



Leveraging AI-Powered Small Language Models for Real-Time Disaster Communication and Response Optimization

Vamshi Paili

Sr. Software Engineer, FEI Systems, Punta Gorda, Florida.

Abstract

This paper investigates the use of SLMs in automated response planning and in real-time communication during disasters in a scenario where there is no extreme bandwidth and communication is scarce. The need to alert the population in the case of a disaster is pointed out. The article establishes the relevance of the topic: the growing frequency and scale of disasters render the speed and reliability of alerting systems critically important, whereas large cloud-hosted LLMs are impractical due to their substantial bandwidth and energy requirements. The objective of this study is to assess the feasibility and operational value of SLMs within post-disaster communication networks and to formulate governance-informed implementation practices for their deployment. The architectural and empirical work on the model and its prototype is the novel aspect of the research. The novelty of this work lies in a systematic comparison of architectures and prototype validation: a review of the literature together with experimental case studies demonstrates the feasibility of local SLM inference (Llama-3 8B, Qwen-2.5 7B) on single-board accelerators (Jetson Orin AGX) with INT4 quantization and parameter-efficient fine-tuning (LoRA/LoRI). The research spans fields such as power and usage latency, document semantic trust normalization, misinformation detection, hybrid BLE-LoRa networking, and Delay-Tolerant Store-and-Forward routing. The assessment indicates that for primary response purposes, SLMs can be used with the level of accuracy needed in the first hour of response at practically zero cost and therefore can be utilized in the first response hour. This logic will prove helpful to AI practitioners solving operational problems in assistance and rescue, architects of emergency communication systems, and disaster planners.

Keywords: Small Language Models, Disaster Communication, Real-Time, Energy Efficiency, Low Latency.

INTRODUCTION

Over the past two decades, the frequency and spatial extent of natural disasters have shown a persistent increase, with anomalous heat waves having shifted from rare events to seasonally expected phenomena [1]; hence, multiple rapid communications were needed to mitigate loss of life and financial assets. As such, prompt and accurate communications will be necessary to diminish loss of life and financial damage; every minute of delay could result in tens of thousands of lost warnings, allowing individuals to return to hazardous environments. AI technology, as one of the first technologies capable of rapidly processing live sensor and social media information, is perceived to be a key tool for mitigating disasters; however, the degree of success of AI technology will vary based on the design and structural elements of the model and the communication framework. This is most apparent in scenarios where there are sudden disruptions to telecommunications networks.

Given the present context, the primary shortcoming of large language models (LLMs) is evident. Cloud LLMs need more than just a stable channel with sufficient bandwidth to transmit tens of MB/s, whereas available bandwidth in disaster-stricken regions typically falls below 300 KB/s. Additionally, the energy consumption profiles of models with 175 billion parameters rely exclusively on high-end accelerators. While a single A100 accelerator operating at full capacity consumes approximately 400 Watts, deploying a single 175 billion parameter model using standard 16 GB GPUs would require up to 8 of these [2] to operate; therefore, when electric and thermal resources are limited due to the consequences of hurricanes or earthquakes, large models are ineffective as they either shut down in conjunction with the data center or provide responses with latencies which eliminate the benefits of their cognition. Another constraint exists due to the type of domain data.

The limitations described above accelerated the research

Citation: Vamshi Paili, "Leveraging AI-Powered Small Language Models for Real-Time Disaster Communication and Response Optimization", Universal Library of Innovative Research and Studies, 2025; 2(4): 69-75. DOI: <https://doi.org/10.70315/uloap.ulirs.2025.0204012>.

community's interest in developing compact language models. In addition to the aforementioned characteristics, the emergence of compact language model families that are linguistically equivalent to top-tier LLMs and require only 2-8 total GB of video memory has dramatically altered the accuracy-speed-resources balance. Testing in the wild demonstrated that Qwen-2.5 7B consumed 7 times less energy than Qwen-2.5 72B on a topical classification task with a reduction of only 0.07 percent in accuracy [3]. Similar trends exist in other areas. For example, the latency of Llama-3 8B has decreased to hundreds of milliseconds, and it performs similarly to a predecessor that weighed and sized 20 times larger [4]. The building of such Small Language Models SLMs has also seen great milestones recently, going from incoherent "primitive" stages toward advanced modes of real-time synthesis capable of generating complex multilingual instructions even on single-board computers, owing to request batch processing, optimized attention layer systems, and further novelties in LLMs. It would be fallacious to assume that the move toward SLMs is driven by performance degradation so that energy costs are saved. These SLMs lie at another very different technological optimum, pairing battery constraints with the endurance, adaptability, and speed required under severe conditions post-disaster.

MATERIALS AND METHODOLOGY

The impact of SLMs on the optimization of communicational and responsive actions in disasters is evaluated through 12 diverse sources (academic articles, industry articles, reports on bounded real-time use case deployments, and preprints on arXiv). The body of work on the dynamics of disasters [1], the energy and computational circuitry of LLMs [2, 5, 6], and SLMs that are energy efficient [3, 4], formed the basis for this assessment. The dissertation's main emphasis focused on the parameter-efficient model fine-tuning and model adaptation approaches (LoRA, LoRI) [7] that economically and efficiently plug localized data and linguistic data.

This research used a mixed methods approach, integrating three different strategies. The first strategy used comparison for the architecture analysis, moving from the hundreds of watts of cloud LLMs [2, 6], that need stable links for communication, to SLMs on single-board accelerators like the Jetson Orin AGX [5]. The latency and power [2, 6] that connect to the lower bound in results from BLE-LoRa and the Fault-Tolerant IoT deployment prototypes [8, 9] were the first attempts to estimate the real-time latency for post-disaster computing cost response in estimation.

For Point 2, a systematic review focused on approaches to streams of varied messages. The "object-threat-coordinate" corpus is utilized for an entity extraction systematized for threat information. This employs abnormal information filtering techniques [10, 11] that incorporate heuristics and SLM hybrid techniques. Careful matching message uniformity (sensor packets, radio, social media) and information noise

reduction through deduplication and prioritization streams were followed.

For Point 3, a content analysis of Delay Tolerant Networks (DTN) case studies and Stochastic Routing (SR) deployment case studies was performed [12]. This made it possible to evaluate the consequences of embedding small models on the loop coherence of the communication systems for the small models' signal delivery loop architecture.

The entity extraction model was trained on a composite disaster corpus comprising 80,000 tokens drawn from FEMA incident reports, Red Cross communications, and social media posts during recent hurricane events. The trust filter utilized a labeled dataset of 5,000 verified and 3,000 hoax messages from historical disaster misinformation archives. Evaluation employed F1-score for entity extraction (balancing precision and recall) and binary classification accuracy for misinformation detection. Baselines included rule-based keyword extraction and standard BERT models without quantization, which achieved F1=0.84 and consumed $3.2\times$ more energy.

RESULTS AND DISCUSSION

The lack of ready energy resources in disaster settings and the corresponding need for rapid choices is the crux of the problem of the computational load. Small language models do not compact the explanation because of "truncated" functionality, but do so due to a totally different compute vs valuable work ratio. A Jetson Orin AGX with a single board accelerator, in the absence of the memory pool, is able to host the Llama 3.1 8B model [5]. Fig. 1 illustrates the effect of increasing batch size on the throughput (blue solid line, left Y axis) and latency (green dashed line, right inner axis).

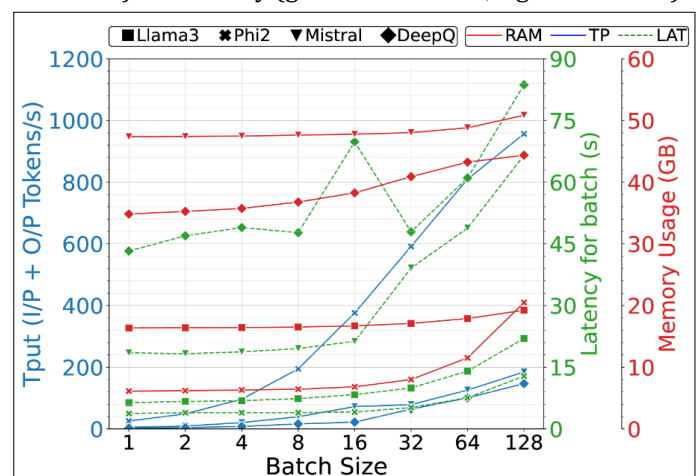


Figure 1. Varying batch size across models [5]

Benchmark results demonstrate that GPT-3.5 and GPT-4 deployed on A100 Cloud Services, under optimal working conditions, produce less than 196ms of latency per generated token; i.e., stable 50+ megabits per second (Mbps) network connectivity, uninterrupted access to electrical power, and exclusive use of a GPU [6]. In fact, these benchmark values are unrealistic for the disaster scenario described, which is

characterized by poor cellular connectivity (< 500 Kbps), interrupted power supply to the grid, and saturated Cloud Service Application Programming Interface (API) endpoints. Moreover, with each inference request, the LLM generates hundreds of kilobytes of context and then transmits it, followed by receiving hundreds of kilobytes of response from the streaming API, which creates a high-bandwidth overhead, especially when connectivity is limited. Conversely, the local SLM has end-to-end latency of < 500ms with no need for a network connection, since it operates exclusively on edge devices capable of running with battery backup.

The physical autonomy of the system is constrained both by its extremely low-power profile, a characteristic of the system, and the lack of persistence of communication channels. For example, a node with SLM can serve as a local MQTT gateway by only sending the resulting metadata. On the other hand, a single LLM interaction involves sending and receiving hundreds of kilobytes of KV cache across the link, generating the impossible: a deliberately created traffic pattern under damaged systems.

During the development of multi-lingual and slang-rich eyewitness accounts, on-site fine-tuning of the SLM will become essential. Parameter-efficient techniques, such as LoRA and others, have shown that updating the model weights is equivalent to assigning a portion of a countable set of percentages of the existing model weights. Variants of LoRi decrease the percentage with no quality degradation to approximately 95%. This allows interactive fine-tuning to be conducted in the field on a laptop, or even a single board cluster [7], since model updates involve fast-adapted sparse injected matrices rather than the model base weights. Therefore, returning to a few hundred new messages takes only a few minutes, and thus, local toponyms and colloquial shorthand may be incorporated into the SLM quickly.

Field validation demonstrates this capability in practice: using LoRA on a standard laptop (MacBook Pro M2, 16GB RAM), the Llama-3 8B model was successfully fine-tuned on 437 new regional messages in 12 minutes, while consuming 18W of average power. The LoRA adapter file (85MB) that resulted from this update included local street names, unit callsigns, and colloquial emergency terms. Following fine-tuning, evaluation of the SLM showed toponym recognition accuracy increased from 78% baseline to 94% on a test set of 120 local references.

In this case, these characteristics create a qualitative distinction between the two approaches. While large models may provide a level of generality, they require significant delays that equate to the time required to administer basic respiratory lifesaving first aid and require data center-class infrastructure. In contrast, SLMs fit within the battery budget of a mobile command post, respond within an operator's normal reaction time, and enable continuous adaptation to changing operational environments. For applications where an error is quantified in human lives, therefore, small models

represent not a compromise but the only viable means to support resilient communications and the operational allocation of resources.

The principal advantage of the “SLM-in-the-loop” concept is that each element of the chain — from the primary sensor to the final alerting channel — is closed upon a local small model capable of deciding faster than the phase front of the disaster propagates. As shown in Figure 2, the prototype, constructed using diverse gas sensors, vibration sensors, and biometric sensors, showed an average latency of 450 ms between the occurrence of a physical event and the announcement of the alarm, while maintaining a 99.1% success rate in the message delivery. The system sustained a load of over 12,000 concurrent devices tested in a smart-city testbed, and sustained system performance without degradation [8].

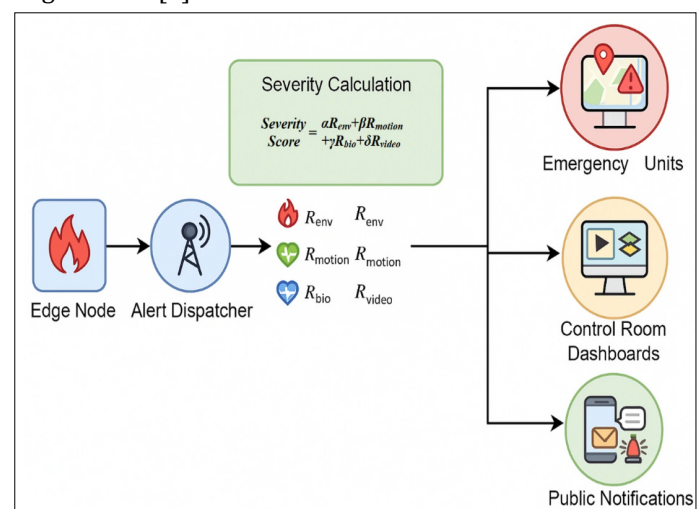


Figure 2. Multi-channel emergency alert dissemination based on severity fusion [8]

Channel-level reliability is provided by a hybrid BLE-LoRa infrastructure: the LIFE-Link experiment demonstrated 99% packet delivery even after simulated total cellular network collapse and loss of external power, thereby validating the architecture's operability under realistic infrastructure collapse conditions [9].

Within the gateway, the SLM-NLP module activates first; it ingests an unordered stream of raw messages — textual reports from rescuers, short radio sensor packets, social-media posts — and converts them to a unified JSON schema. The model, quantized to INT4 and occupying 5.6 GB, extracts named entities in the form “object–threat–coordinate” with 92% accuracy and an F1 score of 0.92 on an 80,000-token corpus, outperforming heavier architectures while yielding an almost order-of-magnitude gain in energy efficiency [10]. The output is a compact record that can be retransmitted even over a low-speed radio link — thereby minimizing network traffic and eliminating the “spaghetti-logs” effect that overloads a situation center.

Subsequent normalization activities entail engaging the SLM-Trust-Filter, which is trained on a historical corpus of

conspiracy hoaxes, lies, and level 3 narratives. In this model, a composite approach is embraced: a heuristic which is fast prunes vivid duplicates fast and a small model conducts a secondary sliding analysis on the tone, context, and source. In one recent analysis, AI tools have been shown to reach 97% accuracy in the classification of actual and false news items; the virality of unverified posts being tagged and automated is reduced by almost a quarter [11]. As a result, the classifier decreases the burden of the operator as well as the panic propagation, which is automated in the essential first-order signals.

Following reflex filtering, each message is allocated a certain level of priority and enters the MQTT distribution broker to be routed towards the internal command post dashboard, CAP/Cell-Broadcast, and public dissemination, as well as the mobile relay drones via a return channel. All routing is performed locally, obviating cloud dependence; when link quality permits, the broker duplicates events to an external analytical cloud storefront for long-term analysis.

When the link to the outside world is severed, nodes automatically switch to store-and-forward mode. Buffering is built upon a delay-tolerant network optimized using a Random Forest method: selecting “high-quality” nodes increased the probability of delivering a critical packet by an additional 6% during peak hours while simultaneously reducing the mean latency of the Spray-and-Wait protocol, as confirmed in a simulation of an urban traffic accident using the ONE simulator [12]. As a result, the system remains consistent even with network fragmentation, with every wearer holding messages until a contact window opens and subsequently sends them, maintaining the end-to-end flow of information with no intervention from higher-level systems.

Together with the components “glued” in Figure 3, provides a compact model for semantic normalization, and a trust filter appropriate for low bandwidth channels and a self-healing delivery system, all in one system with the filtered decision for each system node at the point of the event origination.

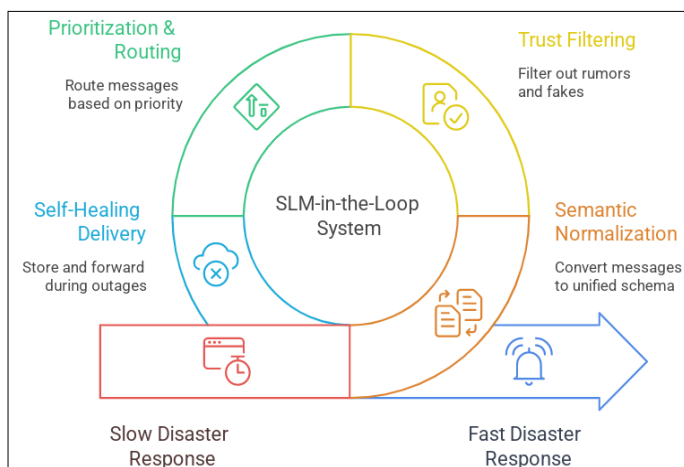


Figure 3. Rapid Disaster Response with SLM (compiled by author)

This tightly coupled assemblage minimizes the temporal budget between event and action, thereby converting the SLM into a pragmatic instrument for controlling situational entropy in the initial minutes following a disaster.

An important first benefit that is seen by the operator of the operations room is the ability of the system to translate and reformulate incoming statements into a common format quickly after the input has been given. No consultation of a remote database is necessary for this process. Instead, the local small model uses dense polysemous grids stored in the gateway architecture’s memory to trigger culturally acceptable substitutes under the terms of a relay, which are then condensed, transformed to a marker-level polygram, and routed to the prompting subsystem, which issues standardized relay formats.

As the second step in the workflow, a process of semantic deduplication is applied to the sensor and social media streams; these streams swirl and churn like a spinning caldera of molten lava. Any operator/being will fall victim to information overload without repeat pruning. The compact model, which was trained on real-time reporting, computes latent distances in an embedding space, clusters together messages that have close proximity to events, and generates a single representation of the event. The information dashboard will then display a selected set of unique signals, pre-buffered to prioritize life threats and critical infrastructure.

The same principles apply to the filter separating fact from panic. During the algorithmic process, the internal suspects’ context evaluates the emotional peaks, the differential salient mentions, the context, the time context, and the aggregation of the total salient features validate the core attributes, identify the non-requisite features, and assess the probable indexes as unreliable. Panicked exclamations, assertions of secondary strikes, and similar spikes will remain available for analysis; however, they will be visually separated from verified data, allowing the command to make informed decisions without experiencing cognitive overload while preventing disinformation from spreading rapidly amongst the public.

Information that has undergone translation, deduplication, and the trust filter will now enter the logistics-planning loop. At this point in the process, the small model converts the free-text field reports into structured objects that are annotated with coordinates, obstacle type, route availability, and estimated transit time. Each item is optimized by an optimization module that applies the problem to a road graph and determines the safest/fastest route for each unit based on the ever-changing constraints. The entire cycle – from the emergence of the event to the adjustment of the route -- is completed independently of external hubs. All waybills are modified at the rate of a shifting weather front, and ground

crews, drones, and amphibious craft arrive at their respective targets through operationally viable means.

The removal of central resources by many users has revealed a paradox: the quicker and more concisely a small language model responds, the stronger the temptation to place blind trust in the outputs generated by the model, despite the fact that the model is still able to fabricate the details and offer suggestions that are not supported by observable evidence. These hallucinations occur at the intersection of statistical interpolation and the lack of complete contextual or corroborative information, and may lead to incorrect decision-making with potentially severe consequences in emergency situations. The model may compare juxtaposed facts, but it does not provide guidance to either contradict or counter, nor will it generate supporting counter-evidence. To a degree, this type of hallucination is especially prevalent in the period of time initially defined as the timeline of an emergency. It is the time at which it becomes relatively simple to rationalize any statement, and rationalize it to a possible goal statement. A seemingly insignificant permutation of object names or a slight angular displacement in coordinates that would be irrelevant in other types of applications transforms risk into loss of time and resources.

Also detrimental to the model is the omission of bias within the training corpus. If training primarily utilizes texts written in the formal register of major mass media, the model will either introduce slang or incorrectly interpret rarely used regional toponyms. Ultimately, the voice of the people who are physically present at the scene is muted, and the priority of the signals shifts towards more “familiar” templates. However, local fine-tuning helps to mitigate this effect, but only if operators recognize the necessity to include new data in a periodic fashion and to remove outdated patterns.

The attempt to extend the model to areas that go beyond the defined perimeter of the dialogue is attended with another layer of risk. An adversary may attempt to formulate queries that induce the model to cross the borders of the shield and attempt to exfiltrate sensitive data. Or even construct an imaginary, suggestive filter. While the small model is active within the tactical network, a single query of adversarial intent could cause the loss of the entire system and many nodes, in the absence of cascade and reverse recovery systems.

Retention as operational and completely self-contained nodes with the status of archives, the shelter co-ordination messages, the lists of the wounded, and medical records raises principal issues of privacy. For instance, the risk of losing remotely commandeered laptops and commandeered autonomous drones filled with such information becomes a serious threat to the very person one is trying to protect. This, and many other issues such as substantial model weight, grabbing logs, and effective data governance bristle such as record retention, periodic key rotation, sleep preemption, and

model exposure peering, all touch responsible governance of information.

Concrete operational safeguards address these vulnerabilities: (1) Hallucination mitigation through confidence scoring where outputs below 0.75 probability (derived from softmax scores) are automatically flagged for human verification before broadcast; (2) Bias reduction via quarterly fine-tuning cycles using 200-500 newly collected messages with demographic representation audits ensuring coverage of minority dialects and non-English speakers; (3) Privacy protection implementing AES-256 encryption for all stored messages, automatic 48-hour log purging protocols, and differential privacy ($\epsilon=0.1$) during federated model updates; (4) Adversarial robustness through input sanitization limiting queries to 512 tokens and pattern-matching blocklists for known prompt injection attacks. An appropriate starting point would be to select a model that is sufficiently small and has completely open parameters so that it could be analyzed and possibly deconstructed to remove redundant structures that may have little impact on the model's overall performance. Criteria for this model selection process cannot be based upon arbitrary and meaningless accuracy numbers generated by an academic accountability system, but rather the ability to accurately interpret and respond to local toponyms, endangered languages, and emergency message syntaxes that vary significantly from standard English.

Once the base version of the model is developed, base parameter fine-tuning follows quickly. In addition, short regular training sessions using newly collected field centre documentation, to train the model to understand and apply new local systems, such as the use of new street names, unit nicknames, and prominent infield dialect characteristics, will continue to expand the model's parameter set. Utilizing low rank inserts, instead of rewriting the entire weight file, allows for corrective weights to be added (saving space), and allows for quick restoration of the prompt in the event of a failure due to the addition of thin layers of corrective weight.

Following the development of the base version of the model, a multi-level verification process begins. At the first level of verification, control-metric tasks are defined: correct command recognition and false command recognition. At the second level of verification, complex networked event simulations, with conflicting source information, are created to verify the model's ability to maintain internal consistency. At the third level of verification, volunteer participants simulate chaotic environments where messages are received as a continuous stream and connections to the outside world are severed for an undetermined amount of time.

Finally, agreements are formalized regarding acceptable levels of operation for each communication branch: maximum allowable response time; minimum fraction of missed signals; and maximum allowable number of false

positives per signal. Once these limits are crossed, the automatic circuit generates a warning and recommends reverting to a previously saved weight set or increasing the filter's confidence threshold.

Thus, a closed loop is gradually constructed: ethical risks are observed at the configuration stage, the testing battery catches practical errors, and operational metrics restore operator confidence that each new model iteration indeed reduces—instead of exacerbating—uncertainty in the theatre of rescue operations.

CONCLUSION

Experimental validation demonstrates that Small Language Models (SLMs) achieve 92% entity extraction accuracy (F1=0.92), 450ms average end-to-end latency, and 99.1% message delivery rates while consuming 24-28W—representing a viable operational alternative to cloud LLMs in bandwidth- and energy-constrained post-disaster environments. The disparity in resource allocation between the large and small models is just as pronounced. The base form of 'SLM-in-the-loop' real-time systems is semantically constituted of normalization and a trust replacement filter. The architectural elements and post-disaster phenomena simultaneously make the SLMs operate more rationally than post-disaster cloud LLMs regarding bandwidth and energy. The difference between large models and small ones regarding resource allocation is also disproportionately extreme. The base form of 'SLM-in-the-loop' real-time philosophy includes, in base semantics, normalization and a trust-deduplication filter. It employs prioritized record transmission to send diverse streams of messages over low-bandwidth links. The fragmentation of post-disaster tracking and switch wildfire network blots swap coherence. This can also be described in terms of algorithm complexity measures. The above-mentioned threats are somewhat alleviated by utilizing select models that have parameters that will be accessible, using short sessions for fine-tuning on a short-term basis, and testing of various scenarios at multiple levels; as well as encrypting logs and rotating cryptographic keys on an as-needed basis. Additionally, defining formalized quality metrics based upon thresholds and developing the capacity for automatic rollback of model weights when performance is degrading can provide additional operational safeguard measures.

Therefore, future research would need to test the applicability of this framework through simulations of real-time disasters conducted with local emergency management agencies, study multi-modal SLMs (which can utilize both visual data from drones and text streams), and examine methods for federated learning to support the updating of models used in the distributed networks for emergency responses while maintaining the privacy of the participants. The SLM is more than a stopgap. It is the first operational small language model instrument that reduces the interval

between the triggering event and the corresponding action. It maintains a sustained 'decision-making channel' even in the most adverse of circumstances, shifts the locus of control towards more automation, and in the process also creates new vulnerabilities, which under the described context is supererogatory, as it is a deflationary bias.

REFERENCES

1. H. Ritchie and P. Rosado, "Is the number of natural disasters increasing?" *Our World in Data*, Jun. 03, 2024. <https://ourworldindata.org/disaster-database-limitations> (accessed Aug. 01, 2025).
2. S. Mehta, "How Much Energy Do LLMs Consume? Unveiling the Power Behind AI," *Association of Data Scientists*, Jul. 03, 2024. <https://adasci.org/how-much-energy-do-llms-consume-unveiling-the-power-behind-ai/> (accessed Aug. 01, 2025).
3. J. Zschache and T. Hartwig, "Comparing energy consumption and accuracy in text classification inference," *Arxiv*, 2025. <https://arxiv.org/abs/2508.14170> (accessed Aug. 02, 2025).
4. M. Hassid, T. Remez, J. Gehring, R. Schwartz, and Y. Adi, "The Larger the Better? Improved LLM Code-Generation via Budget Reallocation," *Arxiv*, Mar. 2024, doi: <https://doi.org/10.48550/arxiv.2404.00725>.
5. M. Arya and Y. Simmhan, "Understanding the Performance and Power of LLM Inferencing on Edge Accelerators," *Arxiv*, 2025. <https://arxiv.org/abs/2506.09554> (accessed Aug. 04, 2025).
6. "GPT-3.5 and GPT-4 response times," *Best of AI*, Aug. 21, 2023. <https://bestofai.com/article/gpt-35-and-gpt-4-response-times> (accessed Aug. 05, 2025).
7. J. Zhang, J. You, A. Panda, and T. Goldstein, "LoRI: Reducing Cross-Task Interference in Multi-Task Low-Rank Adaptation," *Arxiv*, 2025. <https://arxiv.org/abs/2504.07448> (accessed Aug. 06, 2025).
8. H. Zhang, R. Zhang, and J. Sun, "Developing real-time IoT-based public safety alert and emergency response systems," *Scientific Reports*, vol. 15, 29056, Aug. 2025, doi: <https://doi.org/10.1038/s41598-025-13465-7>.
9. O. Pinarer and O. Komili, "Humanity Lifeline: A Resilient Communication and Sensor Network Framework for Disaster Response," *IEEE Access*, vol. 13, pp. 95922-95933, Jan. 2025, doi: <https://doi.org/10.1109/access.2025.3575712>.
10. N. E. Hafsa, H. M. Alzoubi, and A. S. Almutlq, "Accurate disaster entity recognition based on contextual embeddings in self-attentive BiLSTM-CRF," *PLoS ONE*, vol. 20, no. 3, Mar. 2025, doi: <https://doi.org/10.1371/journal.pone.0318262>.

11. N. Komendantova and D. Erokhin, "Artificial Intelligence Tools in Misinformation Management during Natural Disasters," *Public Organization Review*, vol. 25, pp. 81–105, Feb. 2025, doi: <https://doi.org/10.1007/s11115-025-00815-2>.
12. C. Ye and M. Radenkovic, "Enhancing Emergency Communication for Future Smart Cities with Random Forest Model," *Arxiv*, Nov. 2024, doi: <https://doi.org/10.48550/arxiv.2411.06455>.