

# LLMs’ Suitability for Network Security: A Case Study of STRIDE Threat Modeling

Author’s draft for soliciting feedback - May 6, 2025

AbdulAziz AbdulGhaffar  
Dept. of Syst. & Comp. Engineering  
Carleton University  
abdulazizabdulghaff@gmail.carleton.ca

Ashraf Matrawy  
School of Information Technology  
Carleton University  
ashraf.matrawy@carleton.ca

**Abstract**—Artificial Intelligence (AI) is expected to be an integral part of next-generation AI-native 6G networks. With the prevalence of AI, researchers have identified numerous use cases of AI in network security. However, there are almost nonexistent studies that analyze the suitability of Large Language Models (LLMs) in network security. To fill this gap, we examine the suitability of LLMs in network security, particularly with the case study of STRIDE threat modeling. We utilize four prompting techniques with five LLMs to perform STRIDE classification of 5G threats. From our evaluation results, we point out key findings and detailed insights along with the explanation of the possible underlying factors influencing the behavior of LLMs in the modeling of certain threats. The numerical results and the insights support the necessity for adjusting and fine-tuning LLMs for network security use cases.

**Index Terms**—Large Language Model (LLM), STRIDE, threat modeling, suitability of LLM

## I. INTRODUCTION

Future networks, such as Sixth Generation (6G) networks, are envisioned to integrate Artificial Intelligence (AI) into their networks to be *AI-Native* networks [1] to improve performance, efficiency, and scalability [2]. Ericsson’s report [3] indicates that deploying AI in telecom networks will not only reduce the Operational Expenditure (OPEX) of the network but also provide a 5% to 10% return on investment. On the other hand, with the increasing popularity of AI and Large Language Models (LLMs), researchers are identifying potential applications and use cases of AI and LLMs in networks [4]–[6]. These potential use cases include, but are not limited to, network optimization [4], automation of security [5], and threat classification [6].

Upon examining the literature, we notice a significant gap where there is a lack of work analyzing and investigating the suitability of LLMs in the proposed network security use cases. This motivated us to investigate the suitability of LLMs in network security use cases. Due to the importance of threat modeling as a starting point in any security exercise, we focus on the “*Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, and Elevation of Privilege (STRIDE)*” threat model [7], [8].

We have extensive experience with Fifth Generation (5G) threat modeling using STRIDE [9], [10]. Hence, in this work,

we select a case study of STRIDE threat modeling to perform LLM-based classification of 5G threats. We perform the experiments using four prompting techniques with five different LLMs. This work is important as it provides insights on using LLMs for threat classification in next-generation ‘*AI-Native*’ telecom networks.

The main contributions of this work are as follows:

- 1) We investigate the suitability of Large Language Models (LLMs) in network security use cases. For this purpose, we select the case study of STRIDE threat modeling of 5G threats and vulnerabilities. We perform experiments by selecting six 5G threats and utilizing four different prompting techniques with five LLMs.
- 2) We provide detailed insights based on the evaluation results of LLM-based STRIDE classification. We provide detailed discussions on potential underlying factors that influence the behavior of LLMs in modeling certain threats, including: incorrect threat perspective, failure to identify second-order threats, and insights on Few-Shot (FS) prompting positively impacting performance.
- 3) We analyze the suitability of LLMs using numerical testing and various performance metrics, including accuracy, precision, recall, and F1 score. Our results indicate that the performance of the selected LLMs is comparable, highlighting the need for enhancements in these models for STRIDE threat modeling in 5G networks.

The paper is organized as follows: Section II provides the motivation of our work in light of the examined related works. Section III explains the evaluation methodology we use for the STRIDE threat modeling case study. The detailed results and insights of the evaluation are presented in Section IV. Finally, Section V provides the discussion and conclusion of our study.

## II. MOTIVATION AND EXAMINATION OF RELATED WORK

With the advent of LLMs, many researchers have identified various potential use cases of LLMs in telecom networks and cybersecurity. We present the most relevant papers in this section.

**LLM for Networking:** The white paper by Shahid *et al.* [4] presents the concept of “Large Telecom Model (LTM)” for use cases of telecom networks. Some of the potential use

cases they mention include the use of LTM at the network edge, LTM for network optimization, using LTM for network automation tasks, and performing intent-based management using LTMs, etc. Similarly, Zhou *et al.* [5] identify four different areas of telecom networks that may benefit from LLMs. These areas include generation, optimization, classification, and prediction problems in telecom networks. Wu *et al.* [11] present the NetLLM method to utilize LLMs for three different networking problems, namely, “viewport prediction”, “adaptive bitrate streaming”, and “cluster job scheduling”. The authors extensively evaluate the performance of their proposed framework within these problems.

**LLM for Security:** Ferrag *et al.* [6] propose multiple applications of LLMs in cybersecurity, including detection and analysis of threats, incident response, automation of security tasks, verification of security protocols, etc. The work by Guthula *et al.* [12] proposes a foundation model for security that takes into account the distinct nature of network traffic. The aim of the authors is to consider the general applicability of the model.

Sędkowski [13] studies the efficacy of LLMs in recognizing potential threats in the network and recommending countermeasures. Their methodology includes using Nmap reports to classify threats using STRIDE threat modeling with three different LLMs. The author concludes that the performance of LLMs in threat detection is comparable to humans. The scope of their work is focused on testing the application of AI in threat modeling, as opposed to our aim of studying the suitability of LLMs for network security.

Yang *et al.* [14] present “ThreatModeling-LLM”, which is an LLM-based method to perform threat modeling of the banking system using the STRIDE model. Their approach includes various steps to improve the performance of the LLMs. Their scope is limited to banking systems. The main objective of their work is to improve and automate threat modeling using LLMs, instead of investigating the suitability of LLMs for network security.

**Motivation:** After examining the related works, we identify that the suitability of the potential use cases of LLMs needs to be investigated. However, to the best of our knowledge, there are almost nonexistent or very scarce existing studies that are actually testing the suitability of LLMs for these use cases. This motivated us to study the suitability of the LLMs in telecom network security using a case study of STRIDE threat modeling.

### III. EVALUATION METHODOLOGY FOR A CASE STUDY OF STRIDE THREAT MODELING

The aim of this case study is to employ LLMs to categorize the threats and vulnerabilities on 5G interfaces using the STRIDE model. The main objective is to evaluate various prompting and search techniques and investigate the suitability of LLMs for telecommunication tasks.

Our evaluation methodology is shown in Figure 1. We initially select multiple threats and vulnerabilities in the 5G network along with their baseline STRIDE classifications from

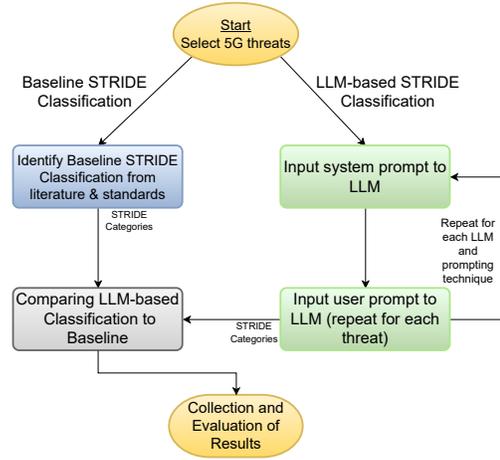


Fig. 1. Evaluation Methodology of STRIDE Modeling

the published literature and standards. Then, we use various prompting techniques to perform the LLM-based STRIDE classification of the selected threats. Finally, we compare the STRIDE classification by LLMs to the baseline and evaluate the results. We will explain each step in the following.

#### A. Selected 5G Threats and Vulnerabilities

To carry out the evaluation, we select six threats and vulnerabilities on 5G interfaces from our previous research work [10] along with their STRIDE classifications as a baseline; this step is shown with a blue block in Figure 1. As we already mentioned in our previous work [10], we want to clarify that the baseline STRIDE classification of the threats may not be unique. We randomly select these attacks from the published literature and standards and ensure that the selected threats cover the six STRIDE categories. Three of the selected threats are on the N1 interface, because N1 is exposed to the Radio Access Network (RAN) and it faces the largest number of threats [10]. In our earlier work [10], we comprehensively explored and identified the threats and vulnerabilities on the critical 5G network interfaces and categorized them based on the STRIDE threat model. The selected threats, along with their description, are explained below. We use the same threat names in the first row of Table I:

- *Access and Mobility Function (AMF) Impersonation on N1 interface:* If a malicious actor is impersonating AMF, it can access sensitive user information through the N1 interface [15], [16]. This is especially important when users send their unique identifiers (e.g., Subscription Permanent Identifier (SUPI)) to the AMF to join the 5G network [10].
- *5G-Globally Unique Temporary Identity (GUTI) and International Mobile Equipment Identity (IMEI) correlation on N1 interface:* If the attacker is able to correlate 5G-GUTI and IMEI of a user, it can trace the present and future mobility and position of the user [10], [17].
- *Bidding down on Xn-handover:* In this threat, insecure algorithms are enforced by the malicious gNodeB (gNB)

```

You are a 5G network security expert.
Your task is to classify a given 5G threat or
vulnerability according to the STRIDE model:
1. Spoofing
2. Tampering
3. Repudiation
4. Information disclosure
5. Denial of service
6. Elevation of privilege
Each threat or vulnerability may belong to one or
more STRIDE categories. Your response should list
only the applicable category or categories without
any additional details or explanations.

```

Listing 1. LLM System Prompt. The text in **Blue** color defines the scope of the LLM task. While the text in **Cyan** color refines the output of the LLM

in the 5G system, resulting in weakening the security of the 5G system [10], [16].

- *Eavesdropping on F1 interface*: On the F1 interface, eavesdropping of the control plane and data plane traffic is a potential threat [10]. This eavesdropping will result in information disclosure and can further lead to threats that can allow spoofing and tampering as well [16], [18].
- *False Single Network Slice Selection Assistance Information (S-NSSAI) on N1 interface*: Providing incorrect S-NSSAI during the Network Slice-Specific Authentication and Authorization (NSSAA) procedure threatens systems resources and may result in escalation of privileges [15], [16].
- *Man-in-The-Middle (MiTM) attack on N3 interface*: The N3 interface between 5G RAN and User Plane Function (UPF) is susceptible to MiTM attack [10].

## B. Large-Language Models (LLMs)

We select the following Large Language Models (LLMs) to perform this evaluation: Sonar by Perplexity [19], GPT-4o by OpenAI [20], Claude 3.7 Sonnet by Anthropic [21], Grok-2 by xAI [22], and Gemini 2.5 Pro by Google [23]. We use these LLMs through the pro version of the Perplexity AI platform that we have access to through our University [24]. We perform the STRIDE classification of the six selected threats using these LLMs in order to evaluate the suitability of the LLMs for network security. Figure 1 shows the LLM-based STRIDE classification methodology with green blocks.

## C. LLM Prompts

We use a combination of system and user prompts to perform the LLM-based STRIDE classification.

**System Prompt:** The experiment is performed by providing a system prompt to the LLM at the beginning, which is an instruction to define the scope of the LLM task and control the output of the LLM [25]. The system prompt we provide to the LLMs is shown in Listing 1. We initially define the **scope** of the LLM task and then outline the instructions to refine the **output** of the LLM.

**User Prompts:** The user prompts are a set of prompts that we run for each of the selected six threats. We use the following two prompting approaches in our evaluation:

```

Classify the following threat/vulnerability:
[NAME_OF_THREAT]

```

Listing 2. User prompt for Zero-Shot (ZS) prompting

```

Classify the following threat/vulnerability:
[NAME_OF_THREAT]
The following are some examples of threat STRIDE
classification. Here, {X} represents that the
threat does not belong to this category, and {O}
means the threat belongs to this category:
1. NAS protocol-based attack on N1 interface:
S{X}, T{X}, R{X}, I{O}, D{O}, E{X}
2. A bidding down of Security features on N1
interface: S{X}, T{O}, R{X}, I{O}, D{O}, E{X}
3. Keystream reuse on Xn interface: S{X}, T{
X}, R{X}, I{O}, D{X}, E{X}
4. Flawed Validation of Client Credentials
Assertion on SBI interface: S{O}, T{X}, R{X}, I{O
}, D{O}, E{O}

```

Listing 3. User prompt for Few-Shot (FS) prompting

- 1) **Zero-Shot (ZS) prompting:** The Zero-Shot (ZS) prompting approach only provides the LLM with a description of the task without including any examples in the prompt [26]. The ZS prompt we use in this case study is shown in Listing 2.
- 2) **Few-Shot (FS) prompting:** This approach of prompting the LLM includes a certain number of examples in the prompt [26]. The FS prompt we use is shown in Listing 3.

We replace the **'[NAME\_OF\_THREAT]'** in these prompts with the name of a specific threat and provide the prompt to the LLM. Then, we record the STRIDE classification of a threat provided by the LLM. We repeat this step for all the six selected threats (see steps in green blocks in Figure 1).

We combine ZS and FS prompting approaches with 'base' LLM knowledge (no internet access) and with 'internet' search, to come up with four prompting techniques, *ZS\_Base*, *ZS\_Internet*, *FS\_Base*, and *FS\_Internet*. Apart from Zero-shot (ZS) and Few-shot (FS) promptings, another approach that is generally used to redefine the scope of the LLMs is 'fine-tuning' [26]. However, fine-tuning requires retraining the LLM with a specific dataset in order to refine its output for a particular use case. Since this approach is expensive (in terms of time and resources), we do not consider LLM fine-tuning for the evaluation in this work.

## IV. RESULTS AND INSIGHTS

While we report some performance metrics on how the LLMs performed in our experiments, we note that our main goal is not to compare the LLMs but rather to study the suitability of these models for network security tasks using a case study of 5G STRIDE modeling. Most importantly, we provide insights on their use in the modeling process.

### A. LLM-based STRIDE Classification

In this section, we provide the results of evaluating the case study of STRIDE threat modeling of 5G threats. The results



‘second-order threat’ of this attack. However, as specified by 3GPP [16], [18], due to the lack of confidentiality and integrity measures, the eavesdropping threat may not only result in ‘information disclosure’ but may also result in ‘spoofing’ and ‘tampering’ threats as well. This shows that LLMs may not always provide a comprehensive threat modeling, specifically when multiple subsequent threats are also possible.

We further see similar behavior in the second selected threat, “5G-GUTI and IMEI correlation on N1 interface”, where the threat is identified correctly as ‘information disclosure’. This ‘information disclosure’ can further lead to ‘tampering’, but it is not categorized as a ‘tampering’ threat by the LLMs. On the positive side, the classification performance of the second threat is very consistent across all prompting techniques and LLMs, and it is slightly improved as we move from the *ZS\_Base* to *FS\_Internet* prompting techniques.

For the “MiTM attack on the N3 interface” (sixth threat), the LLMs mostly identified this threat correctly in the ‘spoofing’, ‘tampering’, and ‘information disclosure’ categories. However, they predominantly failed to identify the MiTM threat in the ‘repudiation’ category. This is similar to the previous results of the fourth threat, where the LLMs did not identify some categories when multiple subsequent threats are also possible. In the case of a MiTM attack, for example, it is possible that an attacker may intercept and modify the content of a packet in transit from the sender to the receiver, if appropriate security measures are not provided. The sender will deny the transmission of the modified content. Nonetheless, due to the lack of security measures, it will be difficult to identify the packet modification or the entity responsible for the modification.

**FS Prompting Improves Performance:** We discover that the classification performance of the third threat, “Bidding down on Xn-handover”, is mostly accurate in all prompting techniques, except *ZS\_Base*. For example, the *FS\_Base* prompting approach achieved an accuracy of 100% with all LLMs. This can be due to the fact that the user prompt we provide for FS prompting includes a similar example of a ‘bidding down of Security features on N1 interface’ with the same STRIDE classification as this threat. We also notice that the performance is increased in FS prompting compared to ZS prompting, similar to the first and second threats.

Similarly, we notice from the first threat that the classification performance is improved as we move from *ZS\_Base* to *FS\_Internet* prompting technique (more green and less red as we move down the ‘I’ column). This also suggests that providing examples to the LLMs improves their performance.

### C. Comparison of Prompting Techniques

We analyze and compare the performance of the prompting techniques in terms of accuracy and F1 score. The results in Figures 2 and 3 are averaged over the six selected threats (36 cells in one row). Figure 2 shows the accuracy of LLMs in 5G STRIDE threat modeling with four different prompting techniques. We see from the figure that with *ZS\_Base* prompting, the accuracy achieved is the lowest compared to other

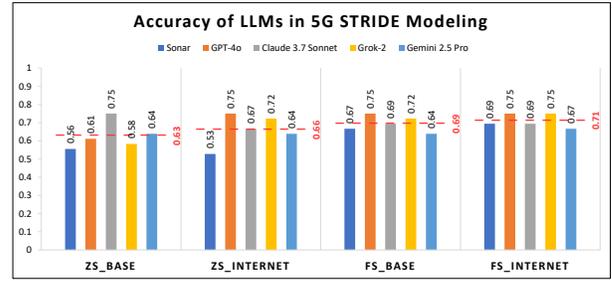


Fig. 2. Accuracy of LLMs in 5G STRIDE Modeling using different prompting techniques

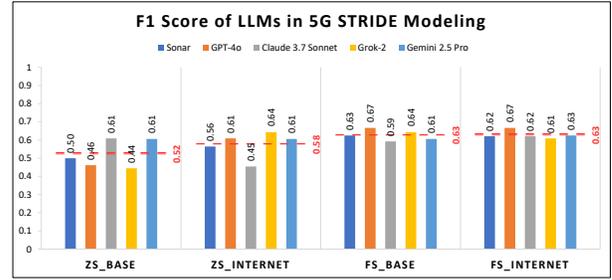


Fig. 3. F1 Score of LLMs in 5G STRIDE Modeling using different prompting techniques

prompting techniques. As we go from *ZS\_Base* prompting to *FS\_Internet* prompting, we observe that the performance (accuracy) is increasing gradually. We plot the average accuracy of each prompting technique and notice that as we move from *ZS\_Base* to *FS\_Internet*, the average accuracy of LLMs in 5G STRIDE threat modeling is increasing from 63% to 71%. These results are in accordance with the evaluation performed by Brown *et al.* [26] and their conclusion that providing examples in the prompts improves the performance.

The F1 score of the LLMs with different prompting techniques is shown in Figure 3. With the *ZS\_Base* prompting, we observe that the lowest average F1 score is recorded with 52%, as compared to the other prompting techniques. We notice that as we go from *ZS\_Base* to *FS\_Internet* prompting technique, the F1 score of the LLMs increases slightly. This is evident from the average line (shown in red), which shows a 10% increase in the average. This slight increase indicates that the FS prompting techniques improve the performance. On the other hand, the performance of the LLMs relative to each other is almost similar and shows no significant difference.

### D. Performance of LLMs in STRIDE Classification

The results show that LLMs’ performance in our experiments is mostly comparable. Figure 4 illustrates the performance of the LLMs in terms of accuracy, precision, recall, and F1 score using a heatmap chart. This result is averaged across all threats and all prompting techniques for a specific LLM. We observe that GPT-4o, Claude 3.7 Sonnet, and Grok-2 show ‘relatively’ higher accuracy and precision but lower recall compared to the other two LLMs. On the other hand, Sonar and Gemini 2.5 Pro achieved lower precision and



Fig. 4. Heatmap showing the performance of LLMs in terms of accuracy, precision, recall, and F1 score. Higher values (dark green color) are better.

higher recall in comparison. The F1 score indicates that the performance of all the LLMs is comparable for this case study. The maximum accuracy achieved is 72%, which highlights that there are opportunities for improvement across all LLMs, perhaps by fine-tuning for the specific application of STRIDE threat modeling in 5G networks.

We note that several articles highlight the prominent challenges and issues with the LLMs [27], [28]. The most relevant challenges to threat modeling include incorrect predictions and LLM hallucinations, which may result in disregarding required countermeasures or implementing unnecessary security measures. Secondly, the adaptability of LLMs for telecom-specific threats and vulnerabilities. Thirdly, improving LLM inference speed in networks to enable rapid threat modeling of the detected threats.

## V. DISCUSSION AND CONCLUSION

In this work, we explore and investigate the suitability of LLMs for network threat modeling. To perform this analysis, we select the case study of STRIDE threat modeling to perform LLM-based classification of 5G threats according to the STRIDE threat model. We observe from our evaluation that providing examples to the LLMs using FS prompting improves their performance. We further notice that LLMs may not always consider the threat classifications as a result of the second-order effect. This will limit the threat identification and may eventually result in not identifying all the possible risks associated with a threat. We hope these insights and results are the starting points to encourage research into fine-tuning LLMs on telecom-specific datasets and to enhance their performance in network security tasks. This is particularly important for future ‘AI-native’ networks, where AI needs to detect and identify threats autonomously and with the highest accuracy.

## ACKNOWLEDGMENT

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC).

## REFERENCES

- [1] Ericsson, “Defining AI native: A key enabler for advanced intelligent telecom networks,” Tech. Rep., 2023. [Online]. Available: <https://www.ericsson.com/49341a/assets/local/reports-papers/white-papers/ai-native.pdf>
- [2] C. K. Thomas *et al.*, “Causal reasoning: Charting a revolutionary course for next-generation ai-native wireless networks,” *IEEE Vehicular Technology Magazine*, 2024.
- [3] Ericsson, “AI business potential: understanding the value of AI for telecom operations,” Tech. Rep., 2022. [Online]. Available: <https://www.ericsson.com/en/reports-and-papers/further-insights/ai-business-potential>
- [4] A. Shahid *et al.*, “Large-scale ai in telecom: Charting the roadmap for innovation, scalability, and enhanced digital experiences,” *arXiv preprint arXiv:2503.04184*, 2025.
- [5] H. Zhou *et al.*, “Large language model (llm) for telecommunications: A comprehensive survey on principles, key techniques, and opportunities,” *IEEE Communications Surveys & Tutorials*, 2024.
- [6] M. A. Ferrag *et al.*, “Generative ai in cybersecurity: A comprehensive review of llm applications and vulnerabilities,” *Internet of Things and Cyber-Physical Systems*, 2025.
- [7] Microsoft, “The stride threat model.” [Online]. Available: [http://msdn.microsoft.com/en-us/library/ee823878\(v=cs.20\).aspx](http://msdn.microsoft.com/en-us/library/ee823878(v=cs.20).aspx)
- [8] L. Kohnfelder *et al.*, “The threats to our products,” *Microsoft Interface, Microsoft Corporation*, vol. 33, 1999.
- [9] D. Sattar *et al.*, “A stride threat model for 5g core slicing,” in *2021 IEEE 4th 5G World Forum (5GWF)*. IEEE, 2021, pp. 247–252.
- [10] M. Mahyoub *et al.*, “Security analysis of critical 5g interfaces,” *IEEE Communications Surveys & Tutorials*, 2024.
- [11] D. Wu *et al.*, “Netllm: Adapting large language models for networking,” in *Proceedings of the ACM SIGCOMM 2024 Conference*, 2024, pp. 661–678.
- [12] S. Guthula *et al.*, “netFound: Foundation model for network security,” *arXiv preprint arXiv:2310.17025*, 2023.
- [13] W. Sędkowski, “Threat Identification using STRIDE and GPT based chatbots,” *Studia Społeczne*, vol. 46, no. 3, pp. 75–86, 2024.
- [14] S. Yang *et al.*, “ThreatModeling-LLM: Automating Threat Modeling using Large Language Models for Banking System,” *arXiv preprint arXiv:2411.17058*, 2024.
- [15] 3rd Generation Partnership Project (3GPP), “Security architecture and procedures for 5G system,” Tech. Rep., TS 33.501, Release 19, 2025, version 19.2.0.
- [16] —, “Security Assurance Specification (SCAS) threats and critical assets in 3GPP network product classes,” Tech. Rep., TS 33.926, Release 19, 2025, version 19.3.0.
- [17] M. Bartock *et al.*, “5G Cybersecurity,” National Institute of Standards and Technology, NIST Special Publication 800-33B, 2022. [Online]. Available: <https://www.nccoe.nist.gov/sites/default/files/2022-04/nist-5G-sp1800-33b-preliminary-draft.pdf>
- [18] 3rd Generation Partnership Project (3GPP), “Study on Security for Next Radio (NR) Integrated Access and Backhaul (IAB) (Release 17),” Tech. Rep., TS 33.824, Release 17, 2022, version 17.0.0.
- [19] “Sonar by Perplexity,” Mar. 2025. [Online]. Available: <https://sonar.perplexity.ai/>
- [20] “Hello GPT-4o | OpenAI,” Mar. 2025. [Online]. Available: <https://openai.com/index/hello-gpt-4o/>
- [21] “Claude 3.7 Sonnet \Anthropic,” Mar. 2025. [Online]. Available: <https://www.anthropic.com/claude/sonnet>
- [22] “Grok-2 Beta Release | xAI,” Mar. 2025. [Online]. Available: <https://x.ai/news/grok-2>
- [23] “Gemini 2.5: Our newest Gemini model with thinking,” Mar. 2025. [Online]. Available: <https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/#gemini-2-5-pro>
- [24] “Perplexity AI,” Mar. 2025. [Online]. Available: <https://www.perplexity.ai/>
- [25] B. Hui *et al.*, “Pleak: Prompt leaking attacks against large language model applications,” in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 2024, pp. 3600–3614.
- [26] T. Brown *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [27] Y. Huang *et al.*, “Large language models for networking: Applications, enabling techniques, and challenges,” *IEEE Network*, 2024.
- [28] G. O. Boateng *et al.*, “A survey on large language models for communication, network, and service management: Application insights, challenges, and future directions,” *IEEE Communications Surveys & Tutorials*, 2025.