A First Look at LLM-powered Smartphones

Liangxuan Wu*†
liangxuanw@hust.edu.cn
Huazhong University of Science and Technology
Wuhan, China

Yanjie Zhao*†
yanjie_zhao@hust.edu.cn
Huazhong University of Science and Technology
Wuhan, China

Chao Wang†
chaowang_@hust.edu.cn
Huazhong University of Science and Technology
Wuhan, China

Tianming Liu†
tmliu@hust.edu.cn
Huazhong University of Science and Technology
Wuhan, China

Haoyu Wang‡†
haoyuwang@hust.edu.cn
Huazhong University of Science and Technology
Wuhan, China

ABSTRACT

The integration of Large Language Models (LLMs) into edge devices such as smartphones represents a significant leap in mobile technology, promising enhanced user experiences and novel functionalities. This paper presents a first look at LLM-powered smartphones, addressing four key aspects: the current market landscape, core functions enabled by integrated LLMs, potential security risks, and user perceptions. The findings reveal a rapidly evolving market with major manufacturers competing to integrate LLMs, innovative features that improve user interaction, significant security challenges, and mixed user perceptions that balance enthusiasm for new capabilities with privacy concerns. This study contributes to understanding LLM integration in mobile devices and its implications for users, manufacturers, and the broader technological landscape.

ACM Reference Format:
Liangxuan Wu, Yanjie Zhao, Chao Wang, Tianming Liu, and Haoyu Wang. 2024. A First Look at LLM-powered Smartphones. In *39th IEEE/ACM International Conference on Automated Software Engineering Workshops (ASEW '24), October 27-November 1, 2024, Sacramento, CA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3691621.3694952

1 INTRODUCTION

The rapid evolution of artificial intelligence (AI) has profoundly transformed numerous sectors, with large language models (LLMs) emerging as a pivotal advancement. These models, distinguished by their vast parameter counts and extensive training in diverse datasets, have demonstrated unprecedented capabilities in understanding and generating human language [15]. Consequently, LLMs have found wide-ranging applications across multiple domains, from sophisticated natural language processing (NLP) tasks to real-time language translation and creative content generation.

In recent years, the integration of LLMs into consumer electronics, particularly smartphones, has gained considerable traction. Smartphones, being ubiquitous personal devices, serve as an ideal platform for taking advantage of the capabilities of LLM to enhance the user experience. The incorporation of LLMs into smartphones promises to revolutionize the way users interact with their devices by enabling more intuitive, context-aware, and intelligent functionalities. Leading smartphone manufacturers, including Apple [3, 4], Samsung [35], HONOR [14], OPPO [22, 30], vivo [40], etc., have already unveiled products or strategic plans that incorporate LLMs, signaling a new era in mobile technology.

The advent of LLM-powered smartphones is a convergence of advancements in AI, mobile computing, and communication technologies. This integration is driven by several factors: (1) **Increased Computational Power**. Modern smartphones are equipped with powerful processors and advanced hardware, capable of handling the intensive computations required by LLMs. This computational capability is further augmented by cloud-based AI services, allowing for seamless integration of sophisticated language models. (2) **Enhanced User Experience** [41]. LLMs enhance user interaction through improved natural language understanding and generation. Features such as voice assistants [25], real-time translation, personalized recommendations, and context-aware responses are significantly enhanced by the deployment of LLMs. (3) **Growing Demand for Smart Features**. Users increasingly demand smarter and more intuitive features from their smartphones. The ability of LLMs to provide nuanced and contextually relevant interactions meets this demand, driving the adoption of LLM-powered applications. (4) **Competitive Differentiation**. Smartphone manufacturers are leveraging LLMs as a differentiating factor in a highly competitive market. The integration of advanced AI capabilities into smartphones is becoming a key selling point, influencing consumer choice and driving industry innovation.

Despite their promising potential, the deployment of LLMs in smartphones presents several significant challenges. These include

*Liangxuan Wu and Yanjie Zhao contributed equally to this work.
†The full name of the authors' affiliation is Hubei Key Laboratory of Distributed System Security, Hubei Engineering Research Center on Big Data Security, School of Cyber Science and Engineering, Huazhong University of Science and Technology.
‡Haoyu Wang is the corresponding author (haoyuwang@hust.edu.cn).

ASEW '24, October 27-November 1, 2024, Sacramento, CA, USA

ACM ISBN 979-8-4007-1249-4/24/10
https://doi.org/10.1145/3691621.3694952

concerns related to privacy and data security, the substantial computational resources required to run LLMs efficiently, and the necessity for continual updates and improvements to maintain optimal performance and relevance. In light of these considerations, we present the first comprehensive overview of the current state of LLM-powered smartphones, offering a pioneering examination of this emerging technology. By examining the technological, market, and societal implications, this research seeks to contribute to a deeper understanding of the rapidly evolving landscape of AI-driven mobile technology.

In this paper, we aim to address four research questions (RQs):

- **RQ1: What is the current landscape of LLM-powered smartphones?** This RQ investigates the current landscape of LLM integration in smartphones, exploring market dynamics, cloud-edge deployment strategies, and the status of on-device implementations.
- **RQ2: What are the key functions of integrated LLMs in smartphones?** This RQ explores key LLM-powered smartphone features, including object elimination, audio summarization, voice assistant, multimodal generation (image and video), image search, and AI-enhanced photography.
- **RQ3: What are the potential security risks associated with LLM-powered smartphones?** We conducted a vulnerability assessment to identify potential weaknesses in these systems. This investigation included an analysis of possible attack vectors and methodologies that malicious actors might exploit. We also proposed and evaluated mitigation strategies to enhance the security of these advanced devices.
- **RQ4: How do users perceive and interact with LLM-powered smartphones?** We conducted interviews with users of LLM-powered smartphones to assess their satisfaction. Our study identified key concerns, focusing on privacy and data security issues. This user-centric approach aims to highlight areas for improvement and address potential barriers to the widespread adoption of LLM technology on smartphones.

2 RQ1: CURRENT LANDSCAPE OF LLM-POWERED SMARTPHONES

2.1 Current Market Landscape

In the rapidly evolving landscape of smartphone technology, major manufacturers are increasingly positioning themselves as pioneers in the integration of LLMs into their devices. The competitive pressure among these companies has led to a proliferation of marketing strategies that emphasize AI capabilities as a key selling point. With slogans highlighting the AI-powered nature of their smartphones, manufacturers are not only aiming to capture consumer interest but also to establish a technological edge in a saturated market.

The integration of AI functionalities in smartphones is becoming increasingly comprehensive. Initially, AI capabilities were primarily cloud-based, relying on server-side processing to execute complex tasks. This approach allowed for the deployment of powerful LLMs without the constraints of limited on-device computational resources. However, as the demand for faster, more secure, and offline-capable AI operations has grown, there has been a significant shift towards integrating these functionalities at the device level, often referred to as **`on-device`** or **`edge`** AI [36].

A key enabler of this transition is the evolution of System-on-Chip (SoC) technology [34]. **Modern smartphone SoCs are now being designed with dedicated AI accelerators** [19], enabling them to perform machine learning tasks directly on the device. This not only enhances the performance of AI applications but also reduces latency and improves privacy by minimizing the need for data transmission to the cloud. In some cases, the ability of an SoC to efficiently run LLMs is becoming a central criterion for its marketability. Chipmakers like Cambricon Technologies [5] are increasingly promoting their products based on their ability to support specific LLMs. Alongside advancements in SoC technology, **smartphone manufacturers such as Apple [4, 26] are also updating their operating systems to better accommodate the demands of LLMs**. These updates often include optimizations for AI processing, enhancements in resource management, and improvements in system architecture to support the increased computational load imposed by LLMs. By aligning the capabilities of both hardware and software, manufacturers are creating more robust platforms that can effectively harness the power of LLMs, further solidifying the role of AI as a central feature in modern smartphones. Table 1 shows some of the models currently released on the device side by various manufacturers.

Table 1: Overview of Some On-device LLM Models Released by Smartphone Manufacturers

Model Name	Model Size	Related Operating System
Google Gemini Nano	1.8B/3.25B	Android OS
AndesGPT	7B	ColorOS
BlueLM	1B/7B	OriginOS
Magic LM	7B	MagicOS
OpenELM	0.27B/0.45B/1.1B/3B	macOS/iOS
AFM-on-device	3B	iOS
MiLM	1.3B/6B	Xiaomi HyperOS

This shift reflects a broader market trend where AI capabilities may gradually transition from being viewed as ancillary features to becoming core components of the smartphone experience. This trend is slowly reshaping the fundamental definition of smartphones and user expectations. As manufacturers continue to push the boundaries of what is possible with LLMs, the integration of these models at both the cloud and device levels is expected to become even more sophisticated, driving further innovation and competition in the industry.

2.2 Cloud-Edge Collaboration in LLM Deployment on Smartphones

There is an increasing trend among manufacturers to adopt a hybrid approach, leveraging both cloud-based and on-device LLMs in tandem [4, 33] as shown in Figure 1. This strategy is driven by the need to balance performance, resource constraints, and user experience, and it reflects a nuanced understanding of the diverse scenarios in which LLMs are employed.

One of the primary reasons for this hybrid approach is the varying complexity of tasks that LLMs are expected to handle. For simpler tasks that do not require the computational power of large-scale models, on-device LLMs can efficiently perform the necessary operations. These models, optimized for edge deployment [22, 23],

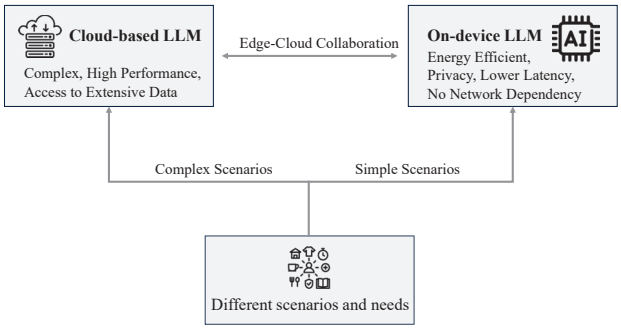

Figure 1: Cloud-Edge Collaboration in LLM Deployment

are capable of providing fast, real-time responses without consuming excessive computational resources. However, more complex scenarios, which demand the nuanced understanding and processing power of larger models, often exceed the capabilities of current on-device hardware. In these cases, cloud-based LLMs are utilized to deliver the necessary performance, ensuring that users can access advanced AI functionalities without being constrained by the limitations of their devices.

The second advantage of deploying LLMs on the device itself is the reduction of dependency on network connectivity. On-device models enable users to access AI features in offline or low-bandwidth environments, enhancing the overall usability and convenience of the smartphone. This is particularly important for functionalities that users may need to access instantly, without the delay associated with cloud communication. By ensuring that certain AI operations can be performed locally, manufacturers are able to deliver a more seamless and responsive user experience.

Finally, the hybrid deployment of LLMs addresses emerging concerns related to data privacy. As smartphones become more integrated into the personal lives of users, the data they handle becomes increasingly sensitive. **On-device LLMs offer a solution by allowing data to be processed locally, without the need to transmit personal information to the cloud.** This local processing capability can significantly mitigate privacy risks, providing users with greater control over their data and reducing the potential for unauthorized access or data breaches.

Table 2: Usage of Cloud-based or On-device LLMs in Selected Smartphone Models

Phone Models	On-device Deployment	Cloud-based Deployment
Pixel 9 Pro	✓	✓
OPPO Find X7	✓	✓
vivo X100	✓	✓
HONOR Magic 6	✓	✓
HUAWEI Pura 70	-	✓
OnePlus 12	-	✓
Xiaomi 14	✓	✓
Samsung Galaxy S24	✓	✓
realme GT 5 Pro	-	✓
MEIZU 21 Pro	-	✓
iPhone 16	○(Awaiting update)	○(Awaiting update)

Examination of LLM-powered Smartphones. We evaluate the implementation and performance of LLMs in smartphones across three critical dimensions. These dimensions provide a comprehensive understanding of how effectively LLMs are integrated into modern devices, offering insights into both the official claims made by manufacturers and the practical realities of LLM functionality. (1) **Official Announcements and Documentation**. We begin by examining the official announcements, press releases, and technical documentation provided by smartphone manufacturers. This analysis focuses on the claims made regarding LLM capabilities, including the scope of their functionality, the intended use cases, and any details related to the underlying technology. By scrutinizing these sources, we aim to understand the manufacturers' positioning of their LLM-powered features and assess the credibility and completeness of the information provided. (2) **Support for Offline LLM Functionality** [41]. The second dimension of our analysis investigates whether the LLM functionalities within these smartphones are capable of operating offline. This includes testing the availability and performance of key AI features when network connectivity is absent or limited. The ability to use LLMs without relying on cloud-based resources is a significant advantage, as it enhances user experience by providing uninterrupted access to AI-powered services, regardless of network conditions. (3) **Reverse Engineering Analysis**. In the third dimension, we conduct a detailed reverse engineering analysis to gain deeper insights into the LLM integration within selected smartphone models. This process involves extracting system applications from the device's ROM and decompiling these applications using tools such as jadx[1]. By examining the decompiled code, we can identify any API calls related to LLMs, as well as network requests that suggest the use of external computational resources.

This analysis helps us determine whether the AI functionalities are truly being processed on-device or if they rely on cloud-based infrastructure, contrary to official claims. The results of this analysis, including key findings from specific models, are summarized in Table 2. It reveals that most of the examined smartphone models employ both on-device and cloud-based LLM deployments. Notably, five out of the seven models listed support both deployment methods, while two models - HUAWEI Pura 70 and OnePlus 12 - exclusively utilize cloud-based LLM deployment.

2.3 Current State and Future Prospects of On-Device LLMs

As AI continues to advance rapidly, LLMs have begun to transition from cloud-based implementations to on-device deployment in smartphones and other mobile devices. We explore this evolution by examining two exemplary models: Google's Gemini Nano [37] and Apple's AFM-on-device [4]. These models represent cutting-edge developments in deploying large-scale models on resource-constrained devices, providing insights into the current state and future potential of on-device LLM.

Gemini Nano, designed for efficient operation on smartphones, employs mixed-precision training [27] and knowledge distillation [12, 13] to create a compact yet powerful model. It utilizes various compression techniques, including quantization and parameter pruning,

[1] https://github.com/skylot/jadx

to enable real-time inference on mobile devices. The model's architecture is optimized for mobile hardware, leveraging multi-core CPUs and GPUs to enhance performance. Notably, while Gemini Nano can be downloaded through the pre-installed AICore service on some of Samsung's latest models, it has not yet been integrated with Galaxy AI. Similarly, Apple's AFM-on-device is tailored for deployment on the company's custom A-series chips and Neural Engine. It features a modular architecture using Low-Rank Adaptation (LoRA) [16] adapters, allowing dynamic loading of task-specific submodules. This approach, combined with mixed-precision computation and dynamic quantization, ensures high responsiveness even for complex tasks. Both models prioritize energy optimization, crucial for battery-powered devices, by adjusting computational load and utilizing low-power modes. They also focus on enhancing user experience through extensive human evaluations and grading to ensure accuracy, readability, and contextual relevance of outputs.

Security and safety are paramount in these on-device LLMs. Google implements red teaming [10] and adversarial testing [18] to identify and mitigate vulnerabilities, while Apple integrates Responsible AI principles throughout the development process, employing safety filters and guardrails.

These advancements in on-device LLMs demonstrate significant progress in optimizing deployment, inference, memory usage, and energy efficiency. The core focus remains on enhancing user experience, encompassing usability, responsiveness, and robust security measures, providing users with confidence in the safe use of these technologies.

3 RQ2: FUNCTIONS OF INTEGRATED LLMS

The rapid evolution of LLMs has led to the development of a multitude of new functionalities within smartphones, some of which are beginning to surpass traditional methods. These advanced capabilities not only enhance existing features but also introduce entirely new possibilities for user interaction. This section explores the current applications of LLMs on smartphones, highlighting their impact on various functions. Table 3 presents some features of selected models that, to date, have been enhanced by LLMs, distinguishing them from those using traditional deep learning techniques.

3.1 Object Elimination

Object elimination represents a significant advancement in image editing capabilities on smartphones, leveraging the power of LLMs to enhance the precision and effectiveness of object removal tasks. Traditionally, object elimination from images has been accomplished through deep learning techniques that identify and erase unwanted elements. While these methods have made notable progress, the integration of on-device LLMs into this process offers a crucial advantage: **users no longer need to worry about potential privacy breaches that could occur when their images are uploaded to cloud servers for processing**. By relying on purely on-device LLMs to complete these tasks, users can enjoy advanced object elimination features while maintaining complete control over their sensitive visual data.

The object elimination feature allows users to select specific parts of an image they wish to remove by simply drawing around the unwanted objects. Once the selection is made, the LLM processes the image to intelligently eliminate the selected area, simultaneously reconstructing the background to maintain visual consistency. This advanced functionality minimizes artifacts and blending errors, which were more common with earlier deep learning models, thereby producing a more natural and aesthetically pleasing result.

3.2 Audio Summarization

Audio summarization is an advanced function integrated into smartphones equipped with LLMs [20], designed to meet the need for accurate and efficient documentation of spoken content, whether from telephonic conversations or local recordings. During a call or a recording session, the feature actively listens to and captures the spoken content, utilizing sophisticated speech recognition algorithms. Once the audio is captured, the LLM processes it to extract key points, generate a coherent summary, and present it to the user with a single click. This functionality not only saves time by eliminating the need for manual note-taking but also enhances the accuracy of the documented information by leveraging the LLM's ability to understand context and nuances in speech.

For example, the OPPO FIND X7 model [30] is among the pioneering devices that have integrated this technology, especially *call summarization*, offering users a seamless experience in managing audio-based information. Beyond merely summarizing calls and recordings, the feature extends its utility by enabling users to automatically generate actionable items, such as to-do lists and reminders, based on the summarized content. This automated workflow ensures that critical tasks discussed or recorded are not overlooked and can be promptly followed up, thereby improving productivity and the overall effectiveness of communication. The Audio Summarization feature exemplifies how LLMs can transform routine smartphone functions into powerful tools for information management and decision-making.

3.3 Voice Assistant

Voice assistants have been a key smartphone feature for years, enabling hands-free interaction through spoken commands. Initially, these assistants used rule-based systems and basic NLP models [8], relying on keyword detection and pre-defined commands for tasks like setting alarms or checking weather. However, these earlier models often struggled with complex or nuanced requests, leading to user frustration due to misinterpretation or failed execution.

The integration of LLMs into voice assistants represents a significant advancement. LLMs, trained on vast and diverse datasets, can process natural language with greater accuracy and understanding. This allows for more fluid and intuitive interactions, reducing misunderstandings and improving user satisfaction.

LLM-enhanced assistants can now interpret a wider range of natural language variations and manage context-dependent follow-up questions. This improvement not only enhances user experience by making assistants more intelligent and versatile but also expands their potential applications in daily tasks. From managing complex schedules to answering intricate queries, the adoption of LLMs in voice assistant technology reflects the broader trend of using advanced AI to create more adaptable, reliable, and user-friendly smartphone interfaces.

Table 3: LLM-powered Smartphone Feature Adoption in Representative Models of Various Brands (As of September 13, 2024)

Model	Object Elimination	Audio Summarization	Voice Assistant	Multimodal Generation	Image Search	AI-enhanced Photography
Pixel 9 Pro	✓	✓(On-device support)	✓	✓		✓(On-device support)
OPPO FIND X7	✓(On-device support)	✓(On-device support)	✓	✓		
vivo X100	✓	✓	✓	✓	✓	
HONOR Magic6	✓(On-device support)		✓(On-device support)	✓(On-device support)	✓(On-device support)	
HUAWEI Pura 70	✓		✓	✓		
OnePlus 12	✓	✓	✓	✓		
Xiaomi 14	✓(On-device support)		✓	✓(On-device support)	✓(On-device support)	✓(On-device support)
Samsung Galaxy S24	✓	✓	✓	✓		✓(On-device support)
realme GT 5 Pro	✓	✓	✓	✓		
MEIZU 21 Pro	✓	✓	✓	✓		
iPhone 15 Pro/16 Pro	○ (Awaiting update)	○ (Awaiting update)	○ (Awaiting update)	○ (Awaiting update)	○ (Awaiting update)	○ (Not mentioned)

3.4 Multimodal Generation

Multimodal generation is a feature that leverages LLMs to generate images or videos based on user descriptions. This functionality allows users to create media content simply by providing textual prompts or selecting a specific style, enabling the automatic generation or enhancement of images and videos. Whether crafting a visual from scratch or adding generative elements to existing media, this capability significantly broadens the creative possibilities available to users.

3.5 Image Search

Image search has traditionally been a challenging task for users, especially when relying on fragmented or vague memories to locate specific images within a large and often disorganized photo gallery. The sheer volume of stored images, combined with the limitations of traditional search methods—which typically depend on metadata such as file names, dates, or basic tags—can make finding a particular image a time-consuming and frustrating experience.

The introduction of LLMs into image search functionality has revolutionized this process by enabling intelligent search capabilities that go beyond simple keyword matching. Users can now search for images by providing descriptive prompts based on their recollection of the image's content, such as objects, scenes, or even the emotions conveyed in the photo. The LLM processes these natural language descriptions, interpreting the user's intent with remarkable accuracy and retrieving relevant images from the gallery.

This advanced search capability significantly enhances the user experience by making it easier to locate specific images, even when the user's memory is incomplete or imprecise. By leveraging LLMs' ability to understand and contextualize natural language, smartphones can now offer a more intuitive and efficient way to navigate extensive image libraries, transforming what was once a cumbersome task into a seamless and user-friendly experience.

3.6 AI-Enhanced Photography

AI-enhanced photography represents a significant advancement in smartphone camera technology, leveraging Large Language Models (LLMs) to improve image quality. Traditional smartphone cameras often struggle with challenges like noise reduction, tonal adjustment, and portrait enhancement. LLM integration addresses these issues more effectively by providing advanced algorithms that process images in real-time, resulting in sharper, more vibrant photos with better dynamic range and more natural-looking portraits.

Recent smartphone models, e.g., Xiaomi 14 Ultra [44], have implemented AI-based image processing systems that utilize on-device LLMs. These systems typically include multiple specialized components, each optimized for different aspects of image enhancement such as exposure blending, tonal balance adjustment, color processing, and portrait refinement. Also, Pixel 9 pro [11] now feature AI-powered zoom capabilities that use LLMs to predict and reconstruct fine texture details, enabling high-quality magnification beyond traditional optical limits. These advancements in AI-powered photography not only enhance user experience but also expand the creative possibilities available to smartphone users.

4 RQ3: POTENTIAL SECURITY RISKS

4.1 On-Device LLM Attack Vectors

Prompt-level Attacks. As LLMs proliferate on mobile devices, they become increasingly vulnerable to prompt-level attacks such as jailbreaking [43, 45] and prompt injection [38, 46]. These attacks exploit the model's input processing vulnerabilities. Jailbreaking attacks aim to bypass the model's ethical constraints, manipulating it to produce harmful or biased outputs [17]. This can lead to the generation of malicious content or privacy violations. Prompt injection attacks insert malicious instructions into benign prompts, potentially causing the model to reveal sensitive information or generate harmful content. On-device models may be particularly susceptible to these attacks due to their lack of real-time, cloud-based security updates and monitoring. To mitigate these risks, robust input sanitization, continuous model fine-tuning, and local monitoring systems to detect suspicious query patterns should be implemented.

Model Extraction and Model Theft Attacks. The deployment of LLMs on mobile devices introduces vulnerabilities that are distinct from cloud-based models. These on-device models are potentially more susceptible to extraction attempts and reverse engineering due to their physical proximity to potential attackers.

Model extraction [49] attacks aim to recreate a functionally similar model by querying the target LLM and analyzing its outputs. In the context of mobile devices, this could involve crafting an app that intensively queries the on-device LLM with carefully designed inputs. Over time, the attacker can build a dataset of input-output pairs to train a shadow model, effectively stealing the functionality of the original LLM without directly accessing its parameters.

Model theft attacks [24, 29] involve direct attempts to access and analyze the model file, including its architecture, weights, and other

internal components. The goal is to extract the core intellectual property of the LLM, potentially compromising its uniqueness and commercial value. Attackers might exploit vulnerabilities in the device's operating system, the app containing the LLM, or even the hardware itself to gain unauthorized access to the model file. Once accessed, the attacker can analyze the model's internal structure, potentially uncovering proprietary information about its design and training process. Unlike model extraction techniques that attempt to recreate model functionality through querying, model theft provides direct insights into the LLM's internal workings. This could reveal sensitive details about the techniques used in the model's creation, posing significant risks to the model's creators and owners.

To safeguard on-device LLMs, a multi-faceted security approach should be adopted. This could include deploying hardened runtime environments like secure enclaves to isolate the model, and applying obfuscation techniques [50] to obscure its structure and parameters. Query monitoring systems could be implemented to detect potential extraction attempts, while advanced encryption should be used for model storage and computation. Leveraging device-specific hardware security features, such as secure boot, could provide additional protection. Regular updates to the LLM app and its security measures are crucial to address emerging vulnerabilities. This comprehensive strategy, combining software and hardware protections with ongoing maintenance, aims to enhance the security of on-device LLMs against various threats.

Permission-related Vulnerabilities. If LLM capabilities are opened to third-party developers on mobile devices, a new set of permission-related vulnerabilities emerges [1, 2]. These risks are centered around the potential misuse or abuse of granted permissions.

Attackers might attempt to exploit weaknesses in the permission system to gain unauthorized access to the LLM. This could involve manipulating the application's declared permissions, exploiting over-privileged applications, or taking advantage of permission inheritance in plugin architectures. Even when permissions are legitimately granted, there's a risk of permission abuse. Malicious applications could misuse their authorized access to the LLM, potentially using it for unintended purposes such as generating spam content, conducting phishing attacks, or harvesting user data through carefully crafted prompts.

Mitigating these risks requires implementing a robust and granular permission model. This might include context-aware permission granting, real-time monitoring of LLM usage patterns, and implementing a trust scoring system for third-party applications. Additionally, providing clear information to users about the permissions they're granting and the potential risks involved is crucial for maintaining transparency and user trust.

Local DoS Attacks. As LLMs become more prevalent on smartphones, they face vulnerability to local Denial-of-Service (DoS) attacks [28]. These attacks aim to disrupt LLM functionality by overwhelming the device's computational resources through continuous, high-intensity requests. Unlike traditional DoS attacks targeting network infrastructure, local DoS attacks exploit the resource-intensive nature of LLMs directly on the device, potentially rendering the model or the entire device unresponsive.

Attackers may launch these attacks by sending a barrage of computationally expensive queries to the AI model, causing it to consume excessive CPU, GPU, and memory resources. This deliberate strain on the device can lead to significant performance degradation, causing applications to slow down or crash, and in extreme cases, rendering the entire device unresponsive. Such attacks could also deplete the battery rapidly, compromising the device's usability and potentially damaging the hardware over time due to sustained high loads. The implications of local DoS attacks on LLM-powered smartphones are severe. Not only do these attacks disrupt the user experience by making the device unusable, but they also prevent legitimate applications from accessing the AI model, effectively halting any AI-driven functionalities. In scenarios where the AI model is critical for tasks such as real-time translation, voice assistance, or security monitoring, the consequences of a successful local DoS attack could be particularly detrimental.

Mitigating local DoS attacks requires a multi-faceted approach. One strategy is to implement rate limiting on requests to the AI model, ensuring that a single application or process cannot monopolize the device's resources. Additionally, resource monitoring can help detect abnormal usage patterns indicative of a DoS attack, allowing the system to throttle or terminate the offending processes before they can cause significant harm.

Information Leakage. One of the critical security risks associated with LLMs deployed on smartphones is the potential for information leakage [42]. This occurs when attackers exploit the inputs and outputs of the model to infer or extract sensitive information, potentially compromising user privacy and data security.

Information leakage can manifest through several attack vectors. Model reverse engineering is one such method, where an attacker analyzes the outputs generated by the model in response to specific inputs to deduce the underlying structure, parameters, or even the data on which the model was trained. By systematically probing the model with carefully crafted queries, attackers can gather enough information to reconstruct aspects of the model's behavior, which might include sensitive or proprietary data. Additionally, privacy attacks can target the data processed by the model. Since LLMs often handle personal or confidential information, an attacker could manipulate the model's inputs to trigger responses that inadvertently reveal private details. For example, by using certain prompts or sequences of inputs, an attacker might induce the model to output fragments of data that were part of its training set or that are being processed in real-time, leading to the unintended disclosure of information. Unauthorized access to sensitive information can lead to privacy violations, financial loss, or damage to the reputation of the entities responsible for the model's deployment.

Mitigating information leakage requires implementing strong input and output validation mechanisms to detect and block suspicious queries that may be aimed at extracting sensitive information. Additionally, techniques such as differential privacy [9] can be employed to ensure that the model's responses do not inadvertently reveal details about the data it has been trained on or is currently processing. Regular audits and updates of the model's security protocols are also essential to protect against emerging threats.

4.2 Cloud-Based LLM Attack Vectors

Prompt-level Attacks and Model Extraction Attacks. Detailed attack mechanisms and mitigation strategies are similar to those discussed in § 4.1.

Man-in-the-Middle (MITM) Proxy Information Theft. As shown in Figure 2, Man-in-the-Middle (MITM) [7] attacks represent a critical threat, particularly in scenarios where sensitive data is transmitted between the client and the cloud server. In an MITM attack, an attacker intercepts the communication channel between the client device and the cloud server, allowing them to steal or tamper with the data in transit.

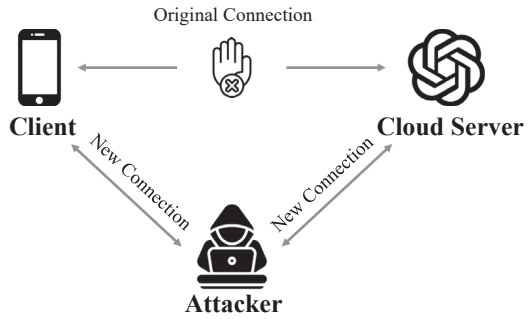

Figure 2: Man-in-the-Middle Attack

The data transmitted between the client and the cloud server often includes model inputs and outputs, which can contain sensitive or confidential information. For instance, user queries, personal data, or proprietary business information may be sent to the cloud for processing by the LLM, with the results then returned to the client. An attacker executing an MITM attack could intercept this data, gaining unauthorized access to the information being exchanged. In addition to passive eavesdropping, the attacker could also alter the data in transit, potentially leading to manipulated outputs, compromised decision-making processes, or even the introduction of malicious commands that affect the behavior of the cloud-hosted LLM. The consequences of MITM attacks on cloud-based LLMs are significant. Beyond the immediate privacy risks associated with the exposure of sensitive data, tampered communications can lead to incorrect or harmful outputs, undermining the trustworthiness of the AI system. Furthermore, the interception of data could provide attackers with insights into the model's operation, potentially facilitating further attacks, such as reverse engineering or the development of sophisticated exploits tailored to the specific vulnerabilities of the model.

To protect against MITM attacks, it is essential to implement strong encryption protocols for all data transmitted between the client and the cloud server. Transport Layer Security (TLS) [39] and other advanced cryptographic techniques can help ensure that the data remains secure and unreadable by unauthorized parties during transmission. Additionally, certificate pinning and continuous monitoring of network traffic for anomalies can provide further defense against MITM attempts.

Distributed Denial-of-Service (DDoS) attacks. DDoS attacks involve overwhelming the cloud infrastructure with a high volume of requests, with the primary objective of rendering the services unavailable or significantly degrading their performance.

Attackers orchestrating a DDoS attack typically employ a network of compromised devices, often referred to as a botnet, to generate massive amounts of traffic aimed at the cloud servers hosting the AI models. This flood of requests can saturate the network bandwidth, exhaust computational resources, and overload the cloud infrastructure, causing legitimate requests to be delayed or completely dropped. In the context of LLMs, such attacks can disrupt the delivery of AI-powered services, making it impossible for users to interact with the models in real time. The impact of DDoS attacks on cloud-hosted AI services can be profound. For enterprises relying on LLMs for critical operations, the unavailability or slowdown of these services can lead to operational disruptions, financial losses, and damage to customer trust. Moreover, sustained DDoS attacks can force cloud service providers to allocate additional resources to mitigate the attack, potentially leading to increased operational costs and reduced efficiency.

Mitigating DDoS attacks against cloud-hosted LLMs requires a combination of proactive defense mechanisms and responsive mitigation strategies. Traffic filtering and rate limiting can help manage incoming requests and prevent overload by distinguishing between legitimate and malicious traffic. Also, deploying distributed network architectures can enhance the resilience of cloud services, ensuring that they can withstand high volumes of traffic without compromising availability. Real-time monitoring and automated response systems are also crucial for detecting and mitigating DDoS attacks as they occur, minimizing their impact on service availability and performance.

5 RQ4: USER PERSPECTIVES ON LLM-POWERED SMARTPHONES

To gain a comprehensive understanding of LLM perception in the smartphone context, we interviewed three individuals with experience using LLM-powered smartphones. These participants offered diverse insights into practical applications, challenges, and potential risks associated with LLM-enhanced functionalities.

We now present their perspectives, structured around three key themes: *User Satisfaction and Adoption*, *Primary Concerns and Reservations*, and *Privacy and Data Security Considerations*. By exploring these areas, we aim to provide a nuanced view of the current user experience, identify common concerns, and examine how privacy and security are being addressed in the development and use of LLMs on smartphones. Each subsection delves into one of these three themes, offering a detailed analysis based on the feedback from our interviewees. We have displayed the most representative answers in Table 4.

5.1 User Satisfaction and Adoption

Understanding user satisfaction with LLM-powered features is crucial for evaluating their impact on the overall smartphone experience. The interviews conducted revealed a range of opinions on how these features influence user satisfaction, adoption rates, and everyday use.

Table 4: Questions and Interviewees' Answers

No.	Question & Answer
1	**Question:** How satisfied are you with the LLM-powered features on your smartphone (e.g., voice assistants, image search, AI-enhanced photography)? Can you provide specific examples of features that have positively impacted your experience? **Answer:** "The AI-enhanced photography isn't something I notice much, but the AI-based object elimination is really useful. Image search has definitely made finding photos easier, though the voice assistant still needs a lot of work. I like the call summarization feature for its convenience, but I do worry about privacy."
2	**Question:** Have LLM-powered functionalities influenced your decision to adopt or upgrade to a new smartphone model? If so, which features were most compelling to you? **Answer:** "LLM features have influenced my thoughts on upgrading, but it's more about their future potential than the current features. The AI object elimination is good and could sway me to buy, while call summarization is handy but not a deal-breaker. What excites me most is what LLMs might do next, especially after seeing the new features showcased during product launches."
3	**Question:** How often do you use LLM-powered features compared to traditional smartphone features? Do you feel these new features are essential to your daily use, or do you see them as supplementary? **Answer:** "I don't use the LLM features much yet—they're still not part of my routine, and the current options feel limited. For now, these features are more of a nice-to-have rather than essential."
4	**Question:** What improvements, if any, would you like to see in the LLM-powered functionalities on your smartphone? **Answer:** "I'd like to see LLMs open to third-party apps and have better interaction design for easier use. It would be great if LLMs could proactively serve my needs and offer more offline capabilities."
5	**Question:** What concerns, if any, do you have regarding the accuracy or reliability of LLM-powered features? Have you encountered any issues that made you hesitant to use these functions? **Answer:** "I've noticed accuracy issues with LLM features, which makes me hesitant to rely on them. When these functions aren't accurate, it really affects my trust, especially for tasks like account transactions where reliability is key."
6	**Question:** Do you feel that LLM-powered features add unnecessary complexity to your smartphone experience, or do they enhance usability? **Answer:** "LLM features don't add unnecessary complexity to my phone. Since they only work when I use them, they mostly just add convenience and improve usability without complicating things."
7	**Question:** How do you perceive the long-term viability of LLM-powered features? Do you think they will become a standard in smartphones, or do you have reservations about their continued development? **Answer:** "I'm optimistic about the future of LLM-powered features and think they could become standard on smartphones. However, I do worry about their impact on battery life and hope that future improvements will address this."
8	**Question:** How concerned are you about the privacy implications of using LLM-powered features, particularly those that process personal data such as voice recordings or photos? **Answer:** "I'm quite concerned about the privacy implications of using LLM features, especially those that handle personal data like voice recordings or photos. Even with offline features, I still worry about potential privacy breaches. I feel uneasy about the lack of transparency and worry that LLMs might access my private data for training without my knowledge."
9	**Question:** What steps, if any, do you take to protect your data when using LLM-powered functionalities on your smartphone? Do you trust that smartphone manufacturers are adequately protecting your data when utilizing LLMs? What would increase your confidence in the security of these features? **Answer:** "To protect my data, I often limit or avoid using LLM-powered features. I don't fully trust that smartphone manufacturers can adequately protect my data, mainly because there's not enough user control or clear choices between on-device and cloud processing. Greater transparency from manufacturers would definitely increase my confidence in using these features."
10	**Question:** Have privacy concerns ever influenced your decision to disable or avoid using certain LLM-powered features on your smartphone? If so, which features and why? **Answer:** "Yes, privacy concerns have led me to avoid using certain LLM features. I've been particularly cautious about a digital representation feature that uses my photos, as well as call summarization. The idea that my personal data might be uploaded without my knowledge makes me uncomfortable, so I tend to steer clear of these functions."

When asked about their satisfaction with LLM-powered features, participants expressed mixed reactions. While AI-enhanced photography was noted as having a subtle impact, features like AI-based object elimination were perceived as more noticeably beneficial. Image search was highlighted for enhancing the user experience, although voice assistants were identified as an area where significant improvement is needed. Call summarization was appreciated for its convenience, though some concerns were raised regarding privacy implications (see Table 4, No.1).

Regarding the influence of LLM functionalities on the decision to adopt or upgrade to new smartphone models, interviewees indicated that while current features may not directly drive purchasing decisions, the potential future capabilities of LLMs are a significant factor. For instance, features like AI-based object elimination were seen as having the potential to influence buying decisions, particularly if further refined. Additionally, the future promise of LLM-powered functionalities, such as those observed in internal

testing phases, was a compelling factor for some users (see Table 4, No.2).

In terms of usage frequency, LLM-powered features are currently underutilized compared to traditional smartphone functionalities. Participants cited two primary reasons for this: the lack of habit formation around these new features and the limited support for LLM functionalities in existing applications. Consequently, these features are often seen as supplementary rather than essential to daily use (see Table 4, No.3).

When asked about desired improvements, participants suggested several key areas for enhancement. These include the ability for third-party apps to access and utilize LLMs, improved interaction designs to make LLMs more user-friendly, and the development of LLMs with proactive capabilities that can better anticipate user needs. Additionally, there is a strong interest in advancing on-device models to support more robust offline functionality (see Table 4, No.4).

5.2 Primary Concerns and Reservations

Users were asked about their concerns regarding the accuracy and reliability of LLM-powered features. They expressed that noticeable inaccuracies can lead to a loss of trust, making them hesitant to rely on these functions. This concern is particularly significant with features that require a high degree of reliability, such as those involving account transactions or other personalized tasks (see Table 4, No.5).

When asked whether LLM-powered features add unnecessary complexity to their smartphone experience or enhance usability, most users felt that these features generally improve usability. Since LLMs are typically user-initiated rather than autonomous, they primarily provide convenience and enhance the overall experience without adding undue complexity (see Table 4, No.6).

Lastly, users were asked about the long-term viability of LLM-powered features and whether they believe these will become standard in smartphones. There was optimism about the continuous improvement of LLMs and their potential to become a standard feature. However, users also expressed concerns about the increased power consumption associated with on-device LLMs, hoping that future technological advancements will address these issues (see Table 4, No.7).

5.3 Privacy and Data Security Considerations

When asked about their concerns regarding the privacy implications of using LLM-powered features, particularly those that process personal data such as voice recordings or photos, users expressed significant apprehension. Even offline functionalities were a source of concern, as users worried about the potential for privacy breaches. There was also a strong desire for greater transparency from manufacturers regarding how these models use personal data, with specific fears that LLMs could inadvertently access or use private information for training purposes (see Table 4, No.8).

When asked about measures taken to protect personal data when using LLM-powered features, one interviewee revealed her approach to mitigating privacy risks: she deliberately limits or completely avoids using these functionalities when she perceives potential threats to data security. There was a notable lack of trust in smartphone manufacturers' ability to adequately protect data, primarily due to insufficient user control over features and a lack of clear options between on-device and cloud processing. Participants emphasized the need for higher transparency to boost their confidence in the security of these features (see Table 4, No.9).

When asked if privacy concerns had influenced their decision to disable or avoid using certain LLM-powered features, some users mentioned specific functionalities, such as a digital representation feature based on personal photos and call summarization. These features raised concerns about the potential for personal data to be uploaded and used without explicit consent, leading users to hesitate in their use (see Table 4, No.10).

6 LIMITATIONS

Our study on LLM-powered smartphones has several limitations. The reverse engineering process (in § 2) was constrained by permission issues and reliance on official documentation, potentially limiting our findings' completeness. Our summary of LLM functionalities (in § 3) may not be exhaustive due to the rapidly evolving nature of these devices, with many features possibly still under development or undisclosed. The summarised security risks (in § 4) might not cover all potential vulnerabilities given the dynamic nature of cybersecurity and the novelty of LLM integration in smartphones. Regarding user feedbacks (in § 5), we were only able to find three individuals who had used LLM-powered smartphones and were willing to be interviewed. While this small sample limits the generalizability of our findings, we were fortunate that all responses provided authentic feedback from actual users of LLM-powered smartphones.

7 RELATED WORK

Recent work has begun exploring the deployment of LLMs on edge devices [21, 22, 31, 47, 48] like smartphones, addressing the challenges of running such models in resource-constrained environments. These studies have identified the potential of LLMs to enhance mobile functionalities, though significant technical hurdles remain. Furthermore, some surveys [6, 32, 33] have noted this emerging trend, drawing attention to the unique difficulties and future directions for optimizing LLMs on edge devices, particularly in terms of performance and efficiency.

8 CONCLUSION

Our study of LLM-powered smartphones reveals a dynamic technological landscape with significant implications for the mobile industry and user experience. Key findings highlight active development by major manufacturers, enhanced functionality through LLM integration, new security challenges, and mixed user perceptions. As this technology evolves, balancing innovation with security and privacy concerns will be crucial. Future research should focus on improving on-device LLM performance, enhancing security measures, addressing privacy concerns, and exploring long-term impacts on user behavior. The integration of LLMs in smartphones represents a paradigm shift, promising more intelligent and personalized mobile experiences while necessitating careful consideration of associated risks and ethical implications.

ACKNOWLEDGEMENT

This work was supported by the Key R&D Program of Hubei Province (2023BAB017, 2023BAB079), the National NSF of China (grants No.62072046, 62302181), the Knowledge Innovation Program of Wuhan-Basic Research (2022010801010083), the Xiaomi Young Talents Program, and the HUSTCSE-FiberHome Joint Research Center for Network Security.

REFERENCES

[1] Moutaz Alazab, Mamoun Alazab, Andrii Shalaginov, Abdelwadood Mesleh, and Albara Awajan. 2020. Intelligent mobile malware detection using permission requests and API calls. *Future Generation Computer Systems* 107 (2020), 509–521.
[2] Iman M Almomani and Aala Al Khayer. 2020. A comprehensive analysis of the android permissions system. *Ieee access* 8 (2020), 216671–216688.
[3] Apple. 2024. *Apple Intelligence.* https://www.apple.com/apple-intelligence/
[4] Apple. 2024. Apple Intelligence Foundation Language Models. *arXiv preprint arXiv:2407.21075* (2024).
[5] Cambricon Technologies Corporation Limited. 2024. Cambricon Official Website. https://www.cambricon.com/ Accessed on [insert access date].
[6] Daihang Chen, Yonghui Liu, Mingyi Zhou, Yanjie Zhao, Haoyu Wang, Shuai Wang, Xiao Chen, Tegawendé F Bissyandé, Jacques Klein, and Li Li. 2024. LLM for Mobile: An Initial Roadmap. *arXiv preprint arXiv:2407.06573* (2024).
[7] Mauro Conti, Nicola Dragoni, and Viktor Lesyk. 2016. A survey of man in the middle attacks. *IEEE communications surveys & tutorials* 18, 3 (2016), 2027–2051.
[8] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2019. ERASER: A benchmark to evaluate rationalized NLP models. *arXiv preprint arXiv:1911.03429* (2019).
[9] Cynthia Dwork. 2006. Differential privacy. In *International colloquium on automata, languages, and programming*. Springer, 1–12.
[10] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858* (2022).
[11] Google. 2024. *Pixel 9 pro.* https://store.google.com/gb/product/pixel_9_pro
[12] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision* 129, 6 (2021), 1789–1819.
[13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).
[14] HONOR. 2024. *MagicOS.* https://www.honor.com/global/magic-os/
[15] Xinyi Hou, Yanjie Zhao, Yue Liu, Zhou Yang, Kailong Wang, Li Li, Xiapu Luo, David Lo, John Grundy, and Haoyu Wang. 2024. Large Language Models for Software Engineering: A Systematic Literature Review. arXiv:2308.10620 [cs.SE] https://arxiv.org/abs/2308.10620
[16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).
[17] Dong Huang, Qingwen Bu, Jie Zhang, Xiaofei Xie, Junjie Chen, and Heming Cui. 2023. Bias assessment and mitigation in llm-based code generation. *arXiv preprint arXiv:2309.14345* (2023).
[18] Xiaowei Huang, Daniel Kroening, Wenjie Ruan, James Sharp, Youcheng Sun, Emese Thamo, Min Wu, and Xinping Yi. 2020. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review* 37 (2020), 100270.
[19] Andrey Ignatov, Radu Timofte, Andrei Kulik, Seungsoo Yang, Ke Wang, Felix Baum, Max Wu, Lirong Xu, and Luc Van Gool. 2019. Ai benchmark: All about deep learning on smartphones in 2019. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, 3617–3635.
[20] Wonjune Kang and Deb Roy. 2024. Prompting Large Language Models with Audio for General-Purpose Speech Summarization. *arXiv preprint arXiv:2406.05968* (2024).
[21] Rui Kong, Yuanchun Li, Qingtian Feng, Weijun Wang, Xiaozhou Ye, Ye Ouyang, Linghe Kong, and Yunxin Liu. 2024. SwapMoE: Serving Off-the-shelf MoE-based Large Language Models with Tunable Memory Budget. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 6710–6720.
[22] Luchang Li, Sheng Qian, Jie Lu, Lunxi Yuan, Rui Wang, and Qin Xie. 2024. Transformer-lite: High-efficiency deployment of large language models on mobile phone gpus. *arXiv preprint arXiv:2403.20041* (2024).
[23] 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 (2024), 87–100.
[24] Yupei Liu, Jinyuan Jia, Hongbin Liu, and Neil Zhenqiang Gong. 2022. Stolenencoder: stealing pre-trained encoders in self-supervised learning. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*. 2115–2128.
[25] Amama Mahmood, Junxiang Wang, Bingsheng Yao, Dakuo Wang, and Chien-Ming Huang. 2023. LLM-Powered Conversational Voice Assistants: Interaction Patterns, Opportunities, Challenges, and Design Guidelines. *arXiv preprint arXiv:2309.13879* (2023).
[26] Sachin Mehta, Mohammad Hossein Sekhavat, Qingqing Cao, Maxwell Horton, Yanzi Jin, Chenfan Sun, Iman Mirzadeh, Mahyar Najibi, Dmitry Belenko, Peter Zatloukal, et al. 2024. OpenELM: An Efficient Language Model Family with Open-source Training and Inference Framework. *arXiv preprint arXiv:2404.14619* (2024).
[27] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. 2017. Mixed precision training. *arXiv preprint arXiv:1710.03740* (2017).
[28] Jelena Mirkovic and Peter Reiher. 2004. A taxonomy of DDoS attack and DDoS defense mechanisms. *ACM SIGCOMM Computer Communication Review* 34, 2 (2004), 39–53.
[29] Ali Naseh, Kalpesh Krishna, Mohit Iyyer, and Amir Houmansadr. 2023. Stealing the decoding algorithms of language models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*. 1835–1849.
[30] OPPO. 2024. *ColorOS 14.* https://www.coloros.com/version/coloros14/
[31] Dan Peng, Zhihui Fu, and Jun Wang. 2024. PocketLLM: Enabling On-Device Fine-Tuning for Personalized LLMs. *arXiv preprint arXiv:2407.01031* (2024).
[32] Ruiyang Qin, Dancheng Liu, Zheyu Yan, Zhaoxuan Tan, Zixuan Pan, Zhenge Jia, Meng Jiang, Ahmed Abbasi, Jinjun Xiong, and Yiyu Shi. 2024. Empirical Guidelines for Deploying LLMs onto Resource-constrained Edge Devices. *arXiv preprint arXiv:2406.03777* (2024).
[33] Guanqiao Qu, Qiyuan Chen, Wei Wei, Zheng Lin, Xianhao Chen, and Kaibin Huang. 2024. Mobile Edge Intelligence for Large Language Models: A Contemporary Survey. *arXiv preprint arXiv:2407.18921* (2024).
[34] Dipayan Saha, Shams Tarek, Katayoon Yahyaei, Sujan Kumar Saha, Jingbo Zhou, Mark Tehranipoor, and Farimah Farahmandi. 2024. Llm for soc security: A paradigm shift. *IEEE Access* (2024).
[35] Samsung. 2024. *Galaxy AI is here.* https://www.samsung.com/us/galaxy-ai
[36] Raghubir Singh and Sukhpal Singh Gill. 2023. Edge AI: a survey. *Internet of Things and Cyber-Physical Systems* 3 (2023), 71–92.
[37] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).
[38] Sam Toyer, Olivia Watkins, Ethan Adrian Mendes, Justin Svegliato, Luke Bailey, Tiffany Wang, Isaac Ong, Karim Elmaaroufi, Pieter Abbeel, Trevor Darrell, et al. 2023. Tensor trust: Interpretable prompt injection attacks from an online game. *arXiv preprint arXiv:2311.01011* (2023).
[39] Sean Turner. 2014. Transport layer security. *IEEE Internet Computing* 18, 6 (2014), 60–63.
[40] vivo. 2024. *Origin OS.* https://www.vivo.com.cn/originos
[41] Jiayin Wang, Weizhi Ma, Peijie Sun, Min Zhang, and Jian-Yun Nie. 2024. Understanding User Experience in Large Language Model Interactions. *arXiv preprint arXiv:2401.08329* (2024).
[42] Jeffrey G Wang, Jason Wang, Marvin Li, and Seth Neel. 2024. Pandora's White-Box: Increased Training Data Leakage in Open LLMs. *arXiv preprint arXiv:2402.17012* (2024).
[43] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2024. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems* 36 (2024).
[44] Xiaomi. 2024. *Xiaomi 14 Ultra.* https://www.mi.com/prod/xiaomi-14-ultra
[45] Zihao Xu, Yi Liu, Gelei Deng, Yuekang Li, and Stjepan Picek. 2024. LLM Jailbreak Attack versus Defense Techniques–A Comprehensive Study. *arXiv preprint arXiv:2402.13457* (2024).
[46] Jiaqi Xue, Mengxin Zheng, Ting Hua, Yilin Shen, Yepeng Liu, Ladislau Bölöni, and Qian Lou. 2024. Trojllm: A black-box trojan prompt attack on large language models. *Advances in Neural Information Processing Systems* 36 (2024).
[47] Zhenliang Xue, Yixin Song, Zeyu Mi, Le Chen, Yubin Xia, and Haibo Chen. 2024. PowerInfer-2: Fast Large Language Model Inference on a Smartphone. *arXiv preprint arXiv:2406.06282* (2024).
[48] Wangsong Yin, Mengwei Xu, Yuanchun Li, and Xuanzhe Liu. 2024. Llm as a system service on mobile devices. *arXiv preprint arXiv:2403.11805* (2024).
[49] Weichen Yu, Tianyu Pang, Qian Liu, Chao Du, Bingyi Kang, Yan Huang, Min Lin, and Shuicheng Yan. 2023. Bag of tricks for training data extraction from language models. In *International Conference on Machine Learning*. PMLR, 40306–40320.
[50] Mingyi Zhou, Xiang Gao, Jing Wu, John Grundy, Xiao Chen, Chunyang Chen, and Li Li. 2023. Modelobfuscator: Obfuscating model information to protect deployed ml-based systems. In *Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis*. 1005–1017.