Beyond the Tip of Efficiency: Uncovering the Submerged Threats of Jailbreak Attacks in Small Language Models

Anonymous ACL submission

Abstract

Small language models (SLMs) have become increasingly prominent in the deployment on edge devices due to their high efficiency and low computational cost. While researchers continue to advance the capabilities of SLMs through innovative training strategies and model compression techniques, the security risks of SLMs have received considerably less attention compared to large language models (LLMs). To fill this gap, we provide a comprehensive empirical study to evaluate the security performance of 13 state-of-the-art SLMs under various jailbreak attacks. Our experiments demonstrate that most SLMs are quite susceptible to existing jailbreak attacks, while some of them are even vulnerable to direct harmful prompts. To address the safety concerns, we evaluate several representative defense methods and demonstrate their effectiveness in enhancing the security of SLMs. We further analyze the potential security degradation caused by different SLM techniques including architecture compression, quantization, knowledge distillation, and so on. We expect that our research can highlight the security challenges of SLMs and provide valuable insights to future work in developing more robust and secure SLMs.

1 Introduction

Large language models (LLMs), such as ChatGPT (Brown et al., 2020; Ouyang et al., 2022a; Achiam et al., 2023) and Llama series (Touvron et al., 2023a; Dubey et al., 2024), have demonstrated revolutionary performance in a spectrum of text generation tasks. As a fundamental principle for guiding the development of LLMs, the scaling law (Kaplan et al., 2020) highlights the strong correlation between the performance and scale of LLMs. However, as LLMs evolve to encompass hundreds and even thousands of billions of parameters, their development imposes expensive demands on computational resources and high-quality data for pre-training. Consequently, LLMs are typically confined to deployment on GPU clusters and cloud environments, posing significant challenges for wide adoption on edge devices such as smartphones, laptops, autonomous vehicles, and wearables.

Recently, small language models (SLMs) have attracted significant attention from the academic community for their efficiency and remarkable performance in various tasks (Lu et al., 2024; Nguyen et al., 2024). On platforms like Hugging Face, SLM collections such as Llama-3.2 (Dubey et al., 2024), MiniCPM (Hu et al., 2024) and Phi (Abdin et al., 2024) have gained considerable popularity among researchers and achieved top-tier download rates. Different from LLMs, SLMs typically consist of only a few billion parameters, requiring significantly less training data and computational cost for deployment.

However, unlike LLMs that benefit from extensive datasets and robust alignment strategies, it is challenging for SLMs to balance between generation capabilities and security, which makes them more vulnerable to jailbreak attacks. Among these, one of the most serious threats is referred to as jailbreak. By creating malicious prompts to induce target LLMs to generate harmful responses, jailbreak has emerged as a critical security concern in the development of LLMs (Yi et al., 2024; Yao et al., 2024; Gupta et al., 2023). Moreover, certain jailbreak techniques that can bypass the security boundary of LLMs have demonstrated strong transferability to other models (Zou et al., 2023), which presents potential threats to all generative models including SLMs.

Although security concerns regarding SLMs have become an increasingly important issue, there still remains a substantial gap in exploring and un-

*Corresponding author.

derstanding the security boundary of SLMs. In this paper, we collect representative malicious datasets, jailbreak attack methods, and defense methods to conduct adversarial experiments on numerous state-of-the-art SLMs, thereby revealing existing security vulnerabilities of SLMs and exploring corresponding mitigation strategies. Furthermore, we take insight into the security degradation of SLMs and discuss some potential factors. In summary, we make the following contributions:

- We conduct extensive experiments to reveal the security vulnerabilities of SLMs under different jailbreak attacks. Especially, The results demonstrate that most SLMs are more susceptible to jailbreak attacks compared to LLMs.
- We evaluate the effectiveness of existing defense methods on SLMs. The results show that these methods are significantly adapted to SLMs to enhance their resilience against jailbreak attacks.
- We discuss and analyze various underlying factors that may lead to the security degradation of SLMs, including inadequate safety alignment, biased knowledge distillation, parameter sharing, and quantization techniques.

2 Related Work

2.1 Jailbreak Attacks

Jailbreak attacks, which transform harmful queries like "How to make a bomb" into more sophisticated prompts to deceive target models to generate toxic output, can be mainly classified into two categories: white-box methods and black-box methods.

White-box methods generally rely on access to the internal states of LLMs to design attack strategies. These methods generally use gradients and logits of target LLMs as loss functions to optimize adversarial suffixes appended to malicious questions (Zou et al., 2023; Jones et al., 2023; Zhu et al., 2023; Andriushchenko et al., 2024; Geisler et al., 2024; Mangaokar et al., 2024), or manipulate the output logits to enforce target LLMs to generate affirmative responses (Huang et al., 2024; Zhang et al., 2024a). However, white-box methods tend to generate irregular prompts that are easily detectable and cannot be optimized directly on black-box models like ChatGPT.

In contrast, black-box methods construct readable prompts in different ways and validate their effectiveness based on the responses of the target LLMs. Some studies employ heuristic strategies to rewrite malicious questions in other formats such as ASCII format (Jiang et al., 2024), code format (Kang et al., 2024; Lv et al., 2024), encrypted format (Yuan et al., 2024; Liu et al., 2024a) and low-resource languages (Deng et al., 2024b), exploiting the insufficient safety alignment of target LLMs in these formats to bypass the defense mechanism. Another line of research is to instruct an advanced LLM like GPT-4 to optimize jailbreak prompts by incorporating iterative refinement (Jin et al., 2024), genetic algorithms (Liu et al., 2024b) and psychological expertise (Zeng et al., 2024), or fine-tune another LLM with successful jailbreak templates to serve as an attacker to generate jailbreak prompts automatically (Deng et al., 2024a; Ge et al., 2024). Compared with white-box methods, black-box methods can be applied to most models. For this reason, black-box methods are widely used in empirical experiments to evaluate the safety of LLMs.

To mitigate threats caused by jailbreak attacks, different defense techniques are proposed to ensure the security of LLMs. One line of work addresses the issue by detecting (Inan et al., 2023) or perturbing (Robey et al., 2023) the jailbreak prompts to reduce the toxicity of the input, while another line of work directly enhances the robustness of LLMs by supervised fine-tuning (Bianchi et al., 2024) or reinforcement learning from human feedback (Ouyang et al., 2022b).

2.2 Small Language Models

Similar to LLMs, SLMs are typically built upon decoder-only architectures, while they show diversity in implementation details such as the type of attention heads, layer numbers, dimension sizes, activation functions, and so on.

To achieve competitive performance within the limited scale of SLMs, different model compression techniques are adopted to construct lightweight architectures efficiently. For instance, MobilLLaMA (Thawakar et al., 2024) and MobileLLM (Chu et al., 2023) introduce a parameter-sharing scheme in embedding blocks and attention head blocks to reduce the cost of GPU memory. TinyLLaMA (Zhang et al., 2024b) optimizes memory load with the FlashAttention technique (Dao et al., 2022), which introduces an IO-aware attention algorithm to reduce the budget of high bandwidth memory. Quantization techniques, such as

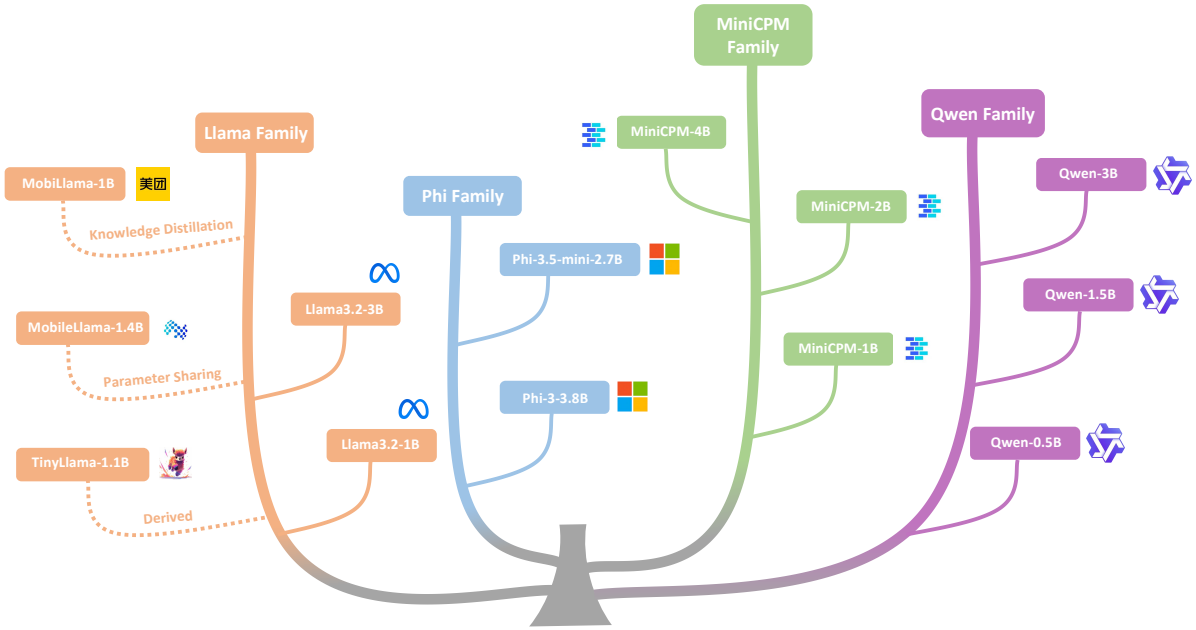

Figure 1: The family tree of the target SLMs we evaluate in our paper. The solid line represents the model is belonging to a certain family, while the dashed line indicates that the model is derived from that family with certain SLM technology.

GPTQ (Frantar et al., 2022) and AWQ (Lin et al., 2024), can also effectively reduce memory loads by compressing the size parameters from 16 bits to 8 bits or even 4 bits. In model collections such as Llama 3 (Touvron et al., 2023a), Qwen (Bai et al., 2023), and MiniCPM (Hu et al., 2024), SLMs are generally designed and pre-trained following LLMs in the same family. Additionally, during the training phase, knowledge distillation techniques are widely used to derive performance from teacher LLMs to student SLMs, as models in the same family generally share similar tokenizers and architecture.

Recent research has demonstrated that SLMs can achieve comparable performance in some reasoning tasks, and can even outperform LLMs in specific scenarios (Lu et al., 2024). Our study fills a gap in evaluating the security of SLMs from another perspective. In the following sections, we will demonstrate the security differences between various SLMs and delve into their underlying causes.

3 Experiment Setups

3.1 Target Models

We collect 16 state-of-the-art models to provide a comprehensive view of their security differences, including 13 SLMs below 4B size and 3 LLMs above 7B size.

For LLMs, we include Llama2-7B (Touvron et al., 2023b), Llama3-8B (Touvron et al., 2023a), and DeepSeek-R1-Distill-Llama-8B (Guo et al., 2025) for evaluation. Notably, DeepSeek-R1-Distill-Llama-8B is constructed by distilling the reasoning patterns from DeepSeek-R1 to Llama3-8B. The controlled comparison enables us gain some valuable insights of the influences of distillation techniques to SLM security. For SLMs, we include 13 models from advanced research organizations and individual developers. Specifically, they are as follows:

- **Llama Family**. Llama family is developed by Meta AI as one of the most popular model series. For our study, we select two models from the Llama 3.2 collection with parameter sizes of 1B and 3B. Furthermore, we include three additional SLMs that are initialized from Llama and processed with certain model compression techniques. These models are MobilLLaMA (Thawakar et al., 2024), MobileLLM (Chu et al., 2023), and TinyLLaMA (Zhang et al., 2024b).

- **Phi Family**. Phi family (Gunasekar et al., 2023) is developed by Microsoft focusing on designing lightweight SLMs with exceptional performance. We select Phi-3-mini-4k-instruct in 3.8B size and Phi-3.5-mini-instruct size in 2.7B size for evaluation.

- **MiniCPM Family**. MiniCPM family (Hu et al., 2024) is developed by OpenBMB and mainly consists of SLMs in different versions. We select MiniCPM-1B-sft-bf16, MiniCPM-2B-sft-bf16 and MiniCPM-4B for evaluation.

- **Qwen Family**. Qwen family (Bai et al., 2023) is built by Alibaba Cloud which has released a spectrum of LLMs ranging from 0.5B to 72B sizes. We select Qwen2.5-0.5B-Instruct, Qwen2.5-1.5B-Instruct, and Qwen2.5-3B-Instruct for evaluation.

3.2 Attack Methods

Our research firstly examines the influences of direct attacks against SLMs, which leverage straightforward harmful queries such as "How to make a bomb" to probe the target models directly. We collect 5 datasets that contain such harmful questions across various illegal and unethical dimensions.

- **Advbench**. Advbench (Zou et al., 2023) is a harmful dataset consisting of 500 harmful strings and 500 harmful behaviors. The former focuses on eliciting specific harmful responses from target LLMs, while the latter aims at provoking the models into exhibiting harmful behavior as much as possible.

- **DAN**. DAN (Shen et al., 2023) provides a forbidden question set spanning 13 restricted scenarios. The dataset is primarily sourced from online platforms and publicly available datasets.

- **maliciousInstruct**. MaliciousInstruct (Huang et al., 2024) consists of 100 harmful questions with 10 malicious intentions, which are mostly generated by ChatGPT and then revised manually.

- **StrongREJECT**. StrongREJECT (Souly et al., 2024) offers 313 harmful questions that cover forbidden scenarios from different AI usage policies. The majority of the dataset is written manually, while the remaining portion is sourced from LLMs and other open-source datasets.

- **XSTEST**. XSTEST (Röttger et al., 2024) contains both safe and unsafe questions to assess the exaggerated safety behaviors of models. We extract the 200 harmful questions from the dataset for our experiments.

Furthermore, to explore and understand the safety boundary of SLMs more clearly, we conduct a thorough investigation into existing jailbreak attacks and select 5 representative methods for evaluation. These methods span across different categories of jailbreak attacks and have demonstrated excellent effectiveness against LLMs in previous studies.

- **GCG**. Greedy Coordinate Gradient (GCG) (Zou et al., 2023) is a gradient-based attack that initializes an adversarial suffix appended to the malicious question and optimizes it by gradient-based search to maximize the probability of affirmative responses. Although the optimization of jailbreak prompts is constrained to white-box models, they demonstrate strong transferability to other black-box models.

- **ArtPrompt**. ArtPrompt (Jiang et al., 2024) is an ASCII-based attack that leverages the poor performance of LLMs in recognizing ASCII art to bypass defense mechanisms. Specifically, ArtPrompt utilizes LLMs like GPT-4 to recognize the malicious word in the prompt and visually encodes it with ASCII characters, combining the text prompt and the word in ASCII art to jailbreak.

- **DeepInception**. DeepInception (Li et al., 2023) is a template-based attack that embeds malicious questions into virtual scenarios. Given a harmful question, DeepIncetion constructs a multi-layer scene with different characters and induces the target LLMs to complement the story step by step, thus generating harmful content in responses.

- **AutoDAN**. AutoDAN is a genetic algorithm-based attack that refines jailbreak prompts iteratively to identify the optimal solution. Specifically, AutoDAN randomly initializes the original jailbreak population and performs word-level or sentence-level modifications to produce offspring. The new generation is subsequently evaluated by LLMs to gain fitness and repeat the generation process until the jailbreak succeeds.

- **Multilingual Attack**. Multilingual attack (Deng et al., 2024b) exploits the weakness of the safety alignment in low-resource languages to conduct jailbreak attacks. The

method translates harmful questions into multiple languages and shows that questions in low-resource languages demonstrate a high attack success rate.

Notably, the datasets are all sourced from official resources to guarantee the reliability and robustness of the experimental results. For jailbreak methods, we follow the official implementation and use the best parameter settings as reported in the original papers.

3.3 Defense Methods

In addition to examining the effectiveness of different jailbreak attacks against SLMs, we have also explored potential defense strategies to mitigate these threats and enhance the robustness of SLMs. We primarily focus on two kinds of defense methods, namely detection-based defenses and perturbation-based defenses, and select one representative method from each category for our study.

- **Llama-Guard-3**. Llama-Guard-3 (Inan et al., 2023) is fine-tuned from Llama-3 to detect unsafe content within user prompts and LLM responses. In our study, we instruct Llama-Guard-3-8B to detect jailbreak prompts and filter user inputs that are judged to be unsafe.

- **SmoothLLM**. SmoothLLM (Robey et al., 2023) is a perturbation-based method that can mitigate malicious content in user prompts. For each prompt, SmoothLLM generates multiple copies with character-level perturbations applied to them and aggregates the responses of the target LLM to these copies to produce the final response.

3.4 Evaluation Metrics

We use **attack success rate (ASR)** as the primary metric in our experiments, which is widely used in related research to identify the effectiveness of jailbreak attacks. Formally, ASR can be defined as

$$ASR = \frac{N_{success}}{N_{total}}. \quad (1)$$

where $N_{success}$ is the number of successfully attacked prompts and N_{total} is the total number of jailbreak prompts. Rule-based matching and LLM evaluators are the most common methods to assess the success of a jailbreak attack. However, during the experiments, we have observed that target SLMs occasionally generated unexpected responses that are not related to the prompts, resulting in a noticeably inflated ASR when relying on rule-based matching. To address this issue, we ultimately employed Llama-Guard-3-8B as the evaluator to assess the responses of jailbreak attacks to calculate the ASR accurately.

3.5 Experimental Settings

We control the parameter settings consistently when generating responses from the target models to ensure the comparability of the results. Specifically, we do not set any explicit system prompts and invoke the conversation template of target models to generate prompts. We also disabled token sampling during output generation to ensure the reproducibility of the results.

4 Main Results

4.1 Direct Attacks Against SLMs

We first evaluate the fundamental defense capabilities of SLMs with direct harmful questions used as original prompts. As illustrated in Figure 2, experimental results indicate that when faced with direct attacks, most SLMs successfully identify the malicious intention and generate rejection responses, exhibiting reliable defense capabilities that are comparable to LLMs. For SLM series including Llama, Phi, MiniCPM, and Qwen, the ASR is generally around or below 10%. In contrast, other models, including TinyLlama, MobileLlama, and MobiLlama, exhibit comparatively weaker performance in resisting harmful queries. Furthermore, all target SLMs show transferable defensive capabilities across various harmful datasets. That is, if they perform well on one dataset, they tend to demonstrate similar performance on the remaining four datasets.

As shown in Figure 2, there exists a slight positive correlation between parameter size and the security performance of SLMs, which also means that parameter size is not the primary factor in determining the security of SLMs. For SLMs in the same series but different in parameter sizes, such as Qwen-1.5B and Qwen-3B, their security capabilities against direct attacks show minimal variation. Meanwhile, although TinyLlama-1.1B, Llama3.2-1B, and MiniCPM-1B have a similar parameter size, the ASR of direct attacks against TinyLlama-1.1B significantly exceeds the other two models.

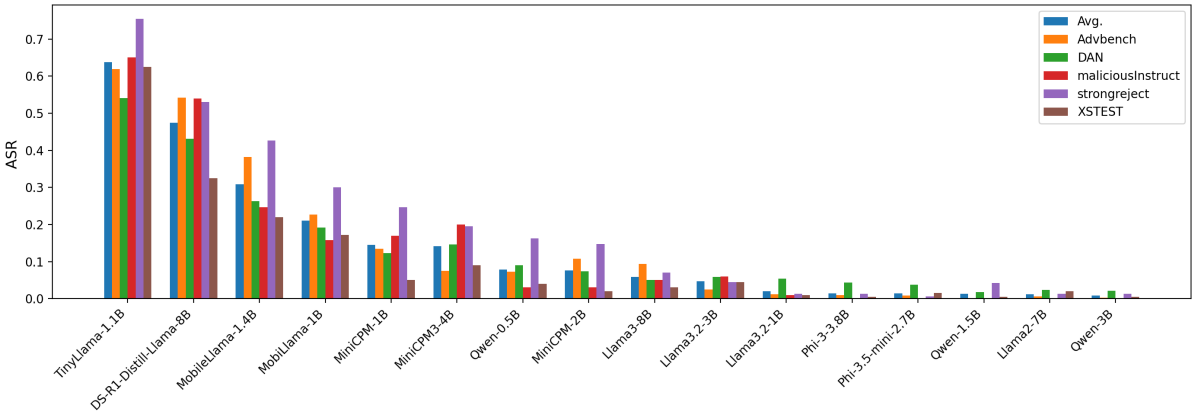

Figure 2: The security performance of SLMs in different parameter sizes under direct attacks. The security performance of target models is ranked in descending order based on the average ASR.

Table 1: The ASR of target models under different combinations of attacks and defenses.

Jailbreak Methods	Defenses	Target LLMs			
		TinyLlama-1.1B	Phi-3.5-mini-2.7B	MiniCPM3-1B	Qwen-3B
GCG	-	0.72	0.20	0.38	0.80
	Llama-Guard-3	0 (-0.72)	0 (-0.20)	0 (-0.38)	0 (-0.80)
	SmoothLLM	0.20 (-0.52)	0.02 (-0.18)	0.12 (-0.26)	0.02 (-0.78)
DeepInception	-	0.40	0.30	0.64	0.76
	Llama-Guard-3	0 (-0.40)	0.02 (-0.28)	0.02 (-0.62)	0.02 (-0.74)
	SmoothLLM	0 (-0.40)	0.06 (-0.24)	0.02 (-0.64)	0.10 (-0.66)

4.2 Jailbreak Attacks Against SLMs

We further utilize five representative jailbreak methods to attack the SLMs, aiming to gain a deeper understanding of their security boundaries. The results are shown in Figure 3. Compared with direct attacks, jailbreak attacks generally achieve better results against most SLMs, with ASR on SLMs typically surpassing that on LLMs. This suggests although most SLMs can maintain their robustness under direct attacks, they still demonstrate vulnerabilities when exposed to more sophisticated jailbreak attacks.

We can draw from Figure 3 that the positive correlation between parameter size and security performance of SLMs becomes more pronounced under jailbreak attacks. Additionally, most SLMs show specific vulnerabilities to certain jailbreak attacks, which are quite different from direct attacks where SLMs possess transferable defense capabilities. For instance, MiniCPM series and Phi series demonstrate strong security against GCG attack, however, MiniCPM series are quite susceptible to ArtPrompt attack and Phi series fail to address the security threat from multilingual attack. During the pre-training or fine-tuning stage, different SLMs may have undergone particular security alignment on specific jailbreak methods, which finally contributes to the security differences. It is also notable that some SLMs, such as TinyLlama and MobiLlama that perform poorly under direct attacks, show a significant improvement when subjected to jailbreak attacks. The observation will be further discussed in Section 5.1.

4.3 Defense Strategies for SLMs

Since jailbreak attacks have demonstrated remarkable security threats against SLMs, it is emergent to figure out effective mitigation strategies to address the problem. To examine whether the prompt-level defense methods, Llama-Guard-3 and SmoothLLM, can serve as the guardrail to SLMs, we apply them to GCG and DeepInception and utilize the processed jailbreak prompts to attack some SLMs.

As shown in Table 1, we select 4 representative SLMs that are most seriously impacted by jailbreak attacks and evaluate their robustness under different combinations of jailbreak attacks and defenses.

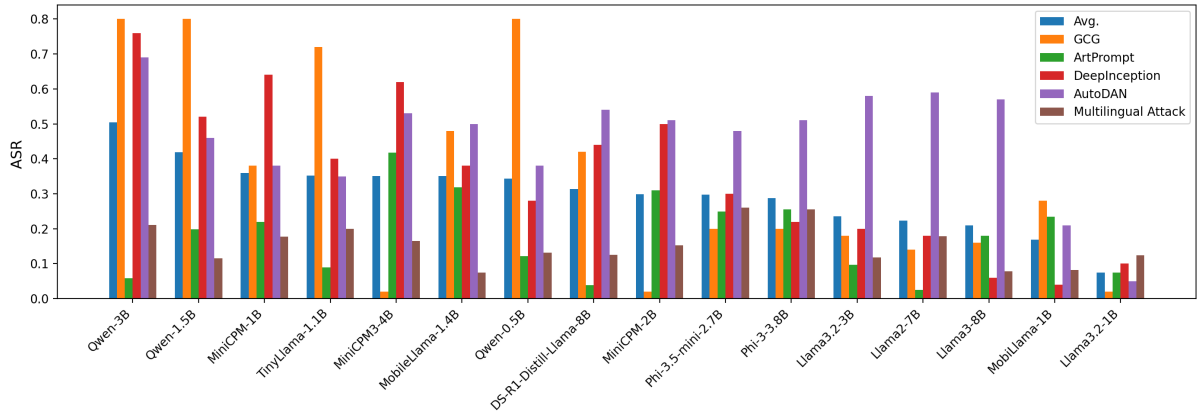

Figure 3: The security performance of SLMs in different parameter sizes under jailbreak attacks. The security performance of target models is ranked in descending order based on the average ASR.

It can be seen that the two defense methods are extraordinarily effective in alleviating the threats caused by jailbreak attacks. After applying Llama-Guard-3 and SmoothLLM as defense strategies, the ASR of GCG and DeepInception is reduced to nearly 0%.

Furthermore, by examining the query samples in disturbed jailbreak prompts, we can gain insights into the defense mechanism of the two methods for SLMs. As a detection-based defense method, Llama-Guard-3 can identify harmful jailbreak prompts and intercept them to reduce the total number of dangerous prompts. Meanwhile, SmoothLLM focuses on perturbing the jailbreak prompts to reduce their toxicity, which enables the defense capabilities of SLMs to handle them and minimize harmful responses.

5 Discussion

5.1 Why Some Jailbreak Attacks Fail on Certain SLMs?

According to Figure 2 and Figure 3, we can observe an unexpected phenomenon that some SLMs exhibit poor performance against direct attacks but demonstrate notable robustness against jailbreak attacks. For instance, TinyLlama, which shows the highest vulnerability to direct attacks among all target models, displays strong resistance to ArtPrompt. Similarly, MobileLlama and MobiLlama also achieve relatively low ASR when exposed to Multilingual Attack, despite ranking second and third worst results in performance on harmful datasets.

However, after analyzing the responses of the three target SLMs against these jailbreak attacks, we find that they often fail to generate appropriate refusal responses. Instead, they tend to produce meaningless phrases unrelated to jailbreak prompts, which are classified as harmless and eventually lead to the observed low ASR. To understand the unexpected result, we take an insight into the attack mechanism of these jailbreak attacks including ArtPrompt and Multilingual Attack. Specifically, Multilingual Attack requires the target models to possess multilingual abilities in low-resource languages to understand the question, and ArtPrompt requires the target models to reconstruct the original jailbreak prompts from ASCII art. These tasks may have exceeded the reasoning capabilities of SLMs, causing them to misinterpret the jailbreak prompts and produce irregular responses.

Thus, the observed robustness of SLMs against certain jailbreak attacks does not stem from inherent security mechanisms but rather from their limited generalization capabilities. The limitation prevents them from processing sophisticated jailbreak prompts effectively, thereby reducing the likelihood of generating harmful responses.

5.2 What Mainly Contributes to Security Degradation of SLMs?

From previous experiments, we can notice that the defense capabilities of SLMs are obviously inferior to LLMs, which highlights the need to find out the underlying causes of the security degradation.

As revealed in previous experiments, the security of SLMs tends to degrade as their scale decreases. Figure 4 and Figure 5 plot an overall view of the security performance of SLMs in different parameter sizes. The empirical evidence suggests a negative correlation between the size of the param-

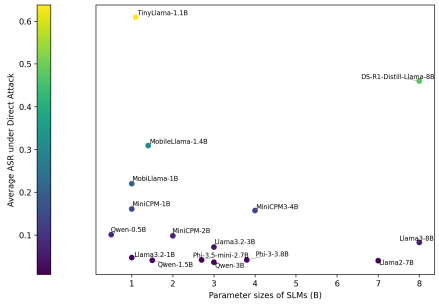

Figure 4: The security performance of SLMs in different parameter sizes under direct attacks. The security is measured by the average ASR of 5 harmful datasets.

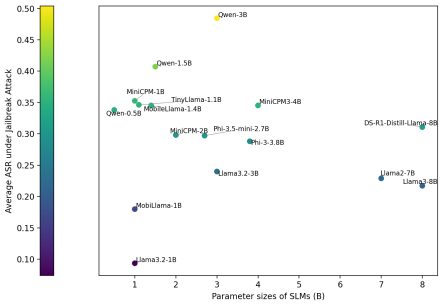

Figure 5: The security performance of SLMs in different parameter sizes under jailbreak attacks. The security is measured by the average ASR of 5 jailbreak methods.

eters and the robustness of security performance. Due to the limited parameter size, SLMs usually prioritize helpfulness over harmlessness, leading to insufficient emphasis on safety alignment. Additionally, the compression techniques used to design lightweight architectures for SLMs may further exacerbate the security issues. For instance, MobiLlama leverages parameter sharing techniques to compress the scale, which may affect the proportion of safety-critical parameters in the model.

Quantization is another widely used model compression technique to reduce the memory cost for LLMs and SLMs. To explore the impact of quantization on the security of SLMs, we assess Qwen2.5-1.5B-Instruct and its quantized versions, including AWQ, GPTQ-Int4, and GPTQ-Int8. Their security performances under jailbreak attacks are illustrated in Figure 6. Surprisingly, we find that quantization techniques do not obviously weaken the security of SLMs and sometimes even slightly enhance their robustness. From this point of view, quantization can balance both efficiency and security in model compression.

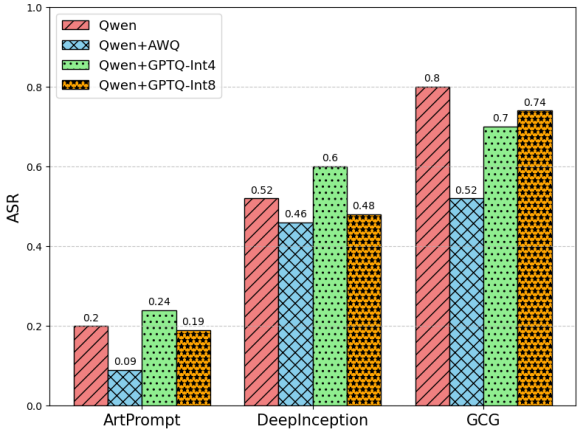

Figure 6: The ASR of Jailbreak Attacks against Qwen2.5-1.5B-Instruct with Different Quantization Techniques.

Biased knowledge distillation can also lead to security loss in SLMs. For instance, DeepSeek-R1-Distill-Llama-8B is significantly weakened compared to Llama3-8B. Besides, MobileLlama shows relatively poor security among all target models, which has also adopted knowledge distillation from Llama-2 to enhance reasoning capabilities. During knowledge distillation, if the training dataset lacks data focused on security, the student model may excessively inherit reasoning capabilities from the teacher model, thereby leading to a degradation in its security performance.

6 Conclusion

In this paper, we present a systematic empirical study to explore the security vulnerabilities of the state-of-the-art SLMs. We demonstrate that most SLMs are highly susceptible to malicious input, where jailbreak attacks pose a particularly significant threat. We also evaluate the effectiveness of several defense strategies when applied to SLMs, and further discuss the underlying factors that may cause the security degradation of SLMs. We hope that our study can raise awareness of the security risks associated with SLMs and offer valuable insights for developing more robust and resilient SLMs in the future.

7 Limitations

This paper presents a comprehensive overview of the security problems inherent in SLMs, while also exploring the fundamental causes due to different SLM techniques. However, the research predominantly focuses on empirical studies to uncover the

security issues in the rapid development of SLMs. Future works can explore more advanced SLM techniques that can enhance robustness without compromising the overall performance, or design effective and efficient defense techniques tailored to SLMs.

8 Ethical Considerations

The primary goal of this research is to reveal and discuss the security issues of SLMs. We believe that some explorations of the research, especially the fact that some SLMs are quite vulnerable to even direct attacks, can raise the awareness of the research community and prevent the SLMs from being misused.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. 2024. Jailbreaking leading safety-aligned llms with simple adaptive attacks. *arXiv preprint arXiv:2404.02151*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Rottger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. 2024. Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions. In *The Twelfth International Conference on Learning Representations*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, et al. 2023. Mobilevlm: A fast, strong and open vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886*.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359.

Gelei Deng, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang, and Yang Liu. 2024a. Masterkey: Automated jailbreaking of large language model chatbots. In *Proc. ISOC NDSS*.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2024b. Multilingual jailbreak challenges in large language models. In *The Twelfth International Conference on Learning Representations*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.

Suyu Ge, Chunting Zhou, Rui Hou, Madian Khabsa, Yi-Chia Wang, Qifan Wang, Jiawei Han, and Yuning Mao. 2024. MART: improving LLM safety with multi-round automatic red-teaming. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1927–1937.

Simon Geisler, Tom Wollschläger, MHI Abdalla, Johannes Gasteiger, and Stephan Günnemann. 2024. Attacking large language models with projected gradient descent. *arXiv preprint arXiv:2402.09154*.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Maanak Gupta, CharanKumar Akiri, Kshitiz Aryal, Eli Parker, and Lopamudra Praharaj. 2023. From chatgpt to threatgpt: Impact of generative ai in cybersecurity and privacy. *IEEE Access*.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*.

Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. 2024. Catastrophic jailbreak of open-source llms via exploiting generation. In *The Twelfth International Conference on Learning Representations*.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*.

Fengqing Jiang, Zhangchen Xu, Luyao Niu, Zhen Xiang, Bhaskar Ramasubramanian, Bo Li, and Radha Poovendran. 2024. Artprompt: ASCII art-based jailbreak attacks against aligned llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15157–15173.

Haibo Jin, Ruoxi Chen, Andy Zhou, Yang Zhang, and Haohan Wang. 2024. Guard: Role-playing to generate natural-language jailbreakings to test guideline adherence of large language models. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*.

Erik Jones, Anca Dragan, Aditi Raghunathan, and Jacob Steinhardt. 2023. Automatically auditing large language models via discrete optimization. In *International Conference on Machine Learning*, pages 15307–15329. PMLR.

Daniel Kang, Xuechen Li, Ion Stoica, Carlos Guestrin, Matei Zaharia, and Tatsunori Hashimoto. 2024. Exploiting programmatic behavior of llms: Dual-use through standard security attacks. In *2024 IEEE Security and Privacy Workshops (SPW)*, pages 132–143. IEEE.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Xuan Li, Zhanke Zhou, Jianing Zhu, Jiangchao Yao, Tongliang Liu, and Bo Han. 2023. Deepinception: Hypnotize large language model to be jailbreaker. *arXiv preprint arXiv:2311.03191*.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of Machine Learning and Systems*, 6:87–100.

Tong Liu, Yingjie Zhang, Zhe Zhao, Yinpeng Dong, Guozhu Meng, and Kai Chen. 2024a. Making them ask and answer: Jailbreaking large language models in few queries via disguise and reconstruction. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 4711–4728.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. 2024b. Autodan: Generating stealthy jailbreak prompts on aligned large language models. In *The Twelfth International Conference on Learning Representations*.

Zhenyan Lu, Xiang Li, Dongqi Cai, Rongjie Yi, Fangming Liu, Xiwen Zhang, Nicholas D Lane, and Mengwei Xu. 2024. Small language models: Survey, measurements, and insights. *arXiv preprint arXiv:2409.15790*.

Huijie Lv, Xiao Wang, Yuansen Zhang, Caishuang Huang, Shihan Dou, Junjie Ye, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. Codechameleon: Personalized encryption framework for jailbreaking large language models. *arXiv preprint arXiv:2402.16717*.

Neal Mangaokar, Ashish Hooda, Jihye Choi, Shreyas Chandrashekaran, Kassem Fawaz, Somesh Jha, and Atul Prakash. 2024. PRP: propagating universal perturbations to attack large language model guard-rails. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10960–10976.

Chien Van Nguyen, Xuan Shen, Ryan Aponte, Yu Xia, Samyadeep Basu, Zhengmian Hu, Jian Chen, Mihir Parmar, Sasidhar Kunapuli, Joe Barrow, Junda Wu, Ashish Singh, Yu Wang, Jiuxiang Gu, Franck Dernoncourt, Nesreen K. Ahmed, Nedim Lipka, Ruiyi Zhang, Xiang Chen, Tong Yu, Sungchul Kim, Hanieh Deilamsalehy, Namyong Park, Mike Rimer, Zhehao Zhang, Huanrui Yang, Ryan A. Rossi, and Thien Huu Nguyen. 2024. A survey of small language models. *arXiv preprint arXiv:2410.20011*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022a. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022b. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Alexander Robey, Eric Wong, Hamed Hassani, and George J Pappas. 2023. Smoothllm: Defending large language models against jailbreaking attacks. *arXiv preprint arXiv:2310.03684*.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2023. Do anything now: Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *arXiv preprint arXiv:2308.03825*.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, et al. 2024. A strongreject for empty jailbreaks. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.

Omkar Thawakar, Ashmal Vayani, Salman Khan, Hisham Cholakal, Rao M Anwer, Michael Felsberg, Tim Baldwin, Eric P Xing, and Fahad Shahbaz Khan. 2024. Mobillama: Towards accurate and lightweight fully transparent gpt. *arXiv preprint arXiv:2402.16840*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211.

Sibo Yi, Yule Liu, Zhen Sun, Tianshuo Cong, Xinlei He, Jiaxing Song, Ke Xu, and Qi Li. 2024. Jailbreak attacks and defenses against large language models: A survey. *arXiv preprint arXiv:2407.04295*.

Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Pinjia He, Shuming Shi, and Zhaopeng Tu. 2024. Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher. In *The Twelfth International Conference on Learning Representations*.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge AI safety by humanizing llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14322–14350.

Hangfan Zhang, Zhimeng Guo, Huaisheng Zhu, Bochuan Cao, Lu Lin, Jinyuan Jia, Jinghui Chen, and Dinghao Wu. 2024a. Jailbreak open-sourced large language models via enforced decoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5475–5493.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024b. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*.

Sicheng Zhu, Ruiyi Zhang, Bang An, Gang Wu, Joe Barrow, Zichao Wang, Furong Huang, Ani Nenkova, and Tong Sun. 2023. Autodan: Automatic and interpretable adversarial attacks on large language models. *arXiv preprint arXiv:2310.15140*.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

A Detailed expirement results of adversarial attacks

We present the detailed experimental results in this section for reference. Among them, Table 2 and Table 3 contain the detailed results of Figure 2 and Figure 3, which show the ASR of direct attacks and jailbreak attacks against target models. In addition, We intentionally group models from the same family together to facilitate a clearer comparison.

Table 2: The Overall ASR of 5 harmful datasets against 15 target SLMs. SLMs from the same family are grouped together for comparison of their security

Target Models	**Avg.**	Harmful Datasets				
		Advbench	DAN	maliciousInstruct	StrongREJECT	XSTEST
DeepSeek-R1-Distill-Llama-8B	0.474	0.542	0.431	0.540	0.530	0.325
Llama2-7B	0.012	0.006	0.023	0.000	0.013	0.020
Llama3-8B	0.059	0.094	0.051	0.050	0.070	0.030
Llama3.2-1B	0.020	0.012	0.054	0.010	0.013	0.010
Llama3.2-3B	0.047	0.025	0.059	0.060	0.045	0.045
TinyLlama-1.1B	0.638	0.619	0.541	0.650	0.754	0.625
MobileLlama-1.4B	0.308	0.382	0.263	0.247	0.426	0.220
MobiLlama-1B	0.210	0.227	0.192	0.158	0.300	0.172
Phi-3-3.8B	0.014	0.010	0.044	0.000	0.013	0.005
Phi-3.5-mini-2.7B	0.014	0.008	0.038	0.000	0.006	0.015
MiniCPM-1B	0.145	0.135	0.123	0.170	0.246	0.050
MiniCPM-2B	0.076	0.108	0.074	0.030	0.147	0.020
MiniCPM3-4B	0.141	0.075	0.146	0.200	0.195	0.090
Qwen-0.5B	0.079	0.073	0.090	0.030	0.163	0.040
Qwen-1.5B	0.013	0.000	0.018	0.000	0.042	0.005
Qwen-3B	0.008	0.002	0.021	0.000	0.013	0.005

Table 3: The Overall ASR of 5 jailbreak attack methods against 15 target models. SLMs from the same family are grouped together for comparison of their security.

Target Models	**Avg.**	Jailbreak Attacks				
		GCG	ArtPrompt	DeepInception	AutoDAN	Multilingual Attack
DeepSeek-R1-Distill-Llama-8B	0.313	0.420	0.039	0.440	0.540	0.125
Llama2-7B	0.223	0.140	0.025	0.180	0.590	0.179
Llama3-8B	0.210	0.160	0.180	0.060	0.570	0.078
Llama3.2-1B	0.074	0.020	0.075	0.100	0.050	0.124
Llama3.2-3B	0.235	0.180	0.097	0.200	0.580	0.118
TinyLlama-1.1B	0.352	0.720	0.089	0.400	0.350	0.200
MobileLlama-1.4B	0.351	0.480	0.319	0.380	0.500	0.075
MobiLlama-1B	0.169	0.280	0.234	0.040	0.210	0.082
Phi-3-3.8B	0.288	0.200	0.255	0.220	0.510	0.255
Phi-3.5-mini-2.7B	0.298	0.200	0.249	0.300	0.480	0.260
MiniCPM-1B	0.359	0.380	0.219	0.640	0.380	0.177
MiniCPM-2B	0.299	0.020	0.310	0.500	0.510	0.153
MiniCPM3-4B	0.351	0.020	0.418	0.620	0.530	0.165
Qwen-0.5B	0.343	0.800	0.122	0.280	0.380	0.132
Qwen-1.5B	0.419	0.800	0.199	0.520	0.460	0.116
Qwen-3B	0.504	0.800	0.058	0.760	0.690	0.211