Network Traffic Foundation Model with Adaptation via In-Context Learning and Mixture-of-Experts

Miru Kim[†] **and Minhae Kwon**[†*]
[†]Department of Intelligent Semiconductor
[*]School of Electronic Engineering
Soongsil University, Seoul, Republic of Korea
mirukim00@soongsil.ac.kr, minhae@ssu.ac.kr

Abstract

Network traffic patterns vary significantly across collection environments, which often degrades the generalization capability of existing models. This motivates the need for a foundation model that can capture the underlying patterns of network traffic independent of collection environments and can be shared across multiple downstream tasks. In this work, we propose a foundation model specialized for network traffic data. After building a foundation model, our framework leverages In-Context Learning (ICL) and incorporates a Mixture-of-Experts (MoE) architecture for downstream tasks, such as intrusion type classification, utilizing the foundation model. To assess the validity of the proposed approach, we conduct ablation studies on ICL and MoE components, demonstrating their respective contributions to adaptability and efficiency. This study highlights the necessity of a universal foundation model for network traffic analysis and suggests a promising direction toward building scalable, general-purpose solutions for future network intelligence applications.

1 Introduction

Network traffic analysis is a cornerstone for a wide range of applications, including intrusion detection, anomaly detection, and traffic forecasting. However, traffic data collected from different environments (e.g., infrastructures or time periods) exhibit highly diverse patterns. This heterogeneity severely limits the generalization of traditional machine learning models, which are often trained on a single dataset or require costly task-specific fine-tuning.

Recent progress in foundation models has demonstrated remarkable success in capturing general-purpose representations that transfer across diverse domains [1, 2, 3]. Nevertheless, building a foundation model for network traffic data presents unique challenges. First, traffic patterns evolve naturally over time and vary across collection environments, demanding models that learn environment-invariant representations. Second, network traffic is inherently sequential, requiring architectures that can capture both short- and long-term temporal dependencies. Finally, continual adaptation to new data sources and tasks must be achieved without sacrificing compatibility with prior knowledge, as newer models should not deteriorate on queries previously handled correctly.

In this work, we aim to address these challenges by developing a transformer-based foundation model for network traffic data, utilizing task adaptation with ICL [4, 5, 6, 7] and MoE [8, 9, 10, 11]. We validate the proposed framework on a representative downstream task: intrusion type classification. Through ablation studies on ICL and MoE components, we investigate their contributions to adaptability, efficiency, and backward compatibility. Our work highlights the importance of designing

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI and ML for Next-Generation Wireless Communications and Networking (AI4NextG).

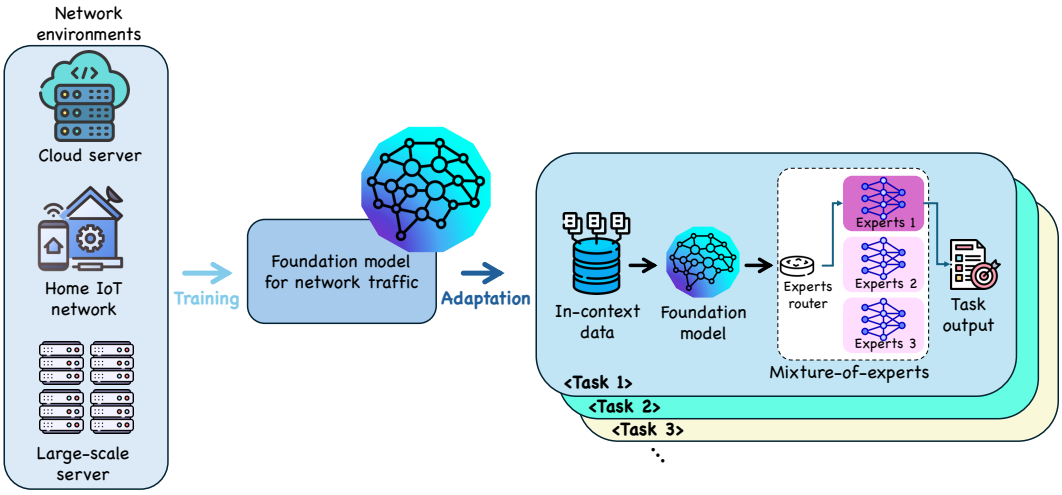

Figure 1: Overview of proposed framework. Data collected from diverse network environments is used to train a foundation model for network traffic. The trained foundation model is then adapted to downstream tasks, where adaptation leverages in-context examples and a MoE module with a learnable router to enhance task performance.

compatible foundation models for evolving data distributions and provides a step toward reliable, general-purpose network intelligence.

2 Proposed Method

Let $\mathcal{D}$ denote a large-scale network traffic dataset, composed of multiple subsets $\mathcal{D}_i$, each collected under a different environment $i \in \mathcal{I}$ (i.e., $\mathcal{D} = \bigcup_{i\in\mathcal{I}} \mathcal{D}_i$, where $\mathcal{D}_i \subseteq \mathcal{X} \times \mathcal{Y}$). Here, $\mathcal{X}$ is the input space of traffic sequences, and $\mathcal{Y}$ is the space of attack labels (e.g., "benign", "DDoS", "DoS"). Each $\mathcal{D}_i$ reflects traffic characteristics under specific collection conditions (e.g., network infrastructure). Each traffic sequence $x = (x_1, x_2, \ldots, x_L) \in \mathcal{X}$ is a multivariate time series of length L with C network traffic features (i.e., $x_t \in \mathbb{R}^C$).

Our first objective is to train a transformer-based foundation model $\mathcal{M}$ on these datasets, enabling it to learn universal, environment-invariant representations of network traffic. The second objective is to leverage the pretrained foundation model $\mathcal{M}$ for downstream tasks. We utilize the foundation model $\mathcal{M}$ as a general-purpose feature extractor and train a lightweight, task-specific model on top of its representations.

2.1 Tokenization of Network Traffic

To consider the temporal dependencies of network traffic data, we tokenize the x into N temporal patches of fixed length P, where $N = \lfloor \frac{L}{P} \rfloor$. Each patch $x_{:t} = \{x_{t-P+1}, ..., x_t\} \in \mathbb{R}^{P\times C}$ is linearly projected into an embedding vector $z_n \in \mathbb{R}^d$ via a 1D convolution, as follows.

$$z_n = W_{\text{token}} \cdot \text{vec}(x_{:n\cdot P}) + b_{\text{token}}, \quad n = 1, \ldots, N \tag{1}$$

Here, $W_{\text{token}} \in \mathbb{R}^{d\times(P\cdot C)}$ and $b_{\text{token}} \in \mathbb{R}^d$ are learnable parameters, and $\text{vec}(\cdot)$ denotes vectorization. A learnable positional embedding $e_n \in \mathbb{R}^d$ is then added to generate a token $h_n = z_n + e_n$. The sequence of tokens $\{h_n\}_{n=1}^N$ serves as the input to the attention block of $\mathcal{M}$. During training, a causal attention mask is applied so that each token attends only to its past and present tokens.

2.2 Training Foundation Model

The foundation model $\mathcal{M}$, parameterized by θ, is trained on $\mathcal{D}$ to capture both: (i) the reconstruction fidelity of benign traffic and (ii) discriminative structure across traffic types. Given an input $x \in \mathcal{X}$, tokenization yields $\{h_n\}_{n=1}^N$, and the foundation model outputs $\{\hat{h}_n\}_{n=1}^N = f_\theta(x)$.

Reconstruction Objective. For benign sequences, the model minimizes the discrepancy between reconstructed patches $\hat{h}_n$ and original embeddings h_n. We adopt a Huber loss with automatically chosen threshold $\delta(x)$, estimated via the median absolute deviation of residuals within each batch, where $\mathcal{D}_{\text{benign}} \subseteq \mathcal{D}$ denotes the set of benign samples.

$$\mathcal{L}_{\text{rec}}(\theta) = \frac{1}{|\mathcal{D}_{\text{benign}}|} \sum_{x \in \mathcal{D}_{\text{benign}}} \begin{cases} \frac{1}{2}\,(\hat{h}_n - h_n)^2, & \text{if } \left|\hat{h}_n - h_n\right| \le \delta(x), \\ \delta(x)\left(\left|\hat{h}_n - h_n\right| - \frac{1}{2}\,\delta(x)\right), & \text{otherwise.} \end{cases} \tag{2}$$

Contrastive Objective. To encourage discriminability across classes (benign and multiple attack types), we adopt a contrastive objective. Let $(\hat{h}_n, \hat{h}_m)$ denote a selected pair of reconstructed patches from original embeddings (h_n, h_m) with class labels c_n, c_m. Here, n is sequentially selected, while m is randomly sampled from remaining elements in $\{1, \ldots, N\}$, excluding n. The pairwise contrastive loss is defined as follows, where ξ is a margin hyperparameter.

$$\mathcal{L}_{\text{cont}}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \sum_{n=1}^{N} \mathbb{E}_{m \sim \mathcal{U}(\{1,\ldots,N\} \setminus \{n\})} \begin{cases} \|\hat{h}_n - \hat{h}_m\|_2, & \text{if } c_n = c_m, \\ \max(0,\, \xi - \|\hat{h}_n - \hat{h}_m\|_2), & \text{if } c_n \neq c_m. \end{cases} \tag{3}$$

Final Objective. The complete learning objective combines reconstruction and contrastive terms, where α and β control the relative importance of reconstruction and discriminative representation, as follows.

$$\mathcal{L}_{\text{FM}}(\theta) = \alpha\, \mathcal{L}_{\text{rec}}(\theta) + \beta\, \mathcal{L}_{\text{cont}}(\theta) \tag{4}$$

2.3 Downstream Task Adaptation Strategies

In-context Support within Environment. Given a query sample $(x, y) \in \mathcal{D}_i$, we construct the in-context set $\mathcal{S}_i(x)$ by drawing S examples from the same environment $\mathcal{D}_i$, as follows.

$$\mathcal{S}_i(x) = \{(x_s, y_s)\}_{s=1}^{S} \subset \mathcal{D}_i \setminus \{(x, y)\} \tag{5}$$

The foundation model then produces representations $\{\tilde{h}_n\}_{n=1}^{N} = f_\theta(x \mid \mathcal{S}_i(x))$, which conditioned on the environment in-context examples. These representations are subsequently passed into a task-specific expert E, which is adapted with $\tilde{h}$ for the target task. The overall training objective using these representations $\tilde{h}$ can be expressed as follows, where $\ell(\cdot)$ denotes task-specific loss.

$$\mathcal{L}_{\text{task}}(E) = \sum_{i \in \mathcal{I}} \mathbb{E}_{(x,y) \sim \mathcal{D}_i}\; \mathbb{E}_{\mathcal{S}_i(x)} \Big[\sum_{n=1}^{N} \ell\big(\tilde{h}_n, E\big) \Big] \tag{6}$$

Mixture-of-Experts with Router Learning. Since network traffic exhibits heterogeneity across environments and intrusion types, we extend the framework to employ multiple experts, $\{E_1, \ldots, E_K\}$, each of which can specialize in different traffic patterns. A router π assigns each input representation $\tilde{h}_n \in \mathbb{R}^d$ to a specific expert as follows, where $\phi \in \mathbb{R}^{K \times d}$ is a learnable parameter.

$$\pi_\phi(\cdot \mid \tilde{h}_n) = \text{softmax}(\phi \cdot \tilde{h}_n) \tag{7}$$

Here, the stochastic policy π_ϕ is optimized via reinforcement learning with state $\tilde{h}$ and action $a \in \{1, ..., K\}$. For each decision, the reward compares the chosen expert's loss $\ell\big(\tilde{h}_n, E_a\big)$ to the best-performing alternative, as follows.

$$R(\tilde{h}_n, a) = \frac{\min_k \ell(\tilde{h}_n, E_k) - \ell(\tilde{h}_n, E_a)}{\ell(\tilde{h}_n, E_a) + \varepsilon} + \zeta \tag{8}$$

In (8), a constant ζ encourages the overall reward to remain positive and stable, and ε is added for numerical stability. Since this setting corresponds to a single-step contextual bandit [12, 13], the router receives an immediate reward after each action, without multi-step dependencies. Thus, the objective reduces to maximizing the expected immediate reward via policy gradients, as follows.

$$\nabla_\phi J(\phi) = \mathbb{E}_{a \sim \pi_\phi(\cdot \mid \tilde{h}_n)} \Big[\Big(R(\tilde{h}_n, a) - b(\tilde{h}_n) \Big)\, \nabla_\phi \log \pi_\phi(a \mid \tilde{h}) \Big] \tag{9}$$

In (9), $b(\cdot)$ denotes the baselines computed as the expected reward over all experts, serving to reduce variance in policy gradient estimation. This update encourages the router to increase the probability of selecting experts that achieve higher relative performance, leading to improved expert assignment and overall model quality.

Table 1: Simulation Results of Intrusion Attack Type Classification across Datasets

Settings			Seen environment			Unsseen environment
	ICL	MoE	UNSW-NB15	BoT-IoT	CICIDS2018	ToN-IoT
case 1	✗	✗	34.17	45.93	46.25	40.68
case 2	✓	✗	41.19	47.73	47.64	45.36
case 3	✓	✓	50.51	50.62	50.66	48.23

3 Simulation

3.1 Simulation Setup

In this work, we evaluate our framework on one representative downstream task, the intrusion type classification task. We employ NF-UQ-NIDS-v3, which includes four widely used intrusion detection datasets—UNSW-NB15 [14], ToN-IoT [15], BoT-IoT [16], and CICIDS2018 [17]—into a NetFlow schema [18]. We designate three environments as seen environments (UNSW-NB15, BoT-IoT, and CICIDS2018) and one as an unseen environment (ToN-IoT). For each seen environment, we split the data into train and test with a ratio of $8:2$. For in-domain evaluation, we use the held-out test sets from the three seen environments. For out-of-domain evaluation, we use the entire unseen environment (ToN-IoT) as a test-only set. To measure the performance of the proposed method, we use classification accuracy.

We compare three configurations to quantify the contributions of ICL and MoE. A baseline model without either component (Case 1), a variant that incorporates only ICL (Case 2), and the proposed approach that combines both ICL and MoE (Case 3).

3.2 Simulation Results

Table 1 summarizes the classification accuracy across different datasets under three settings. Overall, the results demonstrate a consistent improvement when both ICL and MoE are applied jointly (case 3). Specifically, case 3 achieves the highest accuracy in both seen and unseen environments, demonstrating the complementary effect of ICL and MoE.

It is worth noting that the performance on UNSW-NB15 is relatively lower compared to the other datasets. This can be attributed to the fact that UNSW-NB15 contains only about one-tenth of the data volume compared to the other benchmarks, which limits the model's ability to generalize effectively in this setting. However, when both ICL and MoE are employed in the proposed framework, the performance gap between UNSW-NB15 and the other datasets becomes much smaller compared to the other cases. This indicates that the combination of ICL and MoE effectively compensates for the limited data volume in UNSW-NB15, leading to more consistent performance across heterogeneous datasets. These results highlight the importance of jointly leveraging ICL and MoE to improve generalization, even in data-scarce scenarios.

4 Conclusion and Discussion

Conclusion. In this paper, we propose a novel framework for adapting foundation models to network intrusion type classification by integrating ICL and a MoE architecture. Starting from a foundation model trained on diverse network environments, we demonstrated how the combination of ICL and MoE enables effective downstream adaptation while maintaining robustness across heterogeneous settings.

Discussion. This work opens several promising directions for future research. First, extending the category of downstream tasks to include traffic forecasting or resource management could further strengthen its practical benefits for maintaining reliable and efficient network operations in diverse real-world environments. Second, exploring more advanced optimization strategies for the router of MoE or incorporating compatibility-aware objectives may improve stability when adapting foundation models over time.

References

[1] Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, Jiayin Zhang, Juanzi Li, and Lei Hou. Benchmarking foundation models with language-model-as-an-examiner. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 78142–78167, 2023.

[2] Shashank Subramanian, Peter Harrington, Kurt Keutzer, Wahid Bhimji, Dmitriy Morozov, Michael W Mahoney, and Amir Gholami. Towards foundation models for scientific machine learning: Characterizing scaling and transfer behavior. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 71242–71262, 2023.

[3] Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan Ö. Arı k. Chain of agents: Large language models collaborating on long-context tasks. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 132208–132237, 2024.

[4] Julian Coda-Forno, Marcel Binz, Zeynep Akata, Matt Botvinick, Jane Wang, and Eric Schulz. Meta-in-context learning in large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 65189–65201. Curran Associates, Inc., 2023.

[5] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. A survey on in-context learning. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, 2024.

[6] Ruifeng Ren and Yong Liu. Towards understanding how transformers learn in-context through a representation learning lens. volume 37, pages 892–933, 2024.

[7] Kevin Christian Wibisono and Yixin Wang. From unstructured data to in-context learning: Exploring what tasks can be learned and when. *Advances in Neural Information Processing Systems*, 37:16369–16405, 2024.

[8] Jiahui Xu, Lu Sun, and Dengji Zhao. Mome: Mixture-of-masked-experts for efficient multi-task recommendation. In *Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2527–2531, 2024.

[9] Xun Wu, Shaohan Huang, Wenhui Wang, Shuming Ma, Li Dong, and Furu Wei. Multi-head mixture-of-experts. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 94073–94096. Curran Associates, Inc., 2024.

[10] Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. A survey on mixture of experts in large language models. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

[11] Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Yu Wu, et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*, 2024.

[12] Aldo Pacchiano, My Phan, Yasin Abbasi Yadkori, Anup Rao, Julian Zimmert, Tor Lattimore, and Csaba Szepesvari. Model selection in contextual stochastic bandit problems. *Advances in Neural Information Processing Systems*, 33:10328–10337, 2020.

[13] Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[14] Nour Moustafa and Jill Slay. Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set). In *military communications and information systems conference (MilCIS)*, pages 1–6. IEEE, 2015.

[15] Abdullah Alsaedi, Nour Moustafa, Zahir Tari, Abdun Mahmood, and Adnan Anwar. Ton_iot telemetry dataset: A new generation dataset of iot and iiot for data-driven intrusion detection systems. *Ieee Access*, 8:165130–165150, 2020.

[16] Nickolaos Koroniotis, Nour Moustafa, Elena Sitnikova, and Benjamin Turnbull. Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset. *Future Generation Computer Systems*, 100:779–796, 2019.

[17] Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorbani. Toward generating a new intrusion detection dataset and intrusion traffic characterization. In *Proceedings of International Conference on Information Systems Security and Privacy (ICISSP)*, pages 108–116, 2018.

[18] Majed Luay, Siamak Layeghy, Seyedehfaezeh Hosseininoorbin, Mohanad Sarhan, Nour Moustafa, and Marius Portmann. Temporal analysis of netflow datasets for network intrusion detection systems, 2025.