PROTEUS: Protocol-aware Replication using Observational Techniques for Extensible Universal Simulation of ICS Devices

Submission: #001

Abstract

Industrial control systems (ICS) rely on specialized network protocols to coordinate safety and mission critical physical processes through the usage of programmable logic controllers (PLCs). Advancing testing, integration, security analysis, and training requires faithful emulation of protocol behavior; however, progress is constrained by scarce access to hardware and the lack of standardized, machine learning (ML) ready corpora that provide clean request–response (R/R) pairs for supervised generative modeling. We present PROTEUS, a novel fuzzing based methodology to automatically generate ICS protocol datasets suitable for generative modeling. We deliver datasets for representative ICS protocols (Modbus/TCP, S7comm, and DNP3) explicitly designed for response synthesis.

PROTEUS enables fair comparisons, encourages rigorous methodology, and lowers the barrier to building protocol emulators for testing, interoperability validation, honeypot development, and training settings where physical devices are unavailable.

CCS Concepts

• Software and its engineering → Virtual machines; Virtual memory; • Computer systems organization → Heterogeneous (hybrid) systems.

Keywords

Industrial Control Systems, Generative Modeling, Fuzzing, Dataset Generation, Protocol Emulation

ACM Reference Format:

Anonymous Author(s) 2026 PROTEUS: Protocol-aware Replication using Observational Techniques for Extensible Universal Simulation of ICS Devices. In *ACM SIGSAC 33rd ACM Conference on Computer and Communications Security (CCS) (CCS '26), November 15-19, 2026, The Hague, The Netherlands.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/0000000000

1 Introduction

Industrial control systems form the backbone of critical infrastructure spanning energy grids, water treatment facilities, manufacturing plants, transportation networks, and utilities [8]. These systems orchestrate physical processes through programmable logic controllers, supervisory control and data acquisition (SCADA) systems, and field devices that communicate using specialized industrial protocols. The increasing connectivity and digitization of ICS have dramatically expanded their attack surface, exposing critical infrastructure to cyber threats with potentially catastrophic consequences ranging from service disruptions to physical damage [1, 8].

Despite the critical importance of ICS security, research and development are severely hampered by the scarcity of accessible hardware, proprietary protocols, and safety constraints that prevent experimentation on operational systems. This has motivated efforts to develop virtual environments, simulators, and emulators that can faithfully reproduce ICS behavior for testing, training, and security research [2, 7, 10, 11]. Virtual development environments enable software testing without physical hardware [10, 11], while comprehensive testbed frameworks like ICSSIM [2] provide realistic settings for security evaluation. However, these simulation approaches often require extensive domain knowledge, manual configuration, and access to reference implementations or detailed protocol specifications.

The machine learning community has increasingly turned to ICS datasets to develop intelligent security solutions, but the available corpora exhibit significant limitations. Existing publicly available datasets predominantly focus on intrusion detection and anomaly classification [3, 9], providing labeled network traffic captures designed to distinguish normal from malicious behavior. While valuable for training defensive systems, these datasets lack the structured request-response (R/R) pairs necessary for generative modeling tasks. Morris and Gao's industrial control system traffic datasets [9] established early benchmarks for intrusion detection research, and more recent efforts like the anomaly detection dataset by Dehlaghi et al. [3] provide labeled samples for classification. However, neither provides the clean input–output pairs required to train sequence to sequence models that can synthesize protocol compliant responses.

Recent work has begun exploring generative approaches for ICS protocol data. Yang et al. [13] proposed using generative adversarial networks (GANs) to generate fuzzing test cases for industrial protocols, while Zarzycki et al. [14] investigated GAN architectures for testing process control networks against cyber attacks. Despite these promising directions, no prior work has released a standardized, ML ready corpus of paired protocol request–response exchanges suitable for supervised training and rigorous evaluation of generative models. This absence of a benchmark

CCS '26, November 15-19, 2026, The Hague, The Netherlands.

ACM ISBN 979-8-0000-0000-0/00/00...$15.00
https://doi.org/10.1145/0000000000

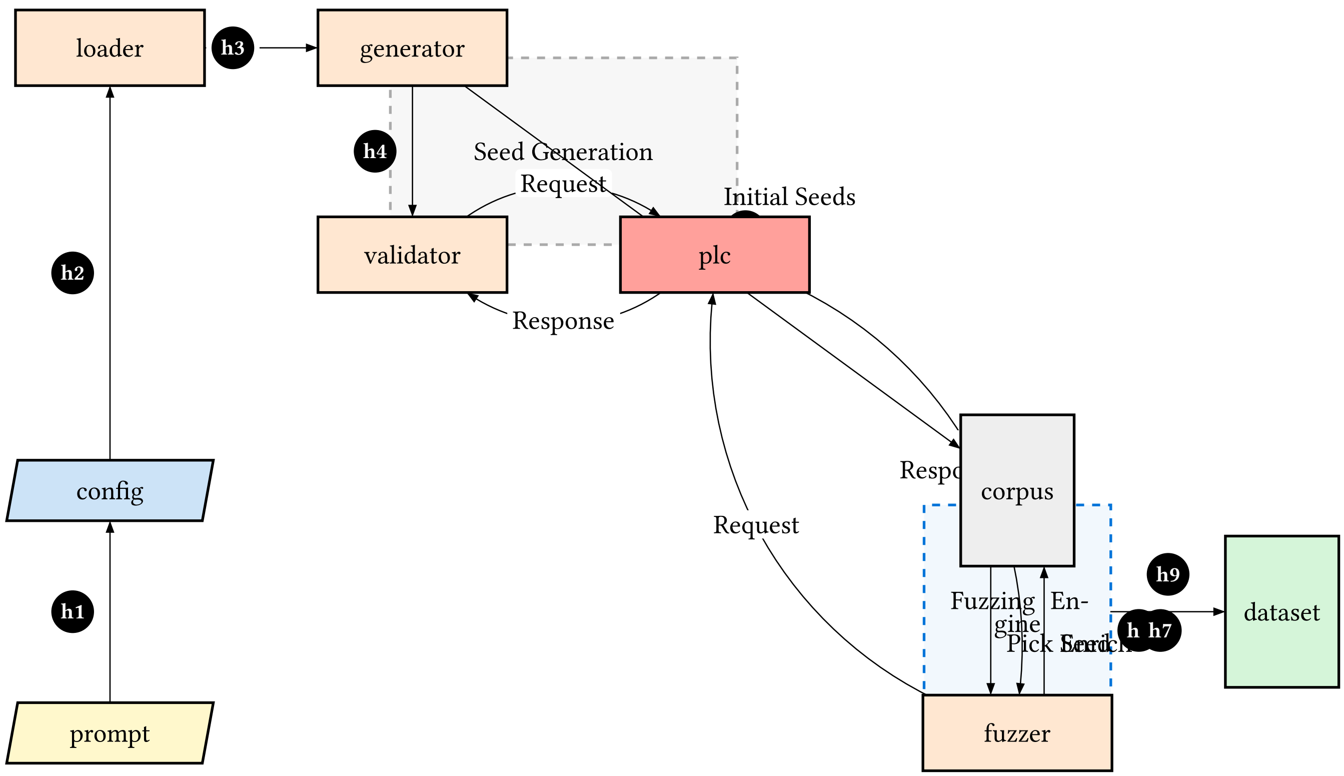

Figure 1: High-level overview of the PROTEUS pipeline illustrating the seed generation and fuzzing components. The Loader ingests the LLM-generated protocol specification, the Seed Generator produces initial valid requests, the Validator checks the generated requests for correctness and enriches with responses, and the Fuzzer iteratively mutates requests to explore protocol behavior, storing interesting request/response pairs in the Dataset for dataset construction.

dataset hinders fair comparison across approaches, prevents reproducible research, and limits the development of practical protocol emulation systems.

The gap between classification oriented datasets and the needs of generative modeling is particularly acute. Training models to emulate protocol behavior, generating correct responses given arbitrary requests, requires clean, protocol faithful R/R pairs that capture the deterministic logic of industrial devices. Such models have applications beyond security, including software testing [7, 11], interoperability validation, honeypot development [12], and training environments where access to physical hardware is constrained. Yet researchers currently lack access to standardized corpora that would enable systematic investigation of generative techniques for ICS protocol emulation.

This paper introduces PROTEUS, a novel fuzzing based methodology to ondemand generate curated request-response pairs for representative ICS protocols, we explicitly created datasets using our framework for Modbus, S7comm and DNP3, nevertheless the framework can be used for additional ones without manual intervention. The datasets are explicitly designed for generative modeling. Our framework repurposes the ideas and notions used in fuzzing to explore valid requests against a ICS device that communicates using a TCP protocol. We provide two interoperable serializations binary preserving (hex/base64) for byte level models and canonical textual JSONL for tokenizer friendly training and frame the benchmark around response synthesis: given a request, produce a protocol conformant response. Informed by dataset quality principles for machine learning [4–6], we provide validation tools to ensure ML suitability.

High Level Pipeline: Figure 1 presents an overview of the PROTEUS pipeline. The process begins with an LLM prompt that yields a protocol specification h1, which is then materialized as a JSON protocol specification h2 and ingested by the Loader. The Loader parses the specification as described in detail in Section ref{sec:seed_generation} before passing it to the Seed Generator h3, which produces an initial set of valid protocol requests h4. These requests are forwarded to the Validator, which interacts with a real ICS device (PLC) to obtain responses and ensure request-response correctness h8. The resulting initial seeds are then transferred to the seed corpus h5. During fuzzing, the Fuzzer picks seeds from the corpus h6, sends mutated requests to the PLC, collects responses h8 and adds interesting new packets back into the seed corpus h7. Finally, the fuzzing engine exports the accumulated request-response pairs into the Dataset h9 for training and evaluation of generative models.

The **main contributions** of this paper are:

- A novel fuzzing based methodology to systematically generate high quality request–response pairs for potentially any ICS protocol, ensuring protocol compliance and diversity.
- Standardized corpus of request–response pairs for Modbus/TCP, S7comm, and DNP3, released in binary preserving and canonical textual forms for generative modeling.
- Quantify quality of the dataset using established data quality metrics for machine learning datasets to ensure its suitability for training robust models.
- Test resulting datasets with multiple baseline models including byte-level sequence models and tokenizer-friendly language models.
- Publicly release the datasets and methodology at: https://anonymous.4open.science/r/icsclone/

Metric	Type	Baseline	Ours
FC entropy	Req	2.1787	**2.2734**
Address Skewness	Both	0.3419	**0.0577**
Address Coverage	Both	300	**63726**
Byte entropy	Req	4.0017	**7.8807**
	Resp	3.8356	**5.7647**
Bigram entropy	Req	6.7598	**15.0316**
	Resp	6.4115	**9.6484**
4-gram entropy	Req	8.8889	**16.6470**
	Resp	7.9201	**10.7034**
Avg length	Req	14.05 ± 2.99	89.23 ± 71.88
	Resp	10.20 ± 1.55	30.68 ± 54.30

Table 1: Comparison of throughput and latency.

2 Methodology

Product Details		Inventory & Price		Region
Type	Item	Stock	Price	
Electronics	**Laptop**	15	$1200	North
Electronics	**Monitor**	30	$350	North
Office	**Desk Chair**	45	$150	South
Office	**Desk Lamp**	120	$45	South
Supplies	**Paper (Ream)**	500	$5	East

References

[1] Allan Cook, Richard Smith, Leandros Maglaras, and Helge Janicke. 2016. Measuring the risk of cyber attack in industrial control systems. In *4th International Symposium for ICS & SCADA Cyber Security Research 2016*, 2016.

[2] Alireza Dehlaghi-Ghadim, Ali Balador, Mahshid Helali Moghadam, Hans Hansson, and Mauro Conti. 2023. ICSSIM—a framework for building industrial control systems security testbeds. *Computers in Industry* 148, (2023), 103906.

[3] Alireza Dehlaghi-Ghadim, Mahshid Helali Moghadam, Ali Balador, and Hans Hansson. 2023. Anomaly detection dataset for industrial control systems. *IEEE Access* 11, (2023), 107982–107996.

[4] Junhua Ding and XinChuan Li. 2018. An approach for validating quality of datasets for machine learning. In *2018 IEEE International Conference on Big Data (Big Data)*, 2018. 2795–2803.

[5] Youdi Gong, Guangzhen Liu, Yunzhi Xue, Rui Li, and Lingzhong Meng. 2023. A survey on dataset quality in machine learning. *Information and Software Technology* 162, (2023), 107268.

[6] Nitin Gupta, Hima Patel, Shazia Afzal, Naveen Panwar, Ruhi Sharma Mittal, Shanmukha Guttula, Abhinav Jain, Lokesh Nagalapatti, Sameep Mehta, Sandeep Hans, and others. 2021. Data Quality Toolkit: Automatic assessment of data quality and remediation for machine learning datasets. *arXiv preprint arXiv:2108.05935* (2021).

[7] Muhammad Zohaib Iqbal, Andrea Arcuri, and Lionel Briand. 2015. Environment modeling and simulation for automated testing of soft real-time embedded software. *Software & Systems Modeling* 14, 1 (2015), 483–524.

[8] Stephen McLaughlin, Charalambos Konstantinou, Xueyang Wang, Lucas Davi, Ahmad-Reza Sadeghi, Michail Maniatakos, and Ramesh Karri. 2016. The Cybersecurity Landscape in Industrial Control Systems. *Proceedings of the IEEE* 104, 5 (2016), 1039–1057. https://doi.org/10.1109/JPROC.2015.2512235

[9] Thomas Morris and Wei Gao. 2014. Industrial control system traffic data sets for intrusion detection research. In *International conference on critical infrastructure protection*, 2014. 65–78.

[10] Pradyumna Sampath and B Rachana Rao. 2011. Efficient embedded software development using QEMU. In *13th Real Time Linux Workshop*, 2011.

[11] Hadipurnawan Satria, Budiono Wibowo, Jin B Kwon, Jeong B Lee, and Young S Hwang. 2009. VDEES: A virtual development environment for embedded software using open source software. *IEEE transactions on consumer electronics* 55, 2 (2009), 959–966.

[12] Christoforos Vasilatos, Dunia J. Mahboobeh, Hithem Lamri, Manaar Alam, and Michail Maniatakos. 2025. LLMPot: Dynamically Configured LLM-based Honeypot for Industrial Protocol and Physical Process Emulation. Retrieved from https://arxiv.org/abs/2405.05999

[13] Hongsen Yang, Yuezhen Huang, Zhiyong Zhang, Fei Li, Brij B Gupta, and P VijayaKumar. 2024. A novel generative adversarial network-based fuzzing cases generation method for industrial control system protocols. *Computers and Electrical Engineering* 117, (2024), 109268.

[14] Krzysztof Zarzycki, Patryk Chaber, Krzysztof Cabaj, Maciej Ławryńczuk, Piotr Marusak, Robert Nebeluk, Sebastian Plamowski, and Andrzej Wojtulewicz. 2023. GAN neural networks architectures for testing process control industrial network against cyber-attacks. *IEEE Access* 11, (2023), 49587–49600.

A Open Science

We are committed to open science principles and will publicly release the datasets, code, and methodology associated with this work upon publication. The datasets will be made available in both binary preserving (hex/base64) and canonical textual (JSONL) formats to facilitate use by a wide range of machine learning models. The code for the PROTEUS pipeline, including the Loader, Seed Generator, Validator, Fuzzer, and dataset construction tools, will be released under an open source license to encourage adoption and further development by the research community. We believe that providing these resources will enable reproducible research, foster collaboration, and accelerate progress in the field of ICS protocol emulation and security analysis.