R01G-SR: Provenance-Preserving Selection over Complementary Evidence Views for Crisis Situation Report Generation

Yuqiu Gao[♡♥], **Yuanhe Tian**[♥*], **Xiaofei Chen**[♡], **Zhentao Yang**[♡], **Yan Song**[♠]
[♡]Southern University of Science and Technology [♥]Zhongguancun Academy
[♠]University of Science and Technology of China
[♡]s-gyq25@bza.edu.cn [♥]yhtian94@gmail.com

Abstract

Crisis situation report generation organizes evolving information into structured summaries with traceable supporting evidence. Generating these reports from noisy social-media streams requires balancing evidence completeness and concentration. Unfiltered evidence retains rare observations but includes competing material and remains subject to finite prompt budgets, while filtering and compression risk discarding useful sparse evidence. In this paper, we propose R01G-SR, a provenance-preserving approach that exploits complementary evidence views through whole-report selection. The approach generates two candidate reports using the same quantized 4B generator, presenting unfiltered chronological evidence to one route and filtered, near-duplicate-compressed evidence to the other. A logistic selector uses input- and report-derived features to choose one complete candidate without merging or rewriting reports, and a separate deterministic compliance layer enforces structural and evidence-ID requirements. Nested crisis-disjoint evaluation shows aggregate improvements over both fixed routes, and the official hidden-test evaluation shows improved report quality over the submitted compressed-route fallback. Analysis reveals larger selection gains on longer inputs and later windows, while early windows account for most remaining loss relative to selecting the better available candidate. These findings support evidence-view selection as a useful component of evidence-linked crisis reporting.[1]

1 Introduction

Situation report generation summarizes the evolving state of a crisis, including its impact, needs, and response, to support a coherent understanding of the situation. Social-media streams provide timely observations about crisis events (Lewis et al., 2025), but their short, informal (Zhang et al., 2018; Nie et al., 2020), fragmented, repetitive, and noisy content complicates the identification and organization of relevant information. A useful report must integrate information scattered across messages while preserving the evidence supporting its claims. The LT4CPR 2026 shared task formalizes these requirements by asking systems to generate a situation report in a fixed schema for each input cell and attach every non-overview claim to message identifiers from that cell (Xia et al., 2026).

Existing studies examine event-oriented social summarization (He et al., 2020), crisis timeline extraction and summarization (Rajaby Faghihi et al., 2022), and attributable crisis facts (Buntain et al., 2023).

For evidence-linked report generation, evidence presentation introduces a trade-off between completeness and concentration. An unfiltered chronological stream retains rare signals alongside duplicated or irrelevant material, increasing the amount of competing information presented to the generator. Finite prompt budgets further limit evidence availability, so even an unfiltered route might exclude messages during prompt packing. Filtering and compression reduce redundant or irrelevant material and concentrate the evidence available for generation (Jiang et al., 2024). However, these operations also risk discarding sparse observations that contribute important information to the report (Tian et al., 2024a; Jin et al., 2026). The value of an evidence view consequently depends on the input and the report produced from that view, motivating selection between complementary outputs. Candidate selection provides a basis for this decision through learned quality signals (Fernandes et al., 2022) or expected utility (Suzgun et al., 2023). Recent document-level candidate selection (Jinnai, 2025) and sentence-level attributed generation (Xia et al., 2025) address complementary aspects of report quality and traceability. Selecting an ex-

*Corresponding author.
[1]Code is released at https://github.com/bjzgcai/R01G-SR.

Proceedings of the First Workshop on Language Technologies for Crisis Preparedness and Response, pages 177–187
November 9, 2026 ©2026 Association for Computational Linguistics

isting report preserves its claim-to-evidence links, whereas synthesizing another report introduces an additional generation and attribution step.

In this paper, we propose R01G-SR, a provenance-preserving approach that generates reports from two complementary evidence views and selects one complete candidate for each input cell. The direct route, R0, uses unfiltered chronological evidence with deterministic prompt packing. The compressed route, R1, applies high-recall filtering and near-duplicate grouping to concentrate the evidence presented to the generator. Both routes use the same generator, prompt, schema, and decoding configuration, so their intended difference lies in evidence preparation. R01G-SR compares the resulting reports using a logistic selector with 22 input- and report-derived features that capture input context and candidate properties. The selector makes its inference decisions without reference text, official scores, crisis identity, or hazard identity. It copies one complete report without merging or rewriting the candidates, preserving the selected candidate's claims and evidence links during selection. A deterministic compliance layer then repairs subsection ownership, sanitizes evidence identifiers, and validates the report schema without generating new claims. We evaluate route selection against both fixed routes under nested crisis-disjoint evaluation and report development and hidden-test results separately. The crisis-disjoint results show aggregate improvements over both fixed routes, while the analysis reveals variation in selection gains across crises, input lengths, and window indices.

Our contributions are summarized as follows.

- We identify and quantify the trade-off between evidence completeness and evidence concentration in structured crisis reporting, and analyze how selection gains vary with input length and window index.
- We propose a provenance-preserving dual-route selection approach that uses reference-free inference features to choose a complete report from complementary evidence views, keeping candidate generation and selection separate from deterministic compliance handling.
- We evaluate route selection against both fixed routes under nested crisis-disjoint evaluation and report development and hidden-test outcomes separately, identifying aggregate improvements and crisis-specific variation in selection gains.

2 Task, Data, and Evaluation

2.1 Task

An LT4CPR *cell* contains crisis metadata, window metadata, and the messages available for one crisis window and replicate. The required output is one JSON situation report per cell, organized into 11 fixed top-level sections and an official subsection vocabulary. Every bullet in Sections 2–11 must cite at least one evidence-message identifier from the same cell. Cells are processed independently: no messages, generated reports, or state are carried across windows, replicates, or crises.

The released training split contains 140 cells from eight crises: blackout, conflict, crowd crush, epidemic, ferry accident, flood, volcano, and wildfire. Development contains 23 cells from two held-out crises, chemical incident and earthquake, and the public test release contains 117 cells from six additional crises. Test references are hidden. The task documentation describes the scenarios as synthetic or fictional and notes that the streams contain deliberately off-topic messages (LT4CPR Shared Task Organizers, 2026a).

2.2 Released evaluation

The released evaluator scores Sections 3–11, aligning system and reference subsections by header and computing text similarity at the subsection level. The official PRIMARY score for an instance is

$$\text{PRIMARY} = \frac{\text{BERTScore-F1} + \text{BLEURT}}{2}, \quad (1)$$

where BERTScore (Zhang et al., 2020) and BLEURT (Sellam et al., 2020) contribute equally. The final system score is the macro-average across scored instances. We use ROUGE (Lin, 2004), evidence alignment, and structure recall as complementary diagnostics when the corresponding artifacts are available, while system comparisons are based on PRIMARY.

3 Related Work

Crisis-stream summarization organizes useful information from noisy social-media messages (Rudra et al., 2016), while broader microblog summarization models conversational discourse and latent topics (Li et al., 2018). Representative resources cover

event-oriented summarization through TWEETSUM (He et al., 2020), crisis timeline extraction and summarization through CrisisLTLSum (Rajaby Faghihi et al., 2022), and attributable facts through CrisisFACTS (Buntain et al., 2023). Broader summarization methods select salient sentences (Chen et al., 2018; Luo et al., 2019) and important medical utterances (Song et al., 2020). LT4CPR couples situation report generation with a fixed schema and explicit evidence identifiers within each input cell (Xia et al., 2026), connecting crisis reporting to attributed and structured generation. ReClaim interleaves reference and claim generation to provide sentence-level citations (Xia et al., 2025). ALCE evaluates answer generation with supporting citations (Gao et al., 2023b), while RARR retrieves evidence and revises unsupported content in existing outputs (Gao et al., 2023a). Attribute First, then Generate selects source segments to guide generation and provide fine-grained attribution (Slobodkin et al., 2024). Grammar-constrained decoding addresses structural validity by enforcing outputs that follow a prescribed grammar (Geng et al., 2023). Related generation methods use expert routing and fusion for dialogue summarization (Tian et al., 2024b), learned quality signals for candidate reranking (Fernandes et al., 2022), or expected utility under minimum Bayes risk decoding (Suzgun et al., 2023). Document-level MBR-OT adapts candidate selection using optimal transport over sentence-level utilities (Jinnai, 2025). Unlike MBR, our selector predicts a route from observable input and candidate features rather than computing expected pairwise utility. For evaluation, BERTScore (Zhang et al., 2020) and BLEURT (Sellam et al., 2020) jointly define the official LT4CPR PRIMARY score, while AlignScore assesses factual consistency through information alignment (Zha et al., 2023). GroUSE examines failure-mode detection by evaluators for grounded question answering (Muller et al., 2025), motivating a distinction between aggregate quality scores and evidence-support diagnostics. Compared with these studies, R01G-SR derives candidate diversity from unfiltered and filtered, near-duplicate-compressed evidence views under the same deterministic generator, and selects one complete report using reference-free inference features. Whole-report selection preserves the chosen candidate's claims and evidence links without post-selection synthesis, while a separate deterministic compliance layer handles structural and evidence-ID requirements.

4 Method

Figure 1 separates evidence preparation, learned whole-report selection, and deterministic compliance checks.

4.1 Problem formulation

For an LT4CPR cell, let E denote the available within-cell evidence and let $V_0(E)$ and $V_1(E)$ denote two evidence views. A fixed generator G produces

$$r_i = G(V_i(E)), \qquad i \in \{0, 1\}. \tag{2}$$

If the reference were available, an oracle could choose the candidate with the larger task score $\mathcal{Q}(r_i)$. At inference time $\mathcal{Q}$ is unavailable, so we learn a selector

$$s(x(E, r_0, r_1)) \in \{0, 1\}, \tag{3}$$

where x contains only observable input and report properties. The selected report is copied intact and then passed through a deterministic compliance function C, giving $\hat{r} = C(r_s)$. This formulation makes the design goal explicit: exploit differences between evidence views without introducing a post-selection generation step that could alter claims or provenance.

4.2 Shared generator

Both routes use Qwen3-4B-Instruct-2507[2] (Yang et al., 2025), quantized as UD-Q4_K_XL GGUF and served locally with `llama.cpp`. Decoding uses temperature 0 and a 6,144-token output cap. The model context allocation and the evidence-packing budget are distinct constraints: a request fitting within the model context does not imply that every source message was included in that request. The direct-route packer limits serialized evidence lines to 85,000 characters, excluding prompt instructions and metadata, and removes middle lines when this budget is exceeded. Inference uses a single NVIDIA A100-SXM4-80GB GPU.

The system prompt restricts the model to the supplied cell evidence and requires a single JSON object. The user prompt specifies crisis and window metadata, the legal section and subsection inventory, confidence labels, evidence-ID rules, and

[2]Official model card: https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507.

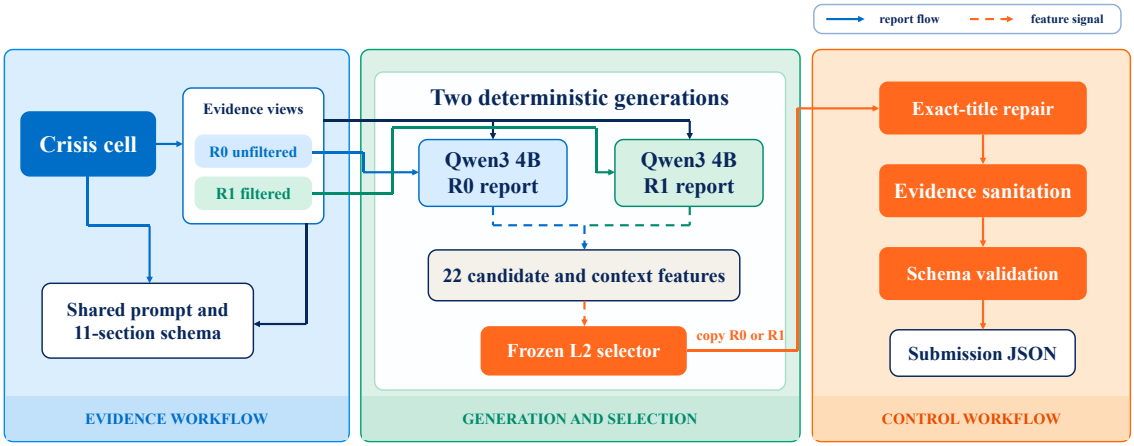


Figure 1: The R01G-SR method. Two complementary evidence views feed independent calls to the same local generator. A frozen selector uses reference-free inference features to choose one complete report without an additional rewrite. Exact-title ownership repair, evidence sanitation, and schema validation form a deterministic compliance layer and do not generate new claims.

instructions for concise atomic claims. Reports are generated automatically from the supplied within-cell evidence. A deterministic normalizer removes illegal or empty structures, merges duplicate subsection titles, and renumbers subsection and bullet identifiers.

4.3 Complementary evidence views

R0: unfiltered chronological evidence. R0 constructs an evidence line for every message in the cell's supplied chronological stream, without relevance filtering or near-duplicate grouping. Each line contains its identifier, timestamp, source label, and text. Before generation, a deterministic packer repeatedly removes the middle line until the sum of line lengths is at most 85,000 characters, or only two lines remain. This preserves the beginning and end of a long stream but can exclude intervening evidence. R0 therefore denotes the unfiltered direct route, not a guarantee that every original message is visible to the model. Both routes still use the same generator, report instructions, and decoding configuration; their evidence preparation differs.

R1: filtered and compressed evidence. R1 favors evidence concentration. It first normalizes retweet markers, URLs, mentions, hashtags, case, and punctuation for matching. A message is retained if it has an annotated source type or overlaps both the crisis metadata and a fixed high-recall list of crisis terms. Retained messages are grouped greedily when the Jaccard similarity of their normalized content tokens is at least 0.70. A group representative is replaced only by a member with a higher fixed source rank, and the compressed line preserves every member identifier. Compression therefore changes the text presented to the model without fabricating evidence identifiers.

The evidence views trade off chronological coverage against reduced redundancy and irrelevant competition.

4.4 Provenance-preserving route selection

The selector combines signals for coverage, evidence support, redundancy, attribution risk, and input context. For each candidate report, we compute five report-level features: the number of scored bullets, the number of represented subsections, mean lexical evidence support, duplicate-bullet count, and the number of numerical bullets containing values absent from their cited evidence. Let $f(r) \in \mathbb{R}^5$ denote these features, $q(r)$ a deterministic candidate-quality score, and $z \in \mathbb{R}^4$ the log input size, relevance ratio, compression ratio, and window index. The selector input is

$$x = [f(r_0), f(r_1), f(r_1) - f(r_0), \\ q(r_0), q(r_1), q(r_1) - q(r_0), z], \tag{4}$$

for a total of 22 dimensions. Section 4.5 defines the features and fitting weights. At inference time, the selector never receives crisis identity, hazard identity, reference text, or official scores.

Training labels record whether R1 attains a higher PRIMARY score than R0; exact ties are omitted from classifier fitting. We standard-

ize the features and fit an L2-regularized logistic regression. Sample weights reflect the crisis/window/replicate hierarchy and observed candidate regret, with regret floored at 0.005. We choose $C \in \{0.03, 0.1, 0.3, 1, 3\}$ and a decision threshold from $\{0.4, 0.5, 0.6\}$ inside the inner loop of nested leave-one-crisis-out (LOCO) validation. The final selector is fitted on all eight training crises with $C = 0.3$ and threshold 0.5. It chooses R1 for 16 development cells and R0 for seven.

Whole-report selection preserves one candidate's evidence trail without another language-model call. It does not itself verify entailment.

4.5 Feature definitions and selector fitting

The five report features are computed over Sections 3–11. For a bullet, lexical support is the maximum Jaccard overlap between its content-token set and that of any valid cited message; it is zero when no valid citation exists. Report-level support is the mean over scored bullets, with zero used for an empty report. A bullet contributes to the duplicate count if its token-set overlap with any earlier scored bullet exceeds 0.75. A numerical-risk flag is set if at least one extracted number in the bullet is absent from the concatenated cited evidence. These are lexical diagnostics, not entailment judgments. Represented subsections are counted by distinct titles in the scored sections.

Let B, H, L, D, and U denote the scored-bullet count, distinct subsection-title count, mean lexical support, duplicate-bullet count, and numerical-risk count, respectively. The deterministic quality feature is

$$q(r) = 0.30\frac{\min(B, 50)}{50} + 0.25\frac{\min(H, 12)}{12} + 0.45L - 0.03D - 0.08U. \quad (5)$$

For N input messages, N_{rel} retained messages, and N_{grp} compressed groups, the context features are $\log(1 + N)$, $N_{\text{rel}}/\max(1, N)$, $N_{\text{grp}}/\max(1, N)$, and the numeric window index. The stored feature vector interleaves each report feature as $(f_j(r_0), f_j(r_1), f_j(r_1) - f_j(r_0))$, followed by the two quality scores, their difference, and the four context features.

Within each fitting fold, exact-score ties are excluded. For a remaining cell i, the unnormalized fitting weight is

$$w_i = \frac{\max(|\Delta_i|, 0.005)}{|\mathcal{C}|\,|\mathcal{W}_{c(i)}|\,n_{c(i),w(i)}}, \quad (6)$$

where Δ_i is the R1-minus-R0 training score, $\mathcal{C}$ contains represented training crises, $\mathcal{W}_{c(i)}$ contains represented windows within its crisis, and $n_{c(i),w(i)}$ counts retained non-tie cells in that crisis-window group. We rescale these weights to mean one. Feature standardization is fitted without sample weights using only the fitting cells; logistic regression then uses the normalized weights. Inner and outer held-out crises are excluded from fitting the scaler and classifier for their respective predictions.

4.6 Deterministic compliance layer

The final layer addresses mechanical submission constraints rather than report quality. From the official subsection vocabulary, we build a title-to-owner map. A subsection is moved only when its title has exactly one legal owner and it appears under a different top-level section. If the destination already contains the same title, the subsections are merged and identifiers are renumbered. Claim text, confidence labels, and evidence content are never rewritten.

Evidence sanitation removes identifiers that do not occur in the cell and drops non-overview bullets left without a valid identifier. Strict schema, coverage, and evidence-ID validation follows. We use R01G-SR to refer to the selected report after these deterministic checks.

5 Experimental Setup

5.1 Research questions

We address three research questions:

RQ1: Does evidence presentation matter, and can direct and compressed views provide complementary report quality under a fixed generator?

RQ2: Can a selector trained without the target crisis exploit that complementarity on unseen crises without reference access at inference time?

RQ3: How do selection gains and remaining regret vary with input size and crisis window?

5.2 Crisis-disjoint selector evaluation

We treat the eight training crises as independent groups. In each outer LOCO fold, all decisions for the held-out crisis are made by a selector whose fitting and inner-loop hyperparameter selection exclude that crisis. We report both the original candidates and a compliance-matched analysis in which

System	PRIMARY	BERT-F1	BLEURT
R01G-SR	**0.524834**	**0.651119**	0.398548
R1-fresh	0.522617	0.645245	**0.399988**
R0	0.515040	0.638329	0.391751

Table 1: Results on all 23 development cells under the current released evaluator. R1-fresh is a regenerated compressed-route comparison.

Method	PRIMARY	Δ vs. R0c	Δ vs. R1c
R0c	0.352297	0.000000	+0.002543
R1c	0.349754	-0.002543	0.000000
R01G-SR	0.363869	**+0.011571**	**+0.014115**
Oracle	0.376693	+0.024396	+0.026939

Table 2: Compliance-matched results on 140 training cells with identical sanitation and hierarchical weights. Both always-R0c and always-R1c are explicit baselines. Selection uses frozen nested-LOCO decisions; oracle uses reference scores. Selector crisis-level wins are 8/8 against R0c and 5/8 against R1c. Deltas use unrounded scores.

both routes receive identical sanitation. A two-route oracle that picks the higher-scoring candidate using reference scores is included only to estimate unattainable headroom.

For this training analysis, scores are averaged first over replicates within each crisis window, then over windows within each crisis, and finally over crises. This hierarchical aggregation differs from the equal-cell averaging used for the current development comparison. The fixed-route baselines and the oracle are recomputed over the same cells with the same weights; only the learned selector requires cross-fitting.

5.3 Development and hidden-test evaluation

After freezing the selector, we evaluate the three development systems on the same 23 cells with the current evaluator. Development scores are never used to alter features, coefficients, or the decision threshold. Development contains two crises. The official hidden test contains 117 cells from six additional crises; its references remain unavailable, so test results are used only for system-level evaluation.

6 Results

6.1 RQ1: Evidence-view complementarity on development

On development, the compressed-route comparison is stronger than the direct route in aggregate. Table 1 shows that R1-fresh reaches 0.522617 PRIMARY compared with 0.515040 for R0. The selector chooses R0 for seven of the 23 cells and R1 for sixteen, and R01G-SR reaches 0.524834. Its advantage over R1-fresh is only +0.002217 (approximately 0.42% relative). Moreover, R1-fresh is a regenerated comparison rather than a verified identical copy of the selector's R1 candidate; the controlled fixed-route test is the crisis-disjoint analysis below.

Relative to R1-fresh, R01G-SR increases BERTScore-F1 by 0.005874 but decreases BLEURT by 0.001440, yielding a net PRIMARY gain of 0.002217. Relative to R0, both components improve. The comparison with the stronger fixed route therefore reflects a trade-off between component metrics, not a uniform improvement. With only two development crises, we treat the aggregate margin as descriptive rather than statistically established superiority.

6.2 RQ2: Crisis-disjoint route selection

R01G-SR reaches 0.363869, exceeding R0c by 0.011571 and R1c by 0.014115. Crisis-level gains occur on 8/8 crises against R0c and 5/8 against R1c. Always choosing R1c therefore does not explain the aggregate benefit, although selection does not beat it on every crisis. Section 6.3 reports each crisis, including losses against R1c on epidemic, volcano, and wildfire. The selector recovers approximately 47% of the oracle improvement over R0c.

6.3 Fixed-route comparisons by crisis

Table 3 reports the 140-cell compliance-matched scores using the same hierarchical weights as Table 2. Each row averages replicates within windows and then windows within one held-out crisis. The macro row averages the eight crises equally. It reveals both the positive aggregate margin and the three crises where always-R1c performs better than the selector.

The crisis-level comparison clarifies the scope of the aggregate improvement. Against the compressed route, the largest positive differences occur for conflict, ferry, and flood, whereas epidemic, volcano, and wildfire favor the fixed compressed route. These mixed outcomes rule out a uniform per-crisis advantage.

Crisis	Cells	R0c	R1c	R01G-SR	Δ vs. R0c	Δ vs. R1c
Blackout	16	0.370619	0.374033	0.387532	+0.016913	+0.013499
Conflict	13	0.358526	0.309418	0.359698	+0.001172	+0.050280
Crowd crush	22	0.366667	0.373663	0.392431	+0.025764	+0.018768
Epidemic	16	0.374077	0.382945	0.375037	+0.000960	-0.007908
Ferry	25	0.352971	0.316636	0.353631	+0.000660	+0.036996
Flood	10	0.336510	0.315041	0.349410	+0.012900	+0.034369
Volcano	16	0.296942	0.345246	0.318324	+0.021382	-0.026923
Wildfire	22	0.362064	0.381050	0.374884	+0.012820	-0.006166
Macro	140	0.352297	0.349754	0.363869	+0.011571	+0.014115

Table 3: Fixed-route comparisons for each held-out training crisis. All values are PRIMARY; deltas are computed before rounding. The selector's aggregate improvement over R1c is not uniform across crises.

System	BERT-F1	BLEURT	Align-F1	PRIMARY
Organizer baseline	0.463445	0.475283	0.208240	0.469364
R1 fallback	0.483191	0.541305	0.278185	0.512248
R01G-SR	**0.501837**	**0.552800**	**0.288103**	**0.527319**

Table 4: Official hidden-test scores for our two submitted systems and the organizer baseline. Align-F1 is Weighted Alignment F1. The organizer baseline is excluded from participant rankings; PRIMARY is computed from BERTScore-F1 and BLEURT.

6.4 Official hidden-test evaluation

The official evaluation uses a hidden-reference test split of 117 cells from six crises. We compare the designated R01G-SR primary with the submitted R1 fallback. The official results workbook reports the scores in Table 4 (LT4CPR Shared Task Organizers, 2026b).

Relative to the submitted R1 fallback, R01G-SR improves PRIMARY by 0.015071. The improvement is not confined to the ranking metric: BERTScore-F1 rises by 0.018646, BLEURT by 0.011495, and Weighted Alignment F1 by 0.009918. The official workbook also shows gains of 0.018577, 0.018827, and 0.020234 in ROUGE-1, ROUGE-2, and ROUGE-L F1, respectively. Thus, the primary submission exceeds the fallback on every reported metric. It also exceeds the organizer baseline by 0.057955 PRIMARY.

R01G-SR ranks sixth of nine teams in both the Primary and Best System Rankings. The test PRIMARY score is numerically close to development (0.5273 versus 0.524834). However, BERTScore-F1 changes from 0.651119 to 0.5018, while BLEURT changes from 0.398548 to 0.5528. Their offsetting movements mean that aggregate similarity should not be interpreted as metric-level stability or as evidence that the same mechanisms generalize across splits. The official comparison with the submitted R1 fallback provides the more direct evidence of system-level benefit on test.

7 Selection Gains by Input Size and Window

The main results establish that route selection can improve report quality. We next examine how selection gains and remaining regret vary as evidence accumulates.

Stratifying the nested-OOF decisions by input size shows descriptive differences across evidence-volume groups. Relative to the compliance-matched direct route, selection changes PRIMARY by -0.0022 for cells with 22–125 messages, +0.0052 for 128–357 messages, and +0.0233 for 362–993 messages. Across windows, the corresponding deltas are -0.0016, +0.0048, +0.0157, and +0.0274 from W1 through W4. These are differences between selected reports and R0c, not isolated compression effects.

Input length, window index, and crisis characteristics vary together, so these subgroup differences are descriptive rather than causal.

The oracle places 82.2% of remaining weighted regret in W1/W2. Regret is the better existing candidate's score minus the selected score, so this identifies missed routing opportunities, not proof that candidate quality dominates. The training gap is 0.012824; the development oracle reaches 0.530747, leaving 0.005913. Future work should separate routing errors from cases where both reports omit or misinterpret evidence.

8 Discussion

Value beyond a fixed evidence view. Under identical sanitation and hierarchical weighting, always-R0c slightly exceeds always-R1c, yet the selector exceeds both (Table 2). Its benefit therefore comes

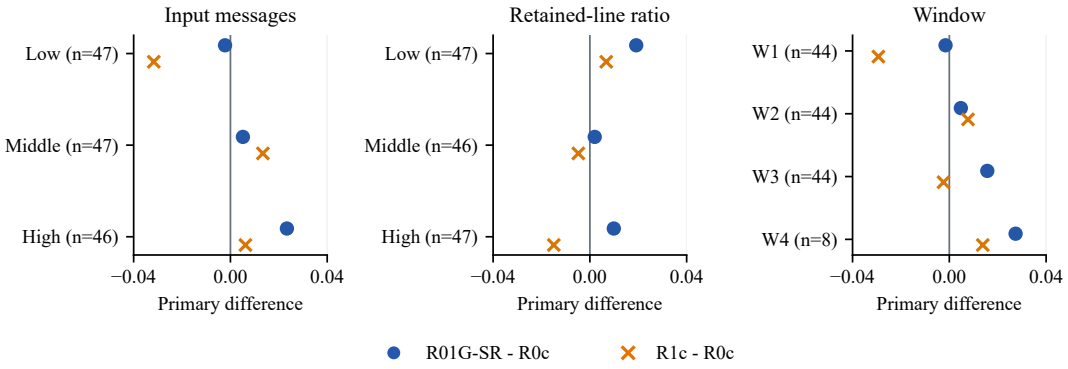


Figure 2: Compliance-matched nested-OOF differences by input-message tercile, retained-line ratio, and crisis window. Retained-line ratio denotes compressed evidence lines per input message, corresponding to $N_{\mathrm{grp}}/\max(1, N)$. Points are descriptive, not causal. W4 has eight cells versus 44 in each earlier window; the ratio trend is non-monotonic.

from conditional decisions rather than a general preference for compression. The selector recovers approximately 47% of the two-route oracle's improvement over R0c. Candidate coverage, lexical support, redundancy, and numerical-risk features describe complementary report properties, while context features describe evidence volume and concentration. We evaluate these signals jointly and make no single-feature attribution. Development, nested-LOCO, and official-test results address different evaluation settings, with candidate relationships and aggregation rules specified in the experimental setup and results.

Crisis-specific differences. For conflict and ferry, the selector stays close to R0c while exceeding R1c; for crowd crush and flood, it exceeds both fixed routes. Epidemic, volcano, and wildfire favor R1c over the selector, showing remaining opportunities for better routing. These crisis-level averages summarize cell-specific decisions made by held-out-crisis selectors. The early-window concentration of regret identifies where routing can improve among the existing candidates, whereas facts omitted by both candidates require improvements to evidence presentation or generation. Packing affects evidence availability, while claim support requires checking the relationship between each claim and its cited messages.

Generation cost. The method requires two generation calls per cell, with cost influenced by the input and output lengths of both routes. Sharing loaded weights does not eliminate the additional decoding workload. Whole-report copying avoids a further synthesis call and preserves the selected candidate's claim–evidence associations before compliance checks. The practical choice is therefore whether the observed quality gain warrants the additional generation budget. A future cost-aware policy could request the second candidate selectively. No latency or energy measurements are available for assessing that trade-off quantitatively.

9 Conclusion

In this paper, we propose R01G-SR to exploit complementary evidence views for evidence-linked crisis situation report generation. The approach generates candidates under the same generator and uses reference-free features to select one complete report without rewriting its claims. Nested crisis-disjoint evaluation shows aggregate gains over both fixed routes, and official hidden-test results show improvements over the submitted compressed-route fallback. Future work should distinguish routing errors from evidence omissions or misinterpretations shared by both candidates.

Ethical Considerations

Crisis streams may contain sensitive information, rumors, distressing accounts, and references to vulnerable people. Our system processes released cells locally and performs no external identity resolution, web enrichment, contact, or profiling. Evidence identifiers improve traceability but do not establish source accuracy; automated reports may still repeat rumors, omit less visible experiences, or overstate uncertain claims. The outputs are research artifacts rather than operational intelligence. Deployment would require source governance, access controls, human review, correction workflows, and evaluation with crisis professionals and affected communities.

Limitations

The nested-LOCO evaluation covers eight training crises, and development covers only two crises. Hidden test references prevent crisis-level oracle and evidence-support analyses on the official test split. The regenerated R1-fresh development comparison is separate from the selector's R1 candidate and the submitted fallback. Automatic similarity metrics do not certify factual entailment, operational usefulness, timeliness, or safety; lexical support and numerical-risk features are shallow attribution proxies. The repair layer handles exact titles with a unique legal owner and leaves ambiguous cases untouched. Quantized local inference also limits conclusions about other model precisions and serving stacks.

Acknowledgments

This work is supported by Zhongguancun Academy (Grant No. XTS0043).

Generative AI tools assisted with manuscript polishing. The authors reviewed all revisions and take full responsibility for the content of this paper.

References

Cody Buntain, Benjamin D. Horne, Amanda Hughes, Muhammad Imran, Richard McCreadie, and Hemant Purohit. 2023. CrisisFACTS 2023 – Overview Paper. In *The Thirty-Second Text REtrieval Conference Proceedings (TREC 2023)*, volume 1328 of *NIST Special Publication*. National Institute of Standards and Technology.

Xiuying Chen, Shen Gao, Chongyang Tao, Yan Song, Dongyan Zhao, and Rui Yan. 2018. Iterative document representation learning towards summarization with polishing. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4088–4097.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023a. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. 2023. Grammar-constrained decoding for structured NLP tasks without finetuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10932–10952, Singapore. Association for Computational Linguistics.

Ruifang He, Liangliang Zhao, and Huanyu Liu. 2020. TWEETSUM: Event oriented social summarization dataset. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5731–5736. International Committee on Computational Linguistics.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2024. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1658–1677, Bangkok, Thailand. Association for Computational Linguistics.

Yiqiao Jin, Kartik Sharma, Vineeth Rakesh, Yingtong Dou, Menghai Pan, Mahashweta Das, and Srijan Kumar. 2026. SARA: Selective and adaptive retrieval-augmented generation with context compression. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14508–14528, San Diego, California, United States. Association for Computational Linguistics.

Yuu Jinnai. 2025. Document-level text generation with minimum Bayes risk decoding using optimal transport. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28260–28279, Vienna, Austria. Association for Computational Linguistics.

William D. Lewis, Haotian Zhu, Keaton Strawn, and Fei Xia. 2025. Tapping into social media in crisis: A survey. In *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, pages 306–331, Vienna, Austria. Association for Computational Linguistics.

Jing Li, Yan Song, Zhongyu Wei, and Kam-Fai Wong. 2018. A joint model of conversational discourse and latent topics on microblogs. *Computational Linguistics*, 44(4):719–754.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.

LT4CPR Shared Task Organizers. 2026a. LT4CPR 2026 AACL Workshop Shared-Task Repository. Released task data, schema, validator, and evaluation code. Test-release snapshot: commit 78dbff586d391d97ec4d11310cf587f43c57ca2a.

LT4CPR Shared Task Organizers. 2026b. LT4CPR Shared Task Official Evaluation Results. Official results workbook. Accessed 2026-09-16.

Ling Luo, Xiang Ao, Yan Song, Feiyang Pan, Min Yang, and Qing He. 2019. Reading like HER: Human reading inspired extractive summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3033–3043, Hong Kong, China. Association for Computational Linguistics.

Sacha Muller, Antonio Loison, Bilel Omrani, and Gautier Viaud. 2025. GroUSE: A benchmark to evaluate evaluators in grounded question answering. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4510–4534, Abu Dhabi, UAE. Association for Computational Linguistics.

Yuyang Nie, Yuanhe Tian, Xiang Wan, Yan Song, and Bo Dai. 2020. Named entity recognition for social media texts with semantic augmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1383–1391.

Hossein Rajaby Faghihi, Bashar Alhafni, Ke Zhang, Shihao Ran, Joel Tetreault, and Alejandro Jaimes. 2022. CrisisLTLSum: A benchmark for local crisis event timeline extraction and summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5455–5477, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Koustav Rudra, Siddhartha Banerjee, Niloy Ganguly, Pawan Goyal, Muhammad Imran, and Prasenjit Mitra. 2016. Summarizing situational and topical information during crises. In *4th International Workshop on Social Web for Disaster Management (SWDM 2016)*, Indianapolis, USA. Co-located with CIKM 2016. Available as arXiv:1610.01561.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892. Association for Computational Linguistics.

Aviv Slobodkin, Eran Hirsch, Arie Cattan, Tal Schuster, and Ido Dagan. 2024. Attribute first, then generate: Locally-attributable grounded text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3344, Bangkok, Thailand. Association for Computational Linguistics.

Yan Song, Yuanhe Tian, Nan Wang, and Fei Xia. 2020. Summarizing medical conversations via identifying important utterances. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 717–729.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2023. Follow the wisdom of the crowd: Effective text generation via minimum Bayes risk decoding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4265–4293, Toronto, Canada. Association for Computational Linguistics.

Yuanhe Tian, Chang Liu, Yan Song, Fei Xia, and Yongdong Zhang. 2024a. Aspect-based sentiment analysis with context denoising. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3083–3095, Mexico City, Mexico. Association for Computational Linguistics.

Yuanhe Tian, Fei Xia, and Yan Song. 2024b. Dialogue summarization with mixture of experts based on large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7143–7155, Bangkok, Thailand. Association for Computational Linguistics.

Fei Xia, Haotian Zhu, Yoo Wun Lee, William D. Lewis, Antonis Anastasopoulos, and Belu Ticona. 2026. Overview of the LT4CPR 2026 shared task: Generating situation reports from social media messages. In *Proceedings of the First Workshop on Language Technologies for Crisis Preparedness and Response (LT4CPR)*, Hengqin, Zhuhai, China. In conjunction with AACL 2026, November 9, 2026.

Sirui Xia, Xintao Wang, Jiaqing Liang, Yifei Zhang, Weikang Zhou, Jiaji Deng, Fei Yu, and Yanghua Xiao.

2025. Ground every sentence: Improving retrieval-augmented LLMs with interleaved reference-claim generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 969–988, Albuquerque, New Mexico. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Yingyi Zhang, Jing Li, Yan Song, and Chengzhi Zhang. 2018. Encoding conversation context for neural keyphrase extraction from microblog posts. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1676–1686.

A Implementation and Resource Disclosure

Item	Setting
Primary generator	Qwen3-4B-Instruct-2507, UD-Q4_K_XL GGUF
Inference engine	`llama.cpp`, local execution
Principal decoding	temperature 0; 6,144 output-token cap
Hardware	one NVIDIA A100-SXM4-80GB
Selector	standardized L2 logistic regression; $C = 0.3$; threshold 0.5
Training data	released LT4CPR training split only

Table 5: Disclosure for the submitted primary system.

B Code Availability

The public repository provides core generation, feature extraction, selector training and application, and deterministic compliance code. Frozen selector coefficients, task data, generated reports, and model weights are not included. Retraining requires access to permitted task data and the corresponding training-candidate scores.