FirstAidQA: A Synthetic Dataset for First Aid and Emergency Response in Low-Connectivity Settings

Saiyma Sittul Muna*
Islamic University of Technology
Dhaka, Bangladesh
saiymasittul@iut-dhaka.edu

Rezwan Islam Salvi*
Islamic University of Technology
Dhaka, Bangladesh
rezwanislam@iut-dhaka.edu

Mushfiqur Rahman Mushfique*
Islamic University of Technology
Dhaka, Bangladesh
mushfique2@iut-dhaka.edu

Ajwad Abrar
Islamic University of Technology
Dhaka, Bangladesh
ajwadabrar@iut-dhaka.edu

Abstract

In emergency situations, every second counts. The deployment of Large Language Models (LLMs) in time-sensitive, low or zero-connectivity environments remains limited. Current models are computationally intensive and unsuitable for low-tier devices often used by first responders or civilians. A major barrier to developing lightweight, domain-specific solutions is the lack of high-quality datasets tailored to first aid and emergency response. To address this gap, we introduce **FirstAidQA**, a synthetic dataset containing 5,500 high-quality question–answer pairs that encompass a wide range of first aid and emergency response scenarios. The dataset was generated using a Large Language Model, ChatGPT-4o-mini, with prompt-based in-context learning, using texts from the Vital First Aid Book (2019). We applied preprocessing steps such as text cleaning, contextual chunking, and filtering, followed by human validation to ensure accuracy, safety, and practical relevance of the QA pairs. FirstAidQA is designed to support instruction-tuning and fine-tuning of LLMs and Small Language Models (SLMs), enabling faster, more reliable, and offline-capable systems for emergency settings. We publicly release the dataset to advance research on safety-critical and resource-constrained AI applications in first aid and emergency response. The dataset is available on Hugging Face at https://huggingface.co/datasets/i-am-mushfiq/FirstAidQA.

1 Introduction

Large Language Models (LLMs) such as GPT-5, Gemini, and Claude have shown great capabilities across a wide range of generalized natural language tasks. However, their practical deployment in safety-critical, real-time applications, such as first aid and emergency response remains limited [1]. Our observation suggests that a primary obstacle is the lack of high-quality, domain-specific QA datasets that contain the unique information and situational knowledge required in emergency settings [2].

First aid response represents a critical domain where time, clarity, and safety are crucial. In many low resource or offline environments, such as disaster zones, rural clinics, remote regions or socio-economically backward regions, access to high-speed internet or modern computing infrastructure is

*These authors contributed equally to the work.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Muslims in ML Workshop.

often unavailable [3]. In these scenarios, having LLMs or SLMs that can accurately, efficiently and safely provide actionable medical guidance on the specific situation could be priceless [4].

Existing QA benchmarks (e.g., BioASQ, MedQA, PubMedQA) are primarily oriented toward clinical diagnostics or academic biomedical literature. However, they do not address the instructional or situational knowledge needed in layperson-administered first aid, where individuals may have little to no formal training or expertise [5]. To address this gap, we introduce **FirstAidQA**, a synthetic dataset of 5500 question–answer pairs spanning a wide range of general and situational first aid and emergency response scenarios. Each QA pair is generated using prompt-based querying of a Large Language Model - ChatGPT-4o-mini, guided via in-context learning with the corpora from the Vital First Aid (2019) book [6]. Specifically, we segmented and pre-processed textual content from the book and supplied these context chunks as input to the LLM to generate realistic and situational questions and answers.

Our dataset is designed to support the fine-tuning and instruction-tuning of LLMs and SLMs for edge deployment in bandwidth-constrained or offline settings. Unlike clinical QA datasets that often mirror academic exam formats, FirstAidQA emphasizes practical, situational, and procedural knowledge, covering topics such as treating burns, managing bleeding, handling animal bites, and other common emergencies. Our main contributions are as follows:

1. We propose and publicly release FirstAidQA, the first synthetic QA dataset specifically tailored to first aid and emergency response, consisting of 5,500 question–answer pairs.
2. We conduct human validation on a randomly sampled subset of the dataset and outline a framework for quality assurance in synthetic data pipelines.
3. We highlight directions for using FirstAidQA in fine-tuning smaller, deployable models for real-time, offline medical assistance.

2 Related Work

Synthetic QA datasets have emerged as a cost-effective, scalable alternative to manual annotation. Self-Instruct demonstrated that LLMs can generate high-quality instruction data via prompt engineering without human need [7]. The authors used GPT-3 to create instruction–input–output triplets, filtered them, and fine-tuned on the resulting data. In medicine, Kotschenreuther's EHR-DS-QA generated 156K QA pairs from discharge summaries, improving retrieval-augmented clinical QA [8]. Cahlen's Offline Practical Skills QA provided a LoRA adapter fine-tuned using TinyLlama-1.1B model for providing information on survival and first-aid in offline settings [9].

Several benchmarks exist in healthcare and related domains. In biomedicine, the BioASQ Challenge provides expert-authored QA sets [10], while COVID-QA offers $\sim$2,000 curated pairs from scientific literature that significantly boost fine-tuned model accuracy [11]. Consumer health resources include MedQuAD (47K QA pairs from NIH websites) [12]. Besides healthcare, datasets such as FinTextQA (finance) [13] and MedMCQA (medical exams) [14] illustrate domain-specific QA's effectiveness.

However, first aid remains underexplored. Existing systems are FAQ-based chatbots (e.g., *Dr.FirstAider*) or evaluations of assistants like Siri, Alexa, and ChatGPT, which often miss key evidence-based steps or provide incomplete guidance. Studies show LLMs can align with clinical guidelines (e.g. burn care) but occasionally omit critical details. In summary, no dedicated first-aid QA dataset exists. Current efforts rely on manual curation or chatbot-style systems. This gap motivates our creation of FirstAidQA.

3 Dataset Design

FirstAidQA is a synthetic dataset tailored for safety-critical, low-connectivity settings. Its development involved source selection, QA generation, post-processing, and pipeline integration, with emphasis on accuracy, usability, and ethics.

3.1 Source Material

The dataset is derived from the certified *Vital First Aid Book 2019*[6], chosen for its structured, comprehensive coverage of emergency care. Topics include general protocols (e.g. DRSABCD),

Table 1: Category Breakdown of the FirstAidQA Dataset

Category	Description / Subcategories
General Emergency Procedures	Preparedness, priorities, DRSABCD protocol, and basic emergency techniques.
CPR	Standard methods for adults, children, and infants; special cases such as drowning, choking, and overdose.
Road Traffic Accidents	Scene safety, casualty assessment and extrication, multi-casualty management.
Moving Casualties	Relocation methods, spinal precautions, and adaptations for solo vs. team responders.
First Aid Equipment & Techniques	Use of kits, dressings, slings, splints, and improvised tools.
Family & Community Safety	Household preparedness and community-level emergency response.
Patient Examination & Monitoring	Vital signs, injury severity, and temperature assessment.
Specific Medical Conditions	Respiratory emergencies, bleeding, burns, fractures, head/facial injuries, temperature-related emergencies, cardiovascular issues, neurological/systemic crises, bites, poisoning, and other acute conditions.
Neck & Spinal Injuries	Assessment, stabilisation, and safe handling.

CPR, accident management, casualty movement, equipment use, patient assessment, and conditions such as asthma, bleeding, burns, fractures, head injuries, and temperature-related emergencies.

Content was segmented for context-preserving QA generation, for example, casualty movement (dragging, spinal precautions) and head injuries (fractures, fluid leakage). The manual also ensures consistency with international standards (American Red Cross, ILCOR).

3.2 Task Taxonomy & Category Breakdown

The dataset follows a structured task taxonomy to ensure comprehensive coverage of emergency scenarios. Table 1 presents the main categories.

3.3 Prompting and Synthetic Data Generation

The core of the dataset construction was the prompt design used to drive synthetic data generation with ChatGPT-4o-mini [7]. The prompt was carefully structured to explicitly define the model's role, provide context from the certified first aid manual, and specify the desired output format. In particular, the model was framed as an expert in synthetic dataset creation for first aid and medical emergencies, which encouraged medically precise and contextually rich answers. Each prompt also included a topic-specific text segment from the manual (such as guidance on moving a casualty or treating burns and scalds). The instructions directed the model to generate detailed question-answer pairs that were medically accurate, step-by-step, and actionable, with questions reflecting diverse perspectives such as those of bystanders, trained responders, or lone rescuers, and covering a wide range of settings including accidents, extreme weather, and confined spaces. The output was required to be in JSON format to enable seamless integration into machine learning pipelines.

The base template was:

> *"Imagine you are a renowned expert in synthetic dataset creation for first aid and medical emergencies. Your task is to generate 20 diverse question–answer pairs from the given corpus of a certified first aid manual. Answers must be detailed, medically accurate, and reflect multiple perspectives and scenarios. Provide the output in JSON format."*

Table 2: Mean Human Evaluation Scores for FirstAidQA (3 evaluators)

Criterion	Mean Score (1–5)
Clarity	4.2
Relevance	4.7
Specificity & Completeness	4.0
Safety & Accuracy	3.7

To expand coverage, the instruction *"Generate 20 more question–answer pairs. Ensure they are not repeated from the previous response"* was iteratively applied until approximately 100 QA pairs per topic were created. This structured prompting enabled systematic synthetic data generation. Each topic block was processed in batches of 20, reviewed for accuracy and diversity, and refined through prompt adjustments when necessary (e.g. explicitly requesting pediatric or elderly cases). Repeating this process across different topics produced 5500 QA pairs in total.

4 Quality Assurance of the Dataset

4.1 Filtering Method

We identified and extracted chunks of text from the book that are most relevant for first aid and avoided less applicable content and theories. Every chunk was evaluated regarding whether it could be used to generate QA pairs that would be applicable in real-world scenarios.

4.2 Safety and Hallucination Check

The LLM (ChatGPT-4o-mini) was given prompts explicitly instructing it to generate the QA pairs only from the provided chunks, so that its responses are firmly in context. To mitigate the risks of bias, we took diversified chunks that capture different situations across the emergency response domain.

4.3 Manual and Expert Review

We randomly selected 200 QA pairs, from different contexts, for expert evaluation. 3 medical professionals reviewed these pairs. They gave scores to every pair on a 1–5 scale for each of the following criteria:

- (a) **Clarity**: The Q&A pair is easy to read and understand.
- (b) **Relevance**: The Q&A pair is directly relevant to first aid scenarios.
- (c) **Specificity & Completeness**: The question is specific and the answer fully addresses it with necessary steps/information.
- (d) **Safety & Accuracy**: The answer is medically accurate and does not suggest unsafe actions.

The reported scores in Table 2 represent the mean ratings given by the 3 evaluators. Examples of QA pairs flagged for potentially unsafe or inaccurate instructions during expert validation are provided in Appendix A. Such occurrences should be taken into caution while using the dataset.

5 Limitations

While being very resourceful, the dataset still has boundaries. First of all, it is **Not a Medical Substitute**. It supports, but does not replace professional care. The dataset has Emergency-Only Focus, i.e. it is not for general health advice. It should be used with caution and professional help should always be sought when available.

6 Conclusion

During emergencies, when every second is critical, the absence of offline and reliable language models has left a vacuum in first aid and emergency response. Current LLMs may be powerful but

falter in low-connectivity settings due to their cloud dependency. Domain-specific datasets are also scarce in this field. Our work bridges this gap with a synthetic dataset that can be used to provide valuable feedback during emergencies.

Developed through prompt-based querying of ChatGPT-4o-mini, and subsequently evaluated by experts, our dataset offers a high-quality, safety-focused tool for deployment in real-world emergencies. By releasing our dataset to the public, we not only provide the first synthetic QA dataset for this valuable domain but also enable potential breakthroughs in low-resource offline AI applications. We envision a world where such models empower individuals in disaster zones or rural communities with real-time medical consultation, saving lives where human assistance cannot reach instantly. This is a call to action for the research community to develop emergency response technology to revolutionize how we deliver care during moments that matter most.

References

[1] Ige Gabriel and Adebayo Precious. Challenges of implementing ai in low-resource healthcare settings. 08 2025.

[2] Jayetri Bardhan, Kirk Roberts, and Daisy Zhe Wang. Question answering for electronic health records: Scoping review of datasets and models. *Journal of Medical Internet Research*, 26:e53636, 2024.

[3] Kinalyne Perez, Daniela Wisniewski, Arzu Ari, Kim Lee, Cristian Lieneck, and Zo Ramamonjiarivelo. Investigation into application of ai and telemedicine in rural communities: a systematic literature review. In *Healthcare*, volume 13, page 324. MDPI, 2025.

[4] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.

[5] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019.

[6] Vital First Aid Training Services Pty Ltd. *New Vital First Aid: First Aid Book*. Vital First Aid Training Services Pty Ltd, Australia, reprinted edition (november 2019) edition, 2019. Illustrated by Nicole Humphries; first published in 2012; for electronic distribution only.

[7] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.

[8] Konstantin Kotschenreuther. Ehr-ds-qa: A synthetic qa dataset derived from medical discharge summaries for enhanced medical information retrieval systems. *PhysioNet*, 2024.

[9] Cahlen Humphreys. Tinyllama-1.1b offline practical skills qa adapter (qlora). `https://huggingface.co/cahlen/tinyllama-offline-practical-skills-qa-qlora`.

[10] George Tsatsaronis, Michael Schroeder, Georgios Paliouras, Yannis Almirantis, Ion Androutsopoulos, Eric Gaussier, Patrick Gallinari, Thierry Artieres, Michael R Alvers, Matthias Zschunke, et al. Bioasq: A challenge on large-scale biomedical semantic indexing and question answering. In *AAAI fall symposium: Information retrieval and knowledge discovery in biomedical text*. Arlington, VA: Citeseer, 2012.

[11] Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. Covid-qa: A question answering dataset for covid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, 2020.

[12] Asma Ben Abacha and Dina Demner-Fushman. A question-entailment approach to question answering. *BMC bioinformatics*, 20(1):511, 2019.

[13] Jian Chen, Peilin Zhou, Yining Hua, Yingxin Loh, Kehui Chen, Ziyuan Li, Bing Zhu, and Junwei Liang. Fintextqa: A dataset for long-form financial question answering. *arXiv preprint arXiv:2405.09980*, 2024.

[14] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR, 2022.

A Validation Flags

Table 3: Flagged Q&A Pairs with Potentially Unsafe Instructions - Requires Cautious Handling.

Item	Question	Answer (as given)	Why Flagged
95	What is the first step in treating a bee sting if the person has a known allergy?	If the person has a known allergy to bee stings, the first step is to have them take any allergy medication, such as an epinephrine injection (EpiPen) if prescribed. Then, monitor their breathing, apply pressure immobilisation if necessary, and seek medical assistance immediately.	Pressure immobilisation is not recommended for anaphylaxis; epinephrine use should be immediate and under medical guidance.
102	How do you adjust CPR for a person with an airway obstruction, such as choking?	If a person is choking and unable to breathe, perform abdominal thrusts (Heimlich maneuver) to attempt to dislodge the object. If they become unconscious, lay them flat and begin CPR. If you find an obstruction in the airway during rescue breaths, attempt to clear it using chest thrusts or finger sweeps before continuing CPR.	Blind finger sweeps are not advised; can push object deeper or cause injury. Guidelines emphasize back blows, abdominal thrusts, then CPR without sweeps.
128	How should I treat a marine life sting?	For a sting from marine life like a stingray, wash the affected area with vinegar to neutralize toxins. Avoid rubbing the sting site and seek medical assistance immediately.	Vinegar is useful for jellyfish, but not for stingrays; hot water immersion is recommended for stingray injuries.
137	What steps should I take if I encounter a tick on an adult?	For an adult, first use a freezing product such as 'Tick OFF' to kill the tick. Avoid pinching or pulling on the tick. After the tick is dead, carefully remove it and check for other ticks in crevices like behind the ears or in the hair. If symptoms of paralysis or weakness appear, seek medical aid immediately.	Freezing sprays are not recommended; safest method is direct removal with fine-tipped tweezers. Killing the tick first increases risk of toxin release.

Item	Question	Answer (as given)	Why Flagged
143	What is the most reliable method to take an adult's body temperature?	The most reliable method to take an adult's body temperature is by using a digital thermometer under the tongue. This provides a quick and accurate measurement. For those who cannot hold a thermometer under the tongue, an alternative is the groin or armpit.	Oral readings may be inaccurate in certain conditions; rectal or tympanic methods are often considered more reliable clinically.
146	What if a child with diarrhoea prefers sugary drinks instead of rehydration fluid?	While tempting, sugary drinks can worsen diarrhoea due to high osmotic load. Only diluted sugary fluids (e.g., juice or lemonade mixed with water at the correct ratio) should be given. Rehydration fluids like Gastrolyte remain the safest and most effective choice.	Even diluted sugary drinks are discouraged; Oral Rehydration Solution (ORS) is the only recommended option.
180	What should you do if someone has been poisoned by swallowing a toxic substance like bleach or drain cleaner?	If someone has swallowed a toxic substance like bleach or drain cleaner, do not induce vomiting. Rinse their mouth with water and seek immediate medical help. These substances can cause severe damage to the throat, esophagus, and stomach, and professional treatment is essential.	Rinsing the mouth may still cause swallowing; drinking water or neutralizers is not recommended. Call poison control/emergency services immediately.
184	How do you safely remove a tick using a first aid kit?	To remove a tick, use fine-tipped tweezers or forceps from the first aid kit to grasp the tick as close to the skin's surface as possible. Pull gently and steadily without twisting to avoid leaving parts of the tick behind. Clean the bite area with antiseptic solution and apply a sterile dressing.	Twisting is sometimes recommended to avoid breaking mouthparts; best practices differ. Leaving fragments behind can cause infection — should follow official local tick-removal guidance.