the Namgang corridor, the Yeongcheongang area, and the Sangpyeong lowland), visualized in Fig. 4.

During graph-walk sampling, rule P5 rejects any walk whose (f, r) pair is absent from the FLOOD_CLOSES set; when no direct edge exists, a NEAR-neighbor fallback (21.8% of scenarios) is used. NEAR edges are derived from Location-Location co-occurrence in the same filtered documents, requiring two or more co-mentioning documents for an edge to be instantiated.

Rule-based validation against the nine constraint rules (Table 3) confirmed zero violations across all 800 generated scenarios.

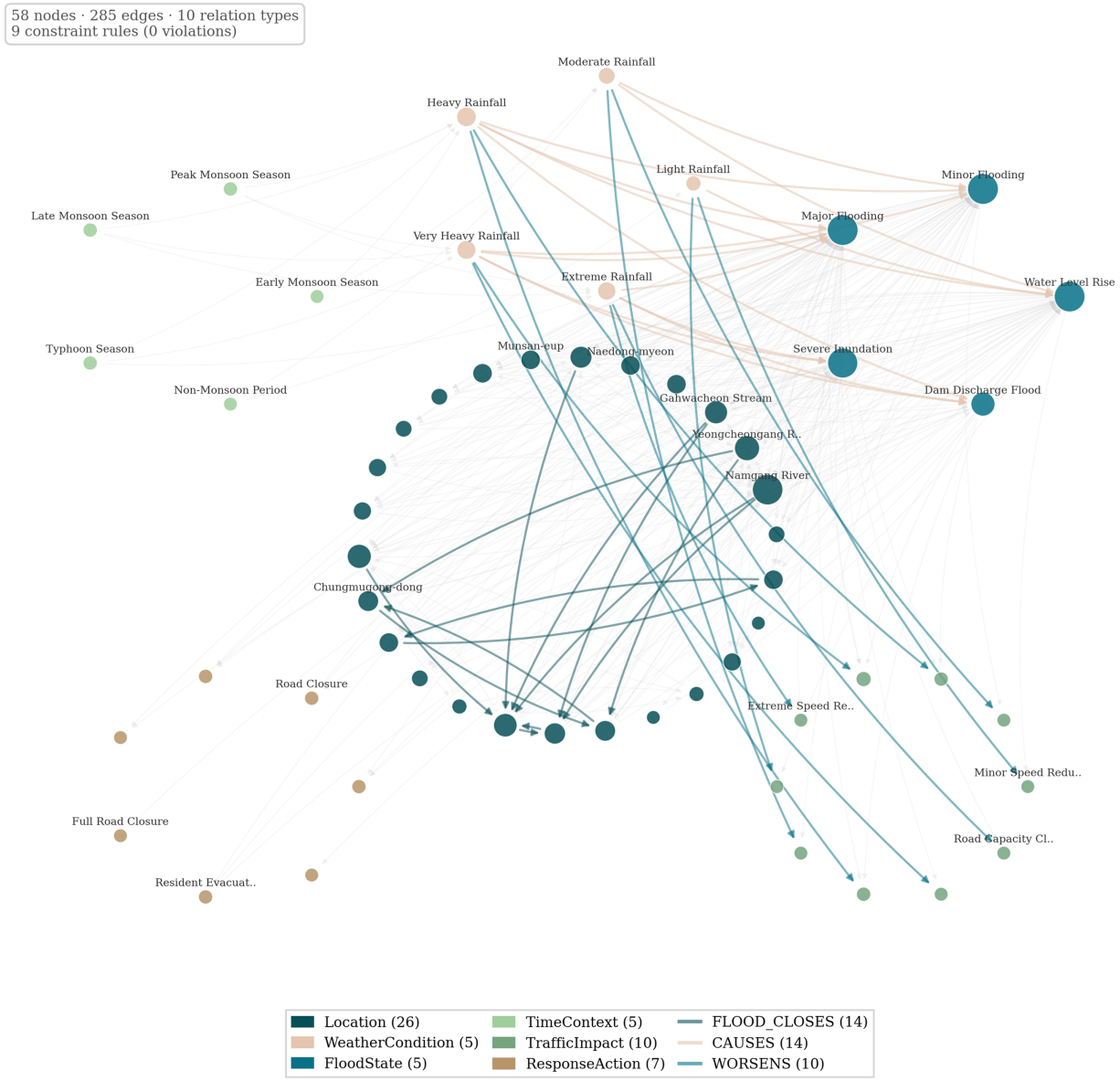


**Figure 3:** Domain knowledge graph visualization for Jinju City urban flood-traffic impact.

**Table 3:** Domain constraint rules (A1–A4 allowance, P1–P5 prohibition).

| ID | Type | Rule |
|---|---|---|
| A1 | Allowance | Light rainfall allows only minor speed reduction (2%–5%). |
| A2 | Allowance | Extreme rainfall with severe inundation requires full road closure. |
| A3 | Allowance | Dam discharge requires at minimum access control and road closure. |
| A4 | Allowance | All rainfall-intensity levels are possible during monsoon season (months 6–9). |
| P1 | Prohibition | Light rainfall cannot cause severe inundation. |
| P2 | Prohibition | Light rainfall cannot trigger full closure or evacuation. |
| P3 | Prohibition | Extreme rainfall is prohibited in the non-monsoon period for Jinju. |

(Continued)

**Table 3 (continued)**

| ID | Type | Rule |
|---|---|---|
| P4 | Prohibition | Speed reduction above 40% requires a flooding condition, not rainfall alone. |
| P5 | Prohibition | Road closures must have an observed FLOOD_CLOSES edge with the flood Location. |

### *3.4 KG-Based Training Data Generation*

Training data was generated through constrained weighted graph walks following a seven-step path: TimeContext → WeatherCondition → FloodState → Location → Closed Road (via FLOOD_CLOSES) → TrafficImpact → ResponseAction. At each step, the next node is selected via weighted probabilistic sampling. For locations without FLOOD_CLOSES edges, an alternative mechanism selects spatially proximate NEAR neighbors.

Each graph walk produces an ordered 7-tuple (TimeContext → WeatherCondition → FloodState → Location → ClosedRoad → TrafficImpact → ResponseAction). The tuple is serialized into a JSON metadata object M with eight fact slots (Section 3.6) and then rendered into a natural-language prompt via a paragraph-level template $t_k$ selected uniformly from $T = \{t_1, \ldots, t_{256}\}$.

The prompt, together with the tuple, forms a (prompt, reference) pair used for instruction fine-tuning; Fig. 5 illustrates the data flow, and Table 4 provides a concrete example of the resulting model input and output formats.

The LLM never accesses the KG at inference time: all KG information is compiled into the training prompt offline, which is precisely the mechanism that allows the fine-tuned sLLM to run locally without a graph store, a prerequisite for disaster-time deployment under connectivity loss.

Each scenario was rendered as a four-paragraph English narrative (situation overview, flood situation, traffic impact, response measures) using four template variants per paragraph (256 combinations). Quantitative values were randomly sampled within KG-defined ranges to produce continuous variation. A total of 800 scenarios were generated with 338 unique entity combinations, validated through 38 automated quality checks with 100% compliance. The dataset was split into training (644), validation (78), and test (78) sets via stratified sampling by weather condition.

### *3.5 Model Architecture and Fine-Tuning*

Three sLLMs were selected to enable systematic comparison across architecture and parameter scale: Flan-T5-Large [13] (770M, encoder-decoder), Qwen2.5-3B-Instruct [14] (3B, decoder-only), and Qwen2.5-7B-Instruct [14] (7B, decoder-only). Flan-T5-Large serves as a cross-family reference baseline. The intra-family pairing of Qwen2.5-3B and Qwen2.5-7B provides a scale-controlled comparison under identical architecture; the implications and limits of this comparison, including the ceiling effect observed on the KG Fact Consistency metric, are discussed in Sections 5.1 and 5.4. All models operate within the 16 GB VRAM constraint of a single commodity GPU, addressing practical requirements for local deployment in disaster response systems.

All models were fine-tuned using QLoRA [15], combining 4-bit NF4 quantization with low-rank adaptation (LoRA) [21] ($r$ = 16, $\alpha$ = 32, dropout = 0.05). Common settings included bfloat16 compute, 8-bit AdamW optimizer, cosine learning rate scheduler (warmup ratio 0.05), and effective batch size of 16.

Model-specific configurations are detailed in Table 2: all three models were trained for 10 epochs with model-specific learning rates. Flan-T5-Large used a learning rate of $3 \times 10^{-4}$ with LoRA on query/value projections (4.7M trainable parameters), while both Qwen models used $2 \times 10^{-4}$ with LoRA on query/key/value/output projections (7.4 and 10.1M trainable parameters, respectively). Experiments were conducted on a single NVIDIA RTX 4060 Ti (16 GB VRAM) with AMD Ryzen 7 7800X3D, using PyTorch 2.9.1 (CUDA 12.8) and HuggingFace Transformers 5.2.0.

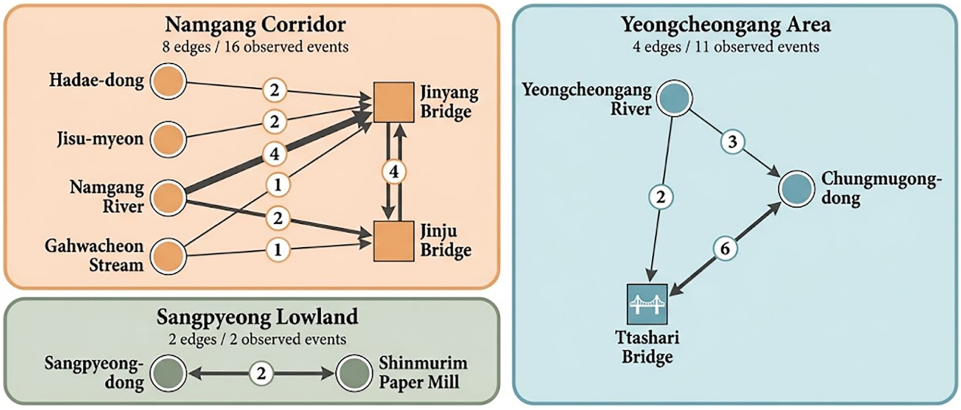


**Figure 4:** FLOOD_CLOSES subgraph.

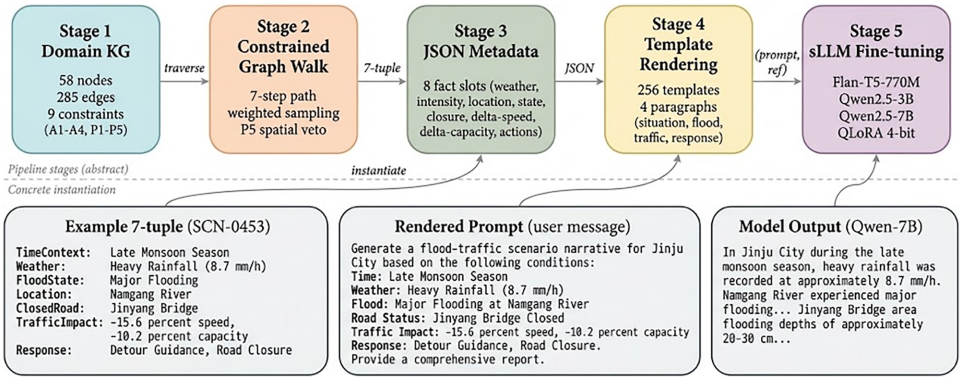


**Figure 5:** Data flow through the KG-driven pipeline.

### *3.6 Evaluation Metrics*

Four complementary metrics were employed: three standard text generation metrics and one domain-specific factual fidelity metric.

BLEU-4 [22] measures $n$-gram precision with a brevity penalty. Given a candidate sentence $c$ and reference sentence $r$ the score is defined as:

$$BLEU - N = BP \times \exp\left(\sum_{n=1}^{N}\left(\frac{1}{N}\log p_n\right)\right) \tag{1}$$

where $p_n$ is the modified $n$-gram precision (clipped to reference counts), and $BP = \min(1, e^{\left(1-\frac{|r|}{|c|}\right)})$ is the brevity penalty. Here, $N = 4$.

**Table 4:** Example of model input and output formats (test sample SCN-0453).

| Component | Content |
|---|---|
| 7-tuple (graph walk) | TimeContext = Late Monsoon Season<br>Weather = Heavy (8.7 mm/h)<br>FloodState = Major Flooding<br>Location = Namgang River<br>ClosedRoad = Jinyang Bridge<br>TrafficImpact = −15.6% speed, −10.2% capacity<br>Response = Detour Guidance, Road Closure |
| JSON metadata | {weather: "Heavy", intensity: 8.7, location: "Namgang River", flood_state: "Major Flooding", closed_road: "Jinyang Bridge", speed_reduction: 15.6, capacity_reduction: 10.2, response_actions: ["Detour Guidance", "Road Closure"]} |
| Rendered prompt | Generate a flood-traffic scenario narrative for Jinju City based on the following conditions:<br>Time: Late Monsoon Season<br>Weather: Heavy Rainfall (8.7 mm/h)<br>Flood: Major Flooding at Namgang River<br>Road Closure: Jinyang Bridge<br>Speed Reduction: 15.6%<br>Capacity Reduction: 10.2%<br>Response Actions: Detour Route Guidance, Road Closure |
| Reference output | Approximately 8.7 mm/h of precipitation classified as heavy rainfall was observed in Jinju City during the late monsoon season . . . [4-paragraph narrative] |
| Model output (Qwen2.5-7B) | In Jinju City during the late monsoon season, heavy rainfall was recorded at approximately 8.7 mm/h. Namgang River experienced major flooding... Jinyang Bridge area flooding depths of approximately 20–30 cm. . . |

ROUGE-L [23] measures the longest common subsequence (LCS) between candidate and reference:

$$R_{lcs} = \frac{LCS\left(r,c\right)}{|r|}, P_{lcs} = \frac{LCS\left(r,c\right)}{|c|}, F_{lcs} = \left(1+\beta^2\right) P_{lcs} R_{lcs} / \left(R_{lcs} + \beta^2 P_{lcs}\right) \tag{2}$$

where $|r|$ and $|c|$ denote reference and candidate lengths, respectively, and $\beta$ is set to favor recall.

BERTScore F1 [24] computes token-level semantic similarity using contextual embeddings. Given reference token embeddings $\mathbb{x} =< x_1, \ldots, x_m >$ and candidate token embeddings $\mathbb{y} =< y_1, \ldots, y_m >$:

$$R_{BERT} = \left(\frac{1}{m}\right) \sum_{i=1}^{m} max_j \cos\left(x_i, y_j\right), P_{BERT} = \left(\frac{1}{n}\right) \sum_{j=1}^{n} max_i \cos\left(x_i, y_j\right) \tag{3}$$

$$F_{BERT} = 2 \times \left(P_{BERT} \times R_{BERT}\right) / \left(P_{BERT} + R_{BERT}\right) \tag{4}$$

The embedding model is roberta-large.

KG Fact Consistency Score: Standard metrics assess text quality but not factual fidelity to input conditions. In data-to-text generation, task-specific factual metrics are commonly employed alongside standard metrics: Wen et al. [25] introduced the slot error rate for dialogue natural language generation (NLG), providing precedent for task-specific evaluation metrics in data-to-text generation. Following this established practice, this study introduces a domain-specific metric that evaluates whether key facts from the input KG triples are faithfully preserved in generated text. For a generated scenario $\hat{y}$ and its corresponding input metadata $m$ containing $K$ fact categories, the score is defined as:

$$KG - FC\,(\hat{y}, m) = \left(\frac{1}{K}\right) \sum_{k=1}^{K} f_k(\hat{y}, m_k) \tag{5}$$

where each $f_k$ is a binary verification function that returns 1 if fact category $k$ is correctly preserved in $\hat{y}$, and 0 otherwise. Eight fact categories ($K = 8$) are defined: (1) weather condition, (2) rainfall intensity (mm/h), (3) flood location, (4) flood state, (5) closed road, (6) speed reduction percentage, (7) capacity reduction percentage, and (8) response actions. Categories 1–7 are verified via exact or partial string matching against the input values, while category 8 computes the fraction of response actions mentioned ($f_8 = \frac{|A_{matched}|}{|A_{total}|}$).

The design uses a uniform averaging scheme: the KG Fact Consistency score is the arithmetic mean of the eight per-category verification values (each in $[0, 1]$), giving equal weight to all facets. Category weights were not learned or tuned, to avoid overfitting the score to the test distribution.

This design is in the same lineage as slot-based factual metrics in data-to-text generation [25,26], which evaluate whether structured input facts are faithfully preserved in surface text. Table 5 specifies the verification rule for each category as implemented in the evaluation code.

**Table 5:** Category-wise verification methods for KG fact consistency.

| No. | Category | Verification | Score Contribution |
|---|---|---|---|
| 1 | Weather condition | Case-insensitive substring match of label | 1 if matched, else 0 |
| 2 | Rainfall intensity (mm/h) | Exact substring match of the numeric value | 1 if matched, else 0 |
| 3 | Flood location | Case-insensitive substring match of Location name | 1 if matched, else 0 |
| 4 | Flood state | Partial-word match: at least 50% of label tokens present | 1 if ≥50% tokens match, else 0 |
| 5 | Closed road | Case-insensitive substring match of road name | 1 if matched, else 0 |
| 6 | Speed reduction (%) | Exact substring match of the numeric value | 1 if matched, else 0 |
| 7 | Capacity reduction (%) | Exact substring match of the numeric value | 1 if matched, else 0 |
| 8 | Resonse actions | Fraction of input action labels present (continuous, 0.0–1.0) | action_matches/ total_actions |

Three consequences of this design are explicitly acknowledged and deferred to future work: (a) the metric measures fact omission but not fabrication (partially addressed via the post-hoc analysis in Section 4.3.4); (b) string-matching verification saturates at 1.0 once a model reliably copies all input facts; and (c) the

metric does not penalize contextual misplacement of otherwise correct values. A subsequent human-expert evaluation targets precisely these three dimensions.

All models were evaluated on the 78-sample test set stratified by weather condition, with both overall and per-condition breakdowns analyzed.

## 4 Results

### *4.1 Data Refinement Results*

From 254 raw documents (89 news articles, 165 disaster alerts), two-stage filtering retained 112 documents (Table 6). Keyword filtering reduced the corpus to 138, and re-validation removed 26 false positives (sinkholes, construction, festivals). LDA topic modeling ($k$ = 15) confirmed four dominant thematic clusters related to urban flood-traffic scenarios with zero off-topic documents. Structured extraction yielded 67 flood-affected locations and 14 road closure records, with Namgang River as the most frequently mentioned location (62 occurrences). Table 6 summarizes the two-stage filtering pipeline and the stage-by-stage corpus reductions.

**Table 6:** Two-stage data filtering results.

| Stage | Description | Input | Removed | Output |
|---|---|---|---|---|
| – | Raw corpus (news articles + disaster alerts) | – | – | 254 (89 + 165) |
| 1 | Hierarchical keyword filtering | 254 | 116 | 138 |
| 2 | Non-flood content removal | 138 | 26 | 112 (82 + 30) |

### *4.2 KG Construction and Data Generation Results*

The domain KG comprises 58 nodes across six entity categories and 285 edges representing ten relationship types (Table 2, Fig. 3). Edge weights were derived from document frequency (76.5%), literature values (8.4%), and domain rules (15.1%). The FLOOD_CLOSES relationship (14 edges) encodes observed flood-to-road-closure spatial pairings, with three spatial clusters identified: the Namgang River corridor, the Yeongcheongang area, and the Sangpyeong lowland. Rule-based validation confirmed zero constraint violations.

Using the KG, 800 scenarios were generated in approximately 5 min (Table 7). The dataset exhibits realistic distributions: heavy to very heavy rainfall constitutes 62.3% of conditions, and Namgang River accounts for 27.5% of locations, proportional to its corpus frequency.

**Table 7:** KG-based data generation statistics.

| Item | Value |
|---|---|
| Total scenarios generated | 800 |
| Unique entity combinations | 338 |
| Generation time | ~5 min |
| Mean text length | 737.9 chars (σ = 65.9) |
| Automated quality checks | 38 checks, 100% compliance |
| Spatial consistency violations | 0 |
| **Distribution** | |

(Continued)

**Table 7 (continued)**

| Item | Value |
|---|---|
| Heavy Rainfall | 305 (38.1%) |
| Very Heavy Rainfall | 194 (24.3%) |
| Moderate Rainfall | 133 (16.6%) |
| Extreme Rainfall | 103 (12.9%) |
| Light Rainfall | 65 (8.1%) |
| **Dataset Split** | |
| Training | 644 |
| Validation | 78 |
| Test | 78 |

### *4.3 Quantitative Evaluation Results*

#### *4.3.1 Overall Model Performance*

Table 8 presents the evaluation results. Qwen2.5-7B achieved the highest scores across all metrics: BLEU-4 of 0.5524, ROUGE-L of 0.6883, BERTScore F1 of 0.9662, and KG Fact Consistency of 1.0000. Performance improves monotonically from 770M through 3B to 7B.

**Table 8:** Overall evaluation metrics for three sLLMs.

| Metric | Flan-T5-Large (770M) | Qwen2.5-3B (3B) | Qwen2.5-7B (7B) |
|---|---|---|---|
| BLEU-4 | 0.4128 | 0.5445 | 0.5524 |
| ROUGE-L | 0.5914 | 0.6744 | 0.6883 |
| BERTScore F1 | 0.9415 | 0.9653 | 0.9662 |
| KG Fact Consistency | 0.8998 | 1.0000 | 1.0000 |

*Best model: Qwen2.5-7B across all metrics; KG Fact Consistency scores of 1.0000 represent the ceiling of the string-matching verification used in Section 3.6; they should not be interpreted as perfect factual quality. See Section 5.4 Limitation #2.

The largest improvement occurs between Flan-T5-Large and Qwen2.5-3B: BLEU-4 increases by 31.9% (0.4128 → 0.5445), ROUGE-L by 14.0%, and KG Fact Consistency by 11.1% (0.8998 → 1.0000). The further improvement from 3 to 7B is modest (BLEU-4: +1.5%, ROUGE-L: +2.1%), suggesting diminishing returns within the same architectural family (Fig. 6).

Both Qwen models achieve KG Fact Consistency of 1.0000, reproducing all input conditions across all 78 test samples.

#### *4.3.2 Training Efficiency*

All models were trained for 10 epochs on a single GPU (Table 9). Training time scales with model size: Flan-T5-Large completes in 45.7 min, Qwen2.5-3B in 104.5 min, and Qwen2.5-7B in 146.2 min. VRAM scales predictably: Flan-T5-Large uses 3.16 GB (19.8%), Qwen2.5-3B uses 7.82 GB (48.9%), and Qwen2.5-7B reaches 14.63 GB (91.4%), representing the practical upper limit for 16 GB GPUs. Total training time for all three models was approximately 5.0 h.

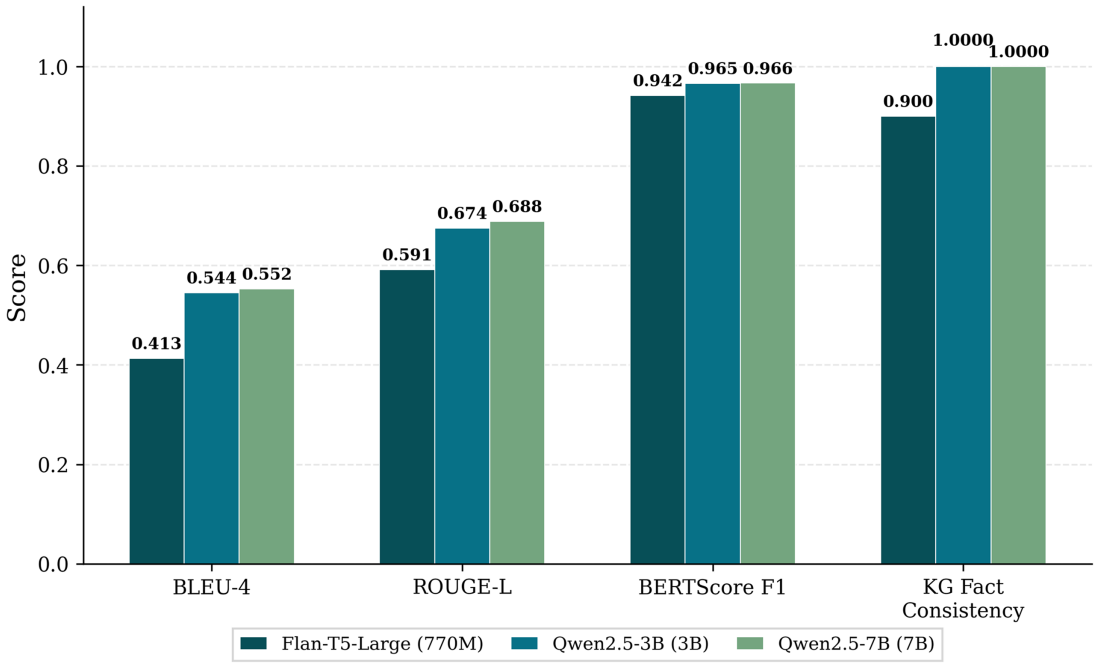


**Figure 6:** Grouped bar chart comparing evaluation metrics across three models.

**Table 9:** Model architecture, QLoRA configuration, and training resource usage.

| Contents | Flan-T5-Large | Qwen2.5-3B-Instruct | Qwen2.5-7B-Instruct |
|---|---|---|---|
| Architecture | Encoder-Decoder | Decoder-Only | Decoder-Only |
| Total Parameters | 498M | 1.71B | 4.36B |
| Trainable Parameters | 4.7M (0.95%) | 7.4M (0.43%) | 10.1M (0.23%) |
| LoRA Target Modules | $q, v$ | $q, k, v, o$ | $q, k, v, o$ |
| Learning Rate | 0.0003 | 0.0002 | 0.0002 |
| Batch × Accum (Effective) | 4 * 4 | 2 * 8 | 1 * 16 |
| Training Time (min) | 45.7 | 104.5 | 146.2 |
| Final Train/Eval Loss | 15.46/0.069 | 0.309/0.100 | 0.283/0.099 |

*Software: PyTorch 2.9.1 (CUDA 12.8), HuggingFace Transformers 5.2.0.

Training convergence analysis (Fig. 7) reveals that Flan-T5-Large starts from a high initial loss (15.15) and reaches a final evaluation loss of 0.069 over 10 epochs, while both Qwen models start from much lower initial losses (0.44 and 0.30, respectively) and converge within 3–4 epochs, with marginal improvement thereafter, reflecting their stronger pre-training foundation and suggesting that the additional training epochs primarily consolidate rather than substantially improve performance.

### *4.3.3 Weather-Condition Breakdown*

Fig. 8 presents the per-weather-condition breakdown of ROUGE-L and KG Fact Consistency. Flan-T5-Large shows notable variation across conditions: ROUGE-L ranges from 0.559 (Very Heavy) to 0.692 (Light Rainfall), and KG Fact Consistency from 0.792 (Light) to 0.961 (Very Heavy), suggesting difficulty with subtler weather conditions.

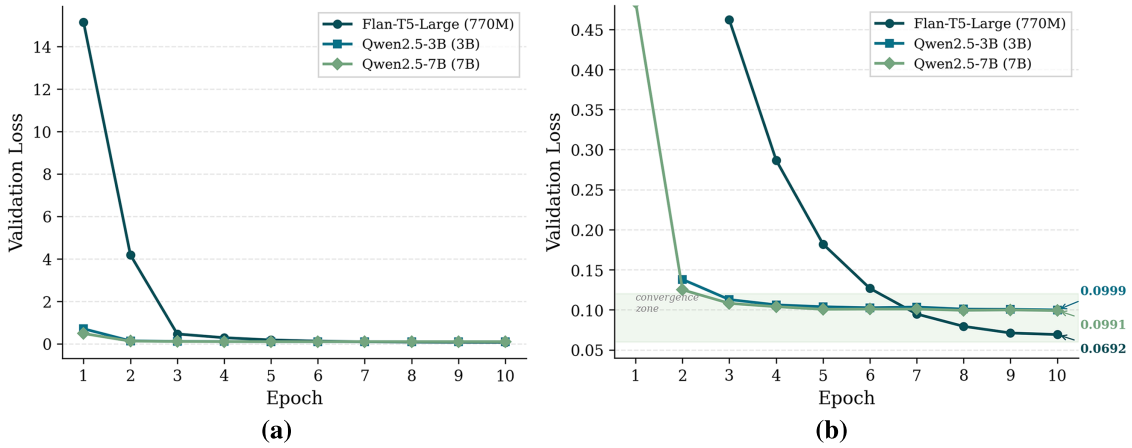


**Figure 7:** Training loss curves for three models: (**a**) Full training curves; (**b**) Convergence detail (Loss < 0.5).

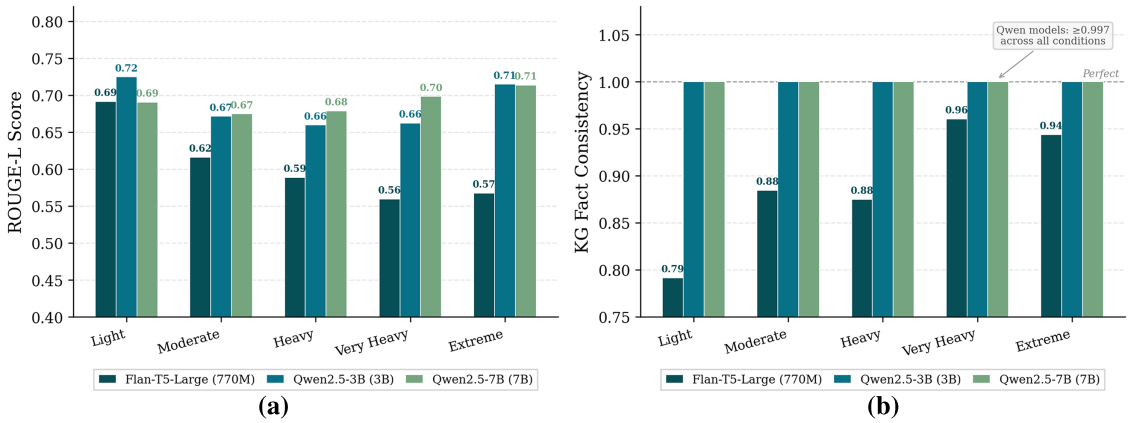


**Figure 8:** Per-weather-condition breakdown: (**a**) ROUGE-L by weather conditions; (**b**) KG fact consistency by weather condition.

In contrast, both Qwen models exhibit weather-invariant performance, achieving KG Fact Consistency of 1.0000 across all five weather categories. Notably, Qwen2.5-3B slightly outperforms Qwen2.5-7B in ROUGE-L for Light Rainfall (0.725 vs. 0.691) and Extreme Rainfall (0.715 vs. 0.714) scenarios, suggesting that the smaller model may capture template patterns slightly differently for edge conditions.

### *4.3.4 Location Fabrication Analysis*

To complement KG Fact Consistency with a measure of fact fabrication (rather than omission), a post-hoc verification step was applied: for each of the 234 generated test outputs (78 samples × 3 models), Location-like tokens were extracted by substring matching against the 26-entry Location gazetteer of the KG and flagged any generated Location not present in the input metadata.

Across all three models (Flan-T5-Large, Qwen2.5-3B, Qwen2.5-7B), the observed cross-reference fabrication rate was 0.00%; no prediction contained a Location outside of its input conditions. This result indicates that KG-grounded fine-tuning effectively constrains the output Location distribution to the set of Locations present in the input conditions, providing a second, independent factual signal complementary to KG Fact Consistency.

This analysis addresses only Location-level fabrication against the KG gazetteer; fabrication of other fact categories (rainfall values, closure status, response directives) and contextual misplacement of otherwise correct values remain subjects for formal human evaluation (Section 5.4).

### *4.4 Scenario Generation Examples*

Table 10 compares model outputs for a representative Heavy Rainfall scenario (8.7 mm/h, Namgang River). All three models correctly preserve key facts: rainfall intensity, flood location, traffic impacts (15.6% speed reduction, 10.2% capacity reduction), and response actions. Qualitative differences emerge in style: Flan-T5-Large produces more verbose narratives with connective phrases, while Qwen models generate more structured, report-like text. Qwen2.5-7B produces the most concise output while preserving all factual elements.

**Table 10:** Scenario output comparison for heavy rainfall condition.

| **Input Conditions** | **Heavy Rainfall, 8.7 mm/h, Late Monsoon Season, Namgang River, Major Flooding, Jinyang Bridge closure, Speed reduction 15.6%, Capacity reduction 10.2%, Response: detour route guidance + road closure** |
|---|---|
| Reference (Ground Truth) | Approximately 8.7 mm/h of precipitation classified as heavy rainfall was observed in Jinju City during the late monsoon season.<br>Due to the persistent rainfall, Namgang River experienced major flooding with significant water accumulation on roadways. The rising water levels surged beyond containment levels, causing widespread overflow, affecting the Jinyang Bridge area with flooding depths of approximately 20–30 cm, sufficient to impede normal vehicle operation.<br>Traffic conditions in the affected area deteriorated significantly: vehicle speeds decreased by approximately 15.6% compared to normal conditions, and road capacity was reduced by approximately 10.2%.<br>In response, authorities implemented detour route guidance and road closure. Alternative routes were communicated to drivers via electronic message signs and local media. |
| Flan-T5-Large | Flooding at Namgang River resulted in major flooding with significant water accumulation on roadways in the Jinju City area. Flood waters reached the Jinyang Bridge area, causing water accumulation of approximately 10–20 cm. Traffic conditions in the affected area deteriorated significantly: vehicle speeds decreased by approximately 15.6% compared to normal conditions, and road capacity was reduced by approximately 10.2%. Emergency management personnel-initiated detour route guidance and road closure. Alternative routes were communicated to drivers via electronic message signs and local media.<br>*Single-paragraph output; omits rainfall intensity (8.7 mm/h) and season context; flood depth slightly different (10–20 cm vs. 20–30 cm); all traffic impact values correctly preserved.* |

(Continued)

**Table 10 (continued)**

| Input Conditions | Heavy Rainfall, 8.7 mm/h, Late Monsoon Season, Namgang River, Major Flooding, Jinyang Bridge closure, Speed reduction 15.6%, Capacity reduction 10.2%, Response: detour route guidance + road closure |
|---|---|
| Qwen2.5-3B | Approximately 8.7 mm/h of precipitation classified as heavy rainfall was observed in Jinju City during the late monsoon season.<br>Major flooding with significant water accumulation on roadways was reported at Namgang River as water levels surged beyond containment levels, causing widespread overflow. The flooding extended to the Jinyang Bridge area, resulting in flooding depths of approximately 20–30 cm, sufficient to impede normal vehicle operation.<br>Traffic flow analysis indicated a speed reduction of approximately 15.6% and a capacity reduction of approximately 10.2% in the flood-affected road segments.<br>The Jinju City traffic management center activated detour route guidance and road closure. Alternative routes were communicated to drivers via electronic message signs and local media.<br>*Four-paragraph structure preserved; all factual elements correctly reproduced; minor stylistic variations (e.g., “traffic flow analysis indicated” vs. “traffic conditions deteriorated”).* |
| Qwen2.5-7B | In Jinju City during the late monsoon season, heavy rainfall was recorded at approximately 8.7 mm/h.<br>Major flooding with significant water accumulation on roadways was reported at Namgang River as water levels surged beyond containment levels, causing widespread overflow. The flooding extended to the Jinyang Bridge area, resulting in flooding depths of approximately 20–30 cm, sufficient to impede normal vehicle operation.<br>Traffic flow analysis indicated a speed reduction of approximately 15.6% and a capacity reduction of approximately 10.2% in the flood-affected road segments.<br>Local authorities responded by implementing detour route guidance and road closure. Alternative routes were communicated to drivers via electronic message signs and local media.<br>*Four-paragraph structure preserved; all factual elements correctly reproduced; most concise opening sentence while retaining all key information (season, intensity, classification).* |

Three qualitative observations complement the quantitative metrics above; they do not constitute a human evaluation and are intended to illustrate cases where automatic metrics align with, or diverge from, intuitive judgment. (a) Convergent case (Extreme Rainfall): Qwen-7B preserves all eight input facts and the narrative coherently escalates to a full closure directive, consistent with its KG Fact Consistency of 1.0000. (b) Metric-insensitive case (Heavy Rainfall): Qwen-3B and Qwen-7B both score 1.0000 on KG Fact Consistency, yet their outputs differ in stylistic register (the 7B variant is noticeably more concise), a distinction that the automatic metrics do not capture. (c) Blind-spot case (Light Rainfall): Flan-T5-Large preserves the numeric values (speed and capacity reductions) correctly, so they pass string-matching verification, yet the connective phrasing places these values in an operationally ambiguous context—a form of contextual misplacement that is invisible to the present metric and motivates the supplementary evaluation avenues discussed in Section 5.4. These qualitative observations are illustrative rather than a formal human

evaluation, acknowledging that automatic metrics have known limitations in capturing factual and semantic adequacy [27,28].

## 5 Discussion

### *5.1 Architectural and Scale Effects on Scenario Generation*

The results reveal two factors potentially influencing performance: architecture and parameter scale. The Flan-T5 vs. Qwen contrast varies both factors simultaneously and cannot isolate an architectural effect; Flan-T5-Large is best interpreted as a cross-family reference baseline.

The intra-family Qwen2.5-3B vs. 7B contrast provides a scale-controlled comparison under identical architecture. The scale effect on KG Fact Consistency is obscured by ceiling saturation (both Qwen models reach 1.0000), and the observable scale effect is limited to modest improvements on lexical metrics (BLEU-4: +1.5%; ROUGE-L: +2.1%).

The controlled within-family comparison (Qwen2.5-3B vs. 7B) provides a scale-controlled contrast under identical architecture: scaling from 3 to 7B yields consistent but modest improve-ments on lexical metrics (BLEU-4: +1.5%, ROUGE-L: +2.1%), while both models saturate at KG Fact Consistency (1.0000), a consequence of the string-matching ceiling effect described in Section 5.4. Under this ceiling, the 3B model already matches the 7B model on the factual-preservation axis, so the additional parameters of the 7B model manifest primarily in text-generation quality rather than factual accuracy. This pattern of diminishing returns is consistent with scaling literature suggesting that task-specific improvements plateau when training data is domain-constrained.

The weather-condition breakdown provides additional insight. Flan-T5-Large's KG Fact Consistency shows an inverse relationship with scenario complexity (0.792 for Light vs. 0.961 for Very Heavy Rainfall), suggesting difficulty with subtler conditions. Both Qwen models maintain consistency of 1.0000 across all weather conditions, demonstrating robust conditional generation. Notably, Qwen2.5-3B slightly outperforms Qwen2.5-7B in ROUGE-L for Light Rainfall (0.725 vs. 0.691), potentially reflecting minor differences in template pattern capture between model scales.

The design role of each constraint rule is analyzed through counterfactual reasoning grounded in the zero-violation outcome of the rule-based validation.

Removing P5 (FLOOD_CLOSES spatial consistency) would admit closures at geographically unrelated bridges (e.g., flooding at Yeongcheongang River paired with closure at Jinyang Bridge), violating the spatial pairings observed in the 112-document corpus. Removing prohibition rules P1–P4 would readmit physically implausible combinations such as Light rainfall producing Severe Inundation, which the corpus never documents. Allowance rules A1–A4 govern upper-bound co-occurrences; their removal would widen the scenario distribution at the cost of operational plausibility, particularly for safety-critical response directives.

A controlled retraining-based ablation quantifying these effects is an important direction for future work.

### *5.2 Effectiveness of KG-Generated Training Data*

The results demonstrate that KG-generated synthetic training data provides sufficient signal for effective domain fine-tuning. Even the smallest model achieves BERTScore F1 of 0.9415, indicating strong semantic alignment. Both Qwen models achieve KG Fact Consistency of 1.0000 across all 78 test samples, preserving all eight fact categories (weather condition, rainfall intensity, flood location, flood state, closed road, speed reduction, capacity reduction, response actions)—validating the KG-driven approach as a reliable method for grounding model outputs in structured domain knowledge.

This design reflects a deliberate separation of concerns between the KG and sLLM components. At the 770M–7B parameter scale, sLLMs lack the capacity for reliable causal reasoning over physical constraints (e.g., inferring that light rainfall cannot cause severe inundation). Rather than relying on the model to learn such relationships from limited training data, the pipeline delegates constraint enforcement entirely to the KG (whose rule-based validation guarantees 100% compliance), while the sLLM is tasked solely with data-to-text generation: transforming structured KG-derived conditions into coherent natural language narratives. This division ensures that factual plausibility is deterministically guaranteed by the KG, independent of model scale or training data volume.

The KG-driven approach offers three key advantages: (1) constraint rules (P1–P5, A1–A4) prevent physically implausible scenarios critical for disaster management; (2) the FLOOD_CLOSES relationship ensures spatial consistency between flood locations and road closures, addressing a common limitation of template-based approaches; and (3) literature-derived traffic impact values provide empirically grounded quantitative parameters distinguishing generated scenarios from those of general-purpose LLMs.

The pipeline generates 800 scenarios in approximately 5 min after the one-time KG construction, enabling on-demand dataset expansion as new empirical data become available.

### *5.3 Practical Trade-Off Analysis for Deployment*

The comparative evaluation provides practical deployment recommendations. Qwen2.5-3B achieves 98.6% of the 7B's BLEU-4 and identical KG Fact Consistency (1.0000) while requiring only 53% of peak VRAM (7.82 vs. 14.63 GB) and 71% of training time. For municipal systems with 8–12 GB VRAM, the 3B model represents the most cost-effective choice, achieving the same factual consistency as the 7B model at substantially lower computational cost.

Qwen2.5-7B achieved the highest text generation scores but operates at the practical limit of 16 GB GPUs (91.4% VRAM utilization), and is therefore recommended for higher-capacity GPUs or applications requiring maximum text quality. Flan-T5-Large (3.16 GB VRAM) is the most resource-efficient but exhibits substantially lower performance, suitable only for severely constrained hardware or as a baseline.

All models were trained for 10 epochs on a single GPU, with per-model training times ranging from 46 to 146 min, demonstrating practical feasibility for municipal disaster response deployment.

### *5.4 Limitations*

Five limitations should be acknowledged.

First, the study focuses on a single city (Jinju), and generalizability to different flooding mechanisms (coastal, snowmelt) has not been validated.

Second, the Flan-T5 vs. Qwen comparison varies architecture and parameter scale simultaneously and cannot isolate an architectural effect. Flan-T5-Large is interpreted as a cross-family reference baseline rather than as evidence of an encoder-decoder vs. decoder-only performance difference; the controlled comparison is restricted to the intra-family Qwen2.5-3B vs. 7B contrast. A matched decoder-only baseline near 770M parameters would further disentangle architectural and scale effects.

Third, the test set (78 samples) limits statistical power for per-weather-condition comparisons.

Fourth, the KG Fact Consistency metric is subject to a ceiling effect: once a model reliably copies all input facts, the score saturates at 1.0 and becomes insensitive to subsequent factual quality dimensions, in particular fabrication of facts not present in the input and contextual misplacement of otherwise correct values.

The post-hoc Location Fabrication Analysis (Section 4.3.4) showed zero cross-reference fabrication across all 234 test predictions, providing a second independent factual signal on the Location-entity axis; this analysis does not cover fabrication of other fact categories or broader semantic errors [5].

Fifth, evaluation relies on automated metrics without human expert assessment. Automatic text-generation metrics have known limitations in capturing deeper semantic dimensions such as rationality, coherence, and information completeness [27,28]. A blinded evaluation with ≥3 domain experts reporting Fleiss' κ would complement the present automated and KG-based fidelity measures.

Notwithstanding these limitations, the study establishes that KG-driven training-data construction enables sLLMs fine-tuned on commodity hardware to achieve KG Fact Consistency of 1.0000 with zero Location-level cross-reference fabrication across 234 test predictions. The FLOOD_CLOSES spatial-consistency rule offers a reusable mechanism for injecting safety-critical spatial constraints into generation, and the nine-rule constraint framework provides a template adaptable to other urban-disaster domains.

The limitations enumerated above frame the scope of validity and motivate methodological refinement, rather than nullify the core finding that KG-grounded fine-tuning can constrain sLLMs to produce factually consistent scenarios for disaster preparedness. Scalability of the manually curated KG is addressed separately in Section 5.5.

### *5.5 KG Transferability and Scalability*

The present KG was partially hand-curated, which raises a legitimate concern about transferability to other cities or flooding regimes beyond the monsoon-driven case of Jinju (e.g., coastal, snowmelt). Three elements are outlined below of a scalable path forward.

(i) Minimum data requirement. In the present study's application to Jinju, reliable KG construction required on the order of 100 filtered local documents and 10+ quantitative literature sources; this serves as an operational lower bound, not a proven threshold, for replicating the pipeline in a new city.

(ii) Automation roadmap. Neural NER pipelines can replace the rule-based Location extraction used in Section 3.3.1 (Stage 1). Pre-trained end-to-end relation-extraction models such as REBEL [29] can bootstrap the relationship layer from a seed of hand-curated triples (Stage 2). The nine domain constraint rules remain expert-defined (Stage 3). Stage 3's expert authorship prevents physically implausible combinations (such as light rainfall paired with severe inundation, or floods paired with closures at geographically unrelated bridges) from propagating into disaster-preparedness guidance.

(iii) Human-in-the-loop for safety. Fully learned constraint generation is an open problem is not resolved here; keeping the constraint layer expert-authored is a deliberate design choice rather than an engineering limitation.

## 6 Conclusions

This study proposed a KG-driven pipeline for automatically generating domain-specific training data and fine-tuning sLLMs for urban flood-traffic scenario synthesis. The domain KG (58 entities, 285 relationships, 9 constraint rules) effectively encoded complex causal, spatial, and quantitative relationships by integrating empirical flood data from 112 local documents with traffic impact values from 14 international studies. The FLOOD_CLOSES spatial consistency rule ensured 100% spatial consistency across all 800 generated scenarios, and the automated pipeline produced the full dataset in approximately 5 min, transforming a manual, labor-intensive authoring process into an on-demand generation step.

The comparative evaluation across three sLLMs (770M, 3B, 7B) revealed monotonic performance improvements with increasing model scale, with the largest gains between Flan-T5-Large and Qwen2.5-3B

(BLEU-4: +31.9%, KG Fact Consistency: +11.1%) and diminishing returns from 3 to 7B. Both Qwen models achieved KG Fact Consistency of 1.0000 across all 78 test samples and all five weather conditions. The 3B model emerges as the optimal choice for municipal deployment, achieving 98.6% of maximum BLEU-4 at 53% of VRAM cost with identical factual consistency. All training was completed on a single GPU within approximately 5.0 h total.

The primary contributions are two-fold: (1) a reusable KG-driven pipeline that transforms domain knowledge into structured training data through constrained graph walks, offering a principled alternative to manual scenario authoring; and (2) empirical evidence that sLLMs fine-tuned with QLoRA on KG-generated data can achieve factual consistency of 1.0000 on commodity hardware, supporting local deployment in connectivity-limited disaster scenarios.

**Acknowledgement:** Not applicable.

**Funding Statement:** This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2026-25494446), and by the KICT Research Program (Project No. 20250284–001, Development of Digital Urban Flood Control Technology for the Realization of Flood Safety City) funded by the Ministry of Science and ICT (MSIT).

**Author Contributions:** The authors confirm their contributions to the paper as follows: study conception and design: Geunhwi Park, Juneyoung Park and Chunjoo Yoon; data collection: Geunhwi Park; analysis and interpretation of results: Geunhwi Park, Juneyoung Park and Jaehong Park; draft manuscript preparation: Geunhwi Park, Juneyoung Park and Chunjoo Yoon. All the authors reviewed and approved the final version of the manuscript.

**Availability of Data and Materials:** Not applicable.

**Ethics Approval:** Not applicable.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. IPCC. Climate change 2023: synthesis report. Geneva, Switzerland: Intergovernmental Panel on Climate Change; 2023. doi:10.59327/IPCC/AR6-9789291691647.
2. Rosenzweig C, Solecki WD, Romero-Lankao P, Mehrotra S, Dhakal S, Ali Ibrahim S, editors. Climate change and cities: second assessment report of the urban climate change research network. Cambridge, UK: Cambridge University Press; 2018. doi:10.1017/9781316563878.
3. Pregnolato M, Ford A, Wilkinson SM, Dawson RJ. The impact of flooding on road transport: a depth-disruption function. Transp Res Part D Transp Environ. 2017;55(1):67–81. doi:10.1016/j.trd.2017.06.020.
4. Ni X, Huang H, Chen A, Liu Y, Xing H. Effect of heavy rainstorm and rain-induced waterlogging on traffic flow on urban road sections: integrated experiment and simulation study. J Transp Eng Part A Syst. 2021;147(10):04021057. doi:10.1061/jtepbs.0000557.
5. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Comput Surv. 2023;55(12):1–38. doi:10.1145/3571730.
6. Tsapakis I, Cheng T, Bolbol A. Impact of weather conditions on macroscopic urban travel times. J Transp Geogr. 2013;28(2):204–11. doi:10.1016/j.jtrangeo.2012.11.003.
7. Koetse MJ, Rietveld P. The impact of climate change and weather on transport: an overview of empirical findings. Transp Res Part D Transp Environ. 2009;14(3):205–21. doi:10.1016/j.trd.2008.12.004.
8. National Academies of Sciences Engineering and Medicine, Transportation Research Board. Highway capacity manual 7th edition: a guide for multimodal mobility analysis. Washington, DC, USA: National Academies Press; 2022. doi:10.17226/26432.

9. Choo KS, Kang DH, Kim BS. Impact assessment of urban flood on traffic disruption using rainfall–depth–vehicle speed relationship. Water. 2020;12(4):926. doi:10.3390/w12040926.
10. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS); 2020 Dec 6–12; Vancouver, BC, Canada. p. 1877–901.
11. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: open foundation and fine-tuned chat models. arXiv:2307.09288. 2023. doi:10.48550/arXiv.2307.09288.
12. Moradi M, Yan K, Colwell D, Samwald M, Asgari R. A critical review of methods and challenges in large language models. Comput Mater Contin. 2025;82(2):1681–98. doi:10.32604/cmc.2025.061263.
13. Chung HW, Hou L, Longpre S, Zoph B, Tay Y, Fedus W, et al. Scaling instruction-finetuned language models. J Mach Learn Res. 2024;25(70):1–53.
14. Yang A, Yang B, Zhang B, Hui B, Zheng B, Yu B, et al. Qwen2.5 technical report. arXiv:2412.15115. 2024. doi:10.48550/arXiv.2412.15115.
15. Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLoRA: efficient finetuning of quantized LLMs. In: Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS); 2023 Dec 10–16; New Orleans, LA, USA. p. 10088–115. doi:10.52202/075280-0441.
16. Shen A, Lai Z, Li D, Hu X. Optimizing fine-tuning in quantized language models: an in-depth analysis of key variables. Comput Mater Contin. 2025;82(1):307–25. doi:10.32604/cmc.2024.057491.
17. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems; 2020 Dec 6–12; Vancouver, BC, Canada. p. 9459–74. doi:10.5555/3495724.3496517.
18. Liu Y, Guo Q, Yang C, Liao Y. TIPS: tailored information extraction in public security using domain-enhanced large language model. Comput Mater Contin. 2025;83(2):2555–72. doi:10.32604/cmc.2025.060318.
19. Chen M, Tao Z, Tang W, Qin T, Yang R, Zhu C. Enhancing emergency decision-making with knowledge graphs and large language models. Int J Disaster Risk Reduct. 2024;113(4):104804. doi:10.1016/j.ijdrr.2024.104804.
20. Hogan A, Blomqvist E, Cochez M, D'amato C, De Melo G, Gutierrez C, et al. Knowledge graphs. ACM Comput Surv. 2022;54(4):1–37. doi:10.1145/3447772.
21. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: low-rank adaptation of large language models. arXiv:2106.09685. 2021.
22. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL); 2002 Jul 6–12; Philadelphia, PA, USA. p. 311–8. doi:10.3115/1073083.1073135.
23. Lin CY. ROUGE: a package for automatic evaluation of summaries. In: Text Summarization Branches Out: Proceedings of the ACL Workshop; 2004 Jul 25–26; Barcelona, Spain. p. 74–81.
24. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: evaluating text generation with BERT. arXiv:1904.09675. 2019.
25. Wen TH, Gašić M, Mrkšić N, Su PH, Vandyke D, Young S. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2015 Sep 17–21; Lisbon, Portugal. p. 1711–21. doi:10.18653/v1/D15-1199.
26. Maynez J, Narayan S, Bohnet B, McDonald R. On faithfulness and factuality in abstractive summarization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL); 2020 Jul 5–10; Online. p. 1906–19. doi:10.18653/v1/2020.acl-main.173.
27. Reiter E. A structured review of the validity of BLEU. Comput Linguist. 2018;44(3):393–401. doi:10.1162/coli_a_00322.

28. Honovich O, Choshen L, Aharoni R, Neeman E, Szpektor I, Abend O. $Q^2$: evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2021 Nov 7–11; Punta Cana, Dominican Republic. p. 7856–70. doi:10.18653/v1/2021.emnlp-main.619.
29. Huguet Cabot PL, Navigli R. REBEL: relation extraction by end-to-end language generation. In: Findings of the Association for Computational Linguistics: EMNLP 2021; 2021 Nov 7–11; Punta Cana, Dominican Republic. p. 2370–81. doi:10.18653/v1/2021.findings-emnlp.204.