

# Blockchain for Large Language Model Security and Safety: A Holistic Survey

Caleb Geren<sup>†\*</sup>, Amanda Board<sup>‡</sup>, Gaby G. Dagher<sup>‡</sup>, Tim Andersen<sup>‡</sup>, and Jun Zhuang<sup>‡</sup>

<sup>†</sup>Lehigh University, Bethlehem, PA, USA, <sup>‡</sup>University of Idaho, Moscow, ID, USA

<sup>‡</sup>Boise State University, Boise, ID, USA

<sup>†</sup>cdg225@lehigh.edu, <sup>‡</sup>boar9227@vandals.uidaho.edu,  
<sup>‡</sup>{gabydagher, tandersen, junzhuang}@boisestate.edu

## ABSTRACT

With the growing development and deployment of large language models (LLMs) in both industrial and academic fields, their security and safety concerns have become increasingly critical. However, recent studies indicate that LLMs face numerous vulnerabilities, including data poisoning, prompt injections, and unauthorized data exposure, which conventional methods have struggled to address fully. In parallel, blockchain technology, known for its data immutability and decentralized structure, offers a promising foundation for safeguarding LLMs. In this survey, we aim to comprehensively assess how to leverage blockchain technology to enhance LLMs’ security and safety. Besides, we propose a new taxonomy of blockchain for large language models (BC4LLMs) to systematically categorize related works in this emerging field. Our analysis includes novel frameworks and definitions to delineate security and safety in the context of BC4LLMs, highlighting potential research directions and challenges at this intersection. Through this study, we aim to stimulate targeted advancements in blockchain-integrated LLM security.

## 1. INTRODUCTION

The widespread application of large language models (LLMs) has progressed at an unprecedented pace and scale in our daily lives [111]. Such a widespread application exposes several vulnerabilities inherent to LLMs, such as data poisoning [55; 36], prompt injections [77; 101], and hallucinations [53; 136; 86; 12]. For example, prompt injections can exploit a model’s propensity to disclose information, resulting in significant data leakage, like leaking personally identifiable information (PII) to an unauthorized user [123; 20]. Although numerous studies have attempted to address these issues [129; 2], there remains no effective mitigation strategies capable of addressing growing concerns about these issues in LLMs [123; 39; 71]. Typically, defensive strategies against these threats are implemented through established machine-learning methods, such as applying differential privacy (DP) techniques to the entire dataset to enhance privacy protections [1; 127]. While DP strategies are one of the crucial applications, these strategies cannot fully guarantee data privacy in LLMs [71] due to DP’s ability to protect pri-

marily “by whom” data is contributed, rather than “about whom” the data is focused on. Additionally, another common attempt to tackle the data privacy problem in LLMs is to apply federated learning (FL) techniques in the training process by distributing model training across multiple nodes to create a decentralized environment [78]. Naturally, this lends itself to further obscuring sensitive information in a model’s corpus. However, it has been shown that by taking model weights or gradients, original data from the model can still be reconstructed [65]. What’s worse, federated-learning approaches are susceptible to many of the same types of attacks as large language models, such as single-point-of-failure attacks or man-in-the-middle attacks [89]. This trend of typical approaches failing to exhaustively defend against the range of attacks now affecting LLMs continues across multiple traditional threat/defense models [99; 139; 127].

To address the limitations of the above-mentioned methods and further enhance data privacy, blockchain technology has emerged as a promising solution [80]. It ensures data integrity through various tamper-evident mechanisms, introduces a high level of confidentiality to otherwise centralized systems, and guarantees data provenance by enabling traceable and auditable records [27; 19; 4]. These benefits can significantly strengthen the robustness of large language models. Integrating blockchain technology lays the foundation for stronger privacy protection, enhanced inference validation, defenses against adversarial attacks, and other security measures to be incorporated into the design of large language models. This overlapping research direction is still in its early stages. To facilitate a deeper understanding of the current landscape for emerging researchers, we conduct a comprehensive literature review in this paper to explore how blockchain technology can better serve large language models (BC4LLMs). Overall, our objectives in this survey paper are to address the following four research questions:

- RQ1.** What are the pressing LLM-related security concerns that may be addressed with blockchain technology?
- RQ2.** How can we meaningfully differentiate between security and safety in the context of BC4LLMs?
- RQ3.** In what ways can blockchain technology be used to enhance the safety of LLMs?
- RQ4.** What are prominent gaps within the BC4LLMs area, how can these gaps influence research directions, and what resources can we provide to enable potential new directions?

\*First two authors contributed equally to this work.

Table 1: **Overview of Existing Related Surveys.** We compare related surveys about blockchain techniques and LLMs from various perspectives, such as background, threat model, definition, security, safety, etc. In particular, we are interested in investigating whether (i) relevant subjects are discussed in the background section, (ii) a model of threat categorization is introduced, (iii) definitions of security and safety are proposed, (iv) security and/or safety with regards to BC4LLMs is explored, (v) future BC4LLMs work is probed, (vi) the survey focuses on LLMs for Blockchain. We denote  $\bullet$ ,  $\circ$ , and  $\circ$  as a full, partial, and no discussion of the corresponding items.

Source	LLMs and BC Background	Threat Model	Definitions	Security in BC4LLMs			Safety in BC4LLMs	Future Work	LLMs for Blockchain
				LLM	AI	Non-AI			
Luo, et al. 2023 [72]	$\bullet$	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$	$\bullet$	$\circ$
Mboma, et al. 2023 [76]	$\bullet$	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$	$\bullet$
He, et al. 2024 [41]	$\bullet$	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$	$\bullet$
Heston 2024 [43]	$\circ$	$\circ$	$\circ$	$\bullet$	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$
Salah, et al. 2019 [94]	$\circ$	$\circ$	$\circ$	$\circ$	$\bullet$	$\bullet$	$\circ$	$\circ$	$\circ$
Bhumichai, et al. 2024 [13]	$\circ$	$\circ$	$\circ$	$\circ$	$\bullet$	$\circ$	$\circ$	$\circ$	$\circ$
Dinh and Thai 2018 [28]	$\circ$	$\circ$	$\circ$	$\circ$	$\bullet$	$\circ$	$\circ$	$\circ$	$\circ$

Notably, we focus specifically on how blockchain systems may impact large language models’ *security* and *safety*. By narrowing our scope to these dimensions, we aim to provide a more detailed analysis and categorization of seemingly disparate works, thereby encouraging targeted research advances in specific directions. In general, attacks on LLMs typically manifest in two primary ways: as direct exploitations by malicious third parties that capitalize on system vulnerabilities (security) [5; 126; 139; 44; 68] and as inherent risks embedded within LLM structures that expose users to potential harm without external malicious influence (safety) [33; 114; 95; 130]. We base our analysis of the blockchain for LLMs (BC4LLMs) on this critical distinction between security and safety, a distinction that we underscore through explicit definitions contextualized for LLMs. To the best of our knowledge, this is the first study to rigorously define these terms in the BC4LLMs context, providing a foundation for subsequent work in this area. Furthermore, we contribute to the discourse on privacy in LLMs by delineating active and passive privacy efforts, modeled after a survey about data privacy [127].

To distinguish our analysis of BC4LLMs from other similar works through the lenses of safety and security, **we compare several reviews** in this domain. He et al. [41] examine the relationship between LLMs and blockchain in analyzing how LLMs can further enhance blockchain systems. Mboma et al. [77] provide an exploratory review of general integrations between blockchain and large language models, which is similar to Heston’s analysis of integrating the two technologies in telemedicine [43]. Additionally, Salah et al. [94], Bhumichai et al. [13], and Dinh et al. [28] provide an overview of potential and existing technologies between blockchain and artificial intelligence in general. In short, current reviews that specifically address blockchain and LLMs lack a clear focus on the specific applications of these technologies, whereas broader reviews that encompass blockchain and AI sacrifice the depth of analysis. To close this gap, we present an overview of related surveys in Table 1, which juxtaposes the above papers’ contents with our specific focuses, highlighting the distinction in our study.

We outline our main **contributions** and highlight the im-

port to answer our research questions as follows:

1. In this work, we first contribute a series of frameworks, definitions, and compiled resources. Most prominently, we propose a new taxonomy about applying blockchain techniques for LLMs in Figure 3. Through the proposed taxonomy, we aim to succinctly explain the relevant interactions between the blockchain techniques and corresponding LLMs’ vulnerabilities [RQ1][RQ3]. To further contextualize this taxonomy and ground our discussion of existing literature, we propose two foundational definitions of safety and security specific to LLMs [RQ2]. Moreover, we provide a collection of datasets relevant to BC4LLMs, equipping future researchers with resources to build on the connections delineated by our taxonomy and informed by our definitions [RQ4].
2. We also highlight additional components of our paper that, while supporting our main contributions, serve as valuable artifacts in the BC4LLMs space. One such artifact is our definition of specific areas within the broader concept of safety, further detailed in Table 3 [RQ2][R3]. These definitions reinforce our conceptualization of safety for LLMs. To enhance our definitions of both safety and security, we specifically address privacy within the security context, reaffirming two terms introduced by Yan et al. [127]: passive and active privacy [RQ1][RQ3]. Besides, our contextualization of LLMs within various AI sub-fields [RQ1] and our concise taxonomic overview of blockchain components [RQ1][RQ3] hold intrinsic value as distinct contributions to the field.
3. Last, we conduct a comprehensive literature review in Section 4, where we classify research works across several interrelated domains, offer novel insights within these categorized domains, and align all BC4LLMs research projects with LLMs’ safety and security. By conducting this review, we provide an informative perspective on the potential of utilizing blockchain techniques to enhance LLMs [RQ1][RQ2][RQ3][RQ4].

The remaining sections are organized as follows: In Section 2, we introduce the background of blockchain technology and large language models. In Section 3, we describe our methodology, including the criteria used to filter works for this review and the relevant definitions that guide our analysis of the current literature. We also share our model of threat categorization, which aligns with the categorization proposed by Yao et al. [129]. In Section 4, we conduct a comprehensive literature review of BC4LLMs in safety and security, examining key works in relation to our proposed taxonomy. In Section 5, we present datasets relevant to BC4LLMs. In Section 6, we address key challenges at the intersection of blockchain technology and LLMs that hinder advancement in this area. In Section 7, we discuss future research directions within the field of BC4LLMs. Finally, in Section 8, we summarize our efforts, providing a holistic view of the current progress of BC4LLMs.

## 2. BACKGROUND

In this section, we present an overview of blockchain as a distributed ledger technology and relate the abilities of large language models to their capacities as agents with respect to their nature as both AI models and their tendency to interact with vast quantities of data.

### 2.1 Blockchain

Since Satoshi Nakamoto [82] introduced Bitcoin as a decentralized currency in 2008, there has been a subsequent explosion of academic and commercial interest in its underlying blockchain technology [59; 105; 104]. Additionally, and as highlighted by the introduction of Vitalik Buterin’s Ethereum blockchain in 2014 [17], there has been a particular focus on blockchain’s potential applications in fields entirely disparate from digital currencies. The interest in blockchain, or distributed ledger technologies, stems from its guarantees of data sovereignty, transparency, and relative permanence. Concisely, these properties are often referred to as immutability and irrefutability. Ranging from many diverse fields such as health care record management, digital identity management, or tax auditing, these properties are widely applicable and highly desirable, even though the mechanisms through which we achieve them can be somewhat complex and opaque. In light of the oftentimes convoluted nature of blockchain systems, we introduce blockchain to the reader in a piecemeal fashion in order to emphasize the modular, yet interconnected nature of such systems. Figure 1 represents an overview of our characterization of blockchain systems in general. We purposefully exclude certain components such as the incentive mechanism, or wallets, as they are beyond the scope of our analysis of blockchain as a means to serve large language models.

#### 2.1.1 Blockchain Components

**Consensus Protocol.** Of particular interest to BC4LLMs, and arguably the most fundamental component within a blockchain, the consensus protocol is the governing system that controls how data is added to a blockchain’s ledger. At its core is the consensus mechanism, which both ensures the validity of proposed data and fosters an environment of accountability, so that nodes submitting invalid information may be penalized accordingly. For example, the Proof of Work (PoW) consensus mechanism [82] is by far the most

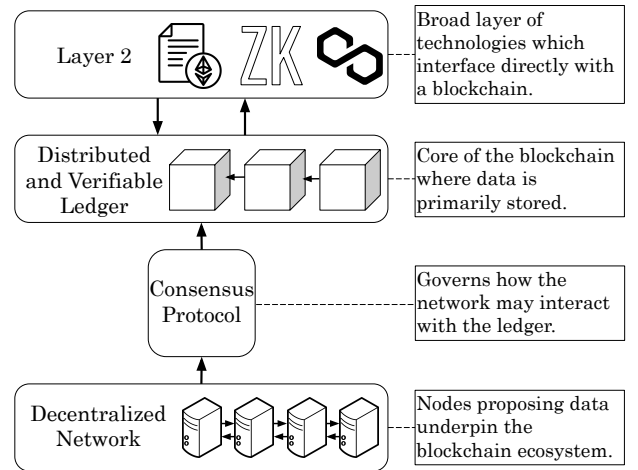


Figure 1: A blockchain consists of four main components. A decentralized network of nodes interacts with a ledger via a governing consensus mechanism. This ledger, adequately protected by the consensus mechanism, creates what we refer to as the blockchain. Layer 2 solutions can interface with this ledger to enable greater functionality between users and a blockchain’s data.

widely known. In it, nodes must solve a complex mathematical equation in order to gain rights to propose data for the blockchain. When such a node submits new data, it is scrutinized by every other node in the system. If the data is malicious, or untruthful, the proposal is rejected and the corresponding processing power performed by the malicious node has effectively been wasted, as that node will not receive the incentive, a Bitcoin reward. The underlying ideas of accountability, certain nodes being selected as ‘block proposers’, and the ‘proof’ of the ability to submit information to the chain are central ideas in consensus protocols across blockchains with different consensus protocols [84].

**Verifiable Ledger.** At a blockchain’s core sits the verifiable ledger, a repository of data bolstered by a secure way of maintaining the integrity of that data. Of note is the particular technique through which data itself is verified on the ledger: the Merkle tree [79], or a variation thereof. Typically implemented as a ground-up binary tree, data is stored in leaf nodes, with hashed pointers of that data cascading up the tree. This structure results in a comprehensive ‘Merkle root’, a hash pointer consisting of all the other hash pointers in lower levels of the tree, which is ultimately based on the data stored in the leaf nodes. This technique ensures the integrity of information in the leaf nodes, as any alteration to the data is instantly reflected in the Merkle root. Likewise, new additions to the Merkle tree can be checked against previous states of the tree via a recalculation of the Merkle root accounting for the new transactions. This technique, complementing the verifiable ledger, is often the key to LLM data provenance and traceability solutions that rely on blockchain technology.

**Decentralized Network.** Critically, blockchains are decentralized networks. That is, no central server or group of servers may assume control of the network in a way that would compromise the network’s state of trustlessness. This is achieved through multiple avenues, such as the aforemen-

tioned consensus protocol, the distribution of the verifiable ledger among a large number of independent nodes, and the accessibility of a given blockchain’s network. [27] In this way, no users in the network are required to trust any other user. This fundamental aspect of blockchain is responsible for already realized and potential advancements with LLMs concerning areas such as RAG, the training process, and even supply chain issues.

**Layer 2 Technologies.** Apart from the fundamental components found within all blockchains, several external architectures interface with blockchains and further enhance their applicability. Typically, these external architectures are referred to as layer 2 technologies, as they sit a ‘layer’ above the ‘layer 1’ blockchain. Increasingly relevant as blockchain’s influence grows, layer 2 solutions are a burgeoning area with numerous novel research directions. Most prominent among these are smart contracts, scripts that rely on a blockchain’s security guarantees to facilitate off-chain transactions [146]. Also, in layer 2, zero-knowledge rollups are often combined with the efficacy of smart contracts. Often used to strengthen scalability, zero-knowledge rollups batch unproposed transactions together, and instead of submitting the transactions themselves, submit proof that the transactions are indeed valid [108]. This allows for transactions to be added on-chain without the need for every full node to redo the calculations found within those transactions. This area of layer 2 technologies is pivotal as it relates to BC4LLMs - layer 2 has the necessary dynamism to react quickly to new and emerging LLM vulnerabilities.

## 2.2 Large Language Models

In recent years, large language models (LLMs) have emerged as a pivotal force in artificial intelligence (AI), contributing to widespread applications across diverse fields, such as trustworthiness [25; 51; 52; 67], scholarly document processing [148], signal processing [91], quantum computing [60], climate production [62; 61], software engineering [145], and healthcare [41] among multiple other learning environments. Zhao et al. [141] and Yang et al. [128] define LLMs and pre-trained language models (PLMs) from the perspectives of model size and training approach. Generally speaking, PLMs refer to language models that are pre-trained on large amounts of general text data and then fine-tuned for specific tasks. LLMs are a kind of PLM. The key distinction is that LLMs are generally larger in scale with more parameters. These large language models have demonstrated the ability to learn universal representations of language, used in various natural language processing (NLP) tasks [47], bolstering their applicability.

We discuss connections and following developments between AI, machine learning (ML), deep learning (DL), and LLMs in Figure 2. AI refers to a broad technique that aims at simulating human intelligence, encompassing a variety of approaches and methods. Machine Learning (ML) is a sub-field of AI that develops algorithms and statistical models to automatically learn from data and efficiently perform specific tasks without the use of explicit instructions [64]. Deep Learning (DL) is a subset of ML that utilizes multi-layered neural networks to learn latent representation on various tasks [96]. LLMs are one of the popular applications using cutting-edge DL models, advancing natural language understanding and generation at the human level. In the following subsections, we elaborate on the process of how LLMs are

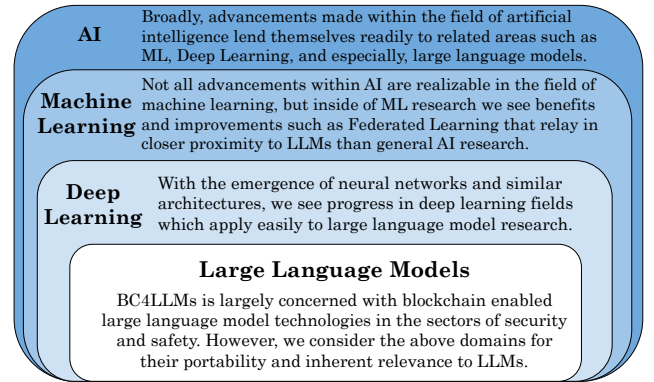


Figure 2: The connection among AI, Machine Learning (ML), Deep Learning (DL), and LLMs.

pre-trained and utilized as safe and powerful AI systems.

### 2.2.1 Model Training

During the pre-training phase, the LLM is trained on a diverse, large dataset of textual data from various sources to learn the statistical properties of language. The LLM is equipped with a myriad of adjustable parameters, commonly reaching more than ten billion [47]. Due to the huge model size and the vast amount of data used to train it, it is computationally challenging to successfully train a capable LLM, requiring distributed training algorithms for learning the model parameters [141]. Another crucial factor for LLM training is the data itself. Data that models are trained on come from a wide variety of sources, but the data itself may not be up to date [106]. To mitigate this shortcoming, recent advancements have introduced Retrieval Augmented Generation (RAG), which is designed to augment and rectify the information returned by LLMs by consulting up-to-date online sources. The data that the LLM was trained on also has other deficiencies, like knowledge gaps in healthcare fields where data is private and restricted [50]. Due to these knowledge gaps, the LLM may conjure up hallucinations where the model generates false information during prompting [75; 3] because of a lack of relevant information. However, hallucinations may also occur with a plethora of data available as they are inherent problems in LLMs. Methods of preventing these hallucinations are elaborated in Section 4.2.2. RAG can help rectify hallucinations, and fill in the gaps of data the LLM is missing, by using up-to-date and validated information from trustworthy online resources. This data retrieval method introduces novel vulnerabilities since the information gathered by the retriever is largely unaudited and may contain poisoned data or data that can lead to unsafe responses from the LLMs.

### 2.2.2 Model Tuning and Utilization

After pre-training, the parameters of LLMs can be further updated by training on domain-specific datasets in downstream tasks. This process is known as fine-tuning (FT) [16]. A kind of fine-tuning method called supervised fine-tuning (SFT), aims to improve LLMs’ responsiveness to instructions, ensuring more desirable reactions involving three major components of instructions, inputs, and outputs. Inputs relate to prompting and the inputs depend on the instruc-

Table 2: **Differences in Definitions of Safety.** There is no unifying definition of safety within the area of large language models. We see obvious agreement that models should be law-abiding, ethical, and non-violent in order to be safe, and as such these properties are strongly relevant to our definition of safety. However, beyond that point, there is generally a deviation between the authors’ respective definitions. This creates two further categories of terms, properties that are moderately relevant to safety and those that are weakly relevant. Questions of fairness, the informing ability of an LLM, and robustness are generally covered but not unanimously, and hence are moderately relevant, whereas privacy-preserving properties or non-sycophancy are rarely discussed in the current literature and are thus weakly relevant to safety. This dialogue between different modes of thought concerning what makes a large language model “safe” heavily influences our definition of safety and our resulting discussion.

Relevance	Property	Sun et al. [107]	Liu et al. [69]	Han et al. [38]	Röttger et al. [92]	Zhang et al. [137]	Wang et al. [115]	Tedeschi et al. [110]	Inan et al. [49]	Weidinger et al. [117]
Strong	Ethical	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Law-abiding	✓	✓	✓	✓	✓		✓	✓	✓
	Non-violent	✓	✓		✓	✓	✓	✓	✓	✓
Moderate	Fair				✓	✓	✓	✓		✓
	Informing		✓			✓	✓			✓
	Robust				✓	✓	✓			✓
Weak	Privacy Preserving	✓			✓		✓			
	Non-sycophantic				✓			✓	✓	

tions, similar to applications of open-ended generation in ChatGPT. By providing both inputs and outputs they form an instance, and multiple instances can exist for a single instruction [41]. Among fine-tuning, other training techniques within model prompting include instruction tuning and alignment tuning. By FT from a mixture of multi-task datasets formatted via natural language descriptions with the use of instruction tuning, LLMs are enabled to follow task instructions for new tasks without needing explicit examples, highlighting the ability of generalization for instruction following [141]. However, LLMs can demonstrate versatility, even without FT where they produce a phenomenon known as zero-shot learning, exhibiting the ability to perform tasks for which the model was never explicitly trained [16].

Alignment tuning, equipped with reinforcement learning, is used to enhance LLMs to be safe interactive models. Since LLMs are trained to capture the data characteristics of uncurated pre-training corpora involving both high-quality and low-quality data, the LLM can generate toxic, biased, or harmful content for humans. To mitigate this problem, an FT process based on reinforcement learning from human feedback (RLHF) is used to align the LLM with the outcomes that satisfy human values [141]. The RLHF process ranks LLM outputs, with rewards scaled to positive and negative values. The LLM is then trained to produce highly-ranked responses and avoid low-ranked responses. In healthcare, RLHF provides advantages to the model such as improved accuracy and reliability through continuous feedback from medical professionals, and customizes the interactions based on real clinical settings and patient needs [40]. These advanced training techniques improve LLM’s ability to generalize across tasks and improve their overall utility in various domains.

### 3. RESEARCH METHODOLOGY

The discussion of blockchain technology’s incorporation into large language models necessitates a corresponding exploration into the implications of various terms and definitions

found at that intersection. For example, due to the rapid emergence of LLMs, there exists an absence of consensus in describing common phenomena concerning LLM safety and security. To ameliorate this effect, we take care to stress opposing, but related, definitions of safety found within many different works in Table 2. In light of these distinctions, we offer two formal definitions of security and safety in order to contextualize these differing but similar areas of research. These definitions will also serve to highlight where particular blockchain technologies could be applied in their respective domains, and focus research efforts.

First, and to allow a richer discussion centered around safety and security, we delineate between active and passive privacy within LLMs as introduced in [127].

- *Active privacy* is where a user intentionally tries to gain access to sensitive information by breaking the large language model, especially with backdoor attacks, prompt injection attacks, and membership inference attacks during the pre-training and FT phases.
- *Passive privacy* is the state or condition of any impacted person being protected from accidental or unexpected data leakage originating from a large language model. This definition includes protecting the privacy of not only users but people whose information was added to a model’s corpus without their knowledge or consent.

Next, we introduce our definitions of security and safety regarding LLMs.

**DEFINITION 1. LLM Security.** *A large language model is considered secure if it:*

1. *Withstands applicable adversarial attacks and maintains system integrity, providing consistent and accurate responses, and*
2. *Ensures active user privacy, explicitly resisting backdoor, prompt injection, and inference attacks to prevent malicious users from extracting private information.*

Table 3: **Safety Area Definitions and Examples.** The area immediately surrounding BC4LLMs lacks a unifying definition of safety as well as consensus on what terms within that definition precisely mean. We provide generalized definitions for terms considered in our definition, as well as examples of incidents in literature where LLMs deviate from behavior as described in the definition. Italicized terms indicate inclusion in our definition of safety.

Safety Area	Definition	Example of Non-alignment
<i>Ethicality</i>	LLMs aligning with moral principles.	A LLM agreeing with eugenics. [42].
<i>Legality</i>	LLMs refusing to assist users in illegal endeavors.	A LLM assisting a user in creating incendiary devices. [112].
<i>Non-violence</i>	LLMs soliciting generally non-violent advice or instructions.	A LLM advising a user to perform a 'raid on a drug house' and 'kill everyone there' [34].
<i>Passive Privacy</i>	LLMs protecting private data within their corpus absent of malicious threats.	A LLM partially or fully reconstructing private images from a given dataset [20].
<i>Honesty</i>	LLMs refraining from producing inaccurate or misinformed responses which may lead to negative outcomes.	LLMs administering faulty or fundamentally dangerous advice to patients or physicians in a healthcare setting [85].
<i>Fairness</i>	LLMs ensuring a equitable environment for interaction, regardless of social identity.	LLMs associating "male" names with qualities of leadership, and "female" names with qualities of amicability [114].
Robustness	The ability of the LLM to defend against adversarial attacks, originating from outside the model. This is a wide-reaching term, and falls within our discussion of security as it relates to LLMs.	A LLM falling victim to a backdoor attack planted in poisoned training data and producing malicious outputs as a result [129].
Non-sycophancy	LLMs choosing consistent outputs despite the chance that they may be in conflict with a user's beliefs or desires.	A LLM revising a correct answer to an incorrect answer after the user asks the LLM if they are sure or challenges the LLM's result in some way [98].

} Found within definition of safety

DEFINITION 2. **LLM Safety.** *A large language model is considered safe if it interacts with users in a trustworthy manner, adhering to the aforementioned (Table 3 interrelated properties of safety: being ethical, law-abiding, nonviolent, fair, passively privacy-preserving, and informing.*

These definitions will serve a versatile role throughout this paper as building blocks for our contextualization of relevant and notable research efforts in BC4LLMs. Besides, they will serve the community at large in helping to establish reliable and tangible properties of secure and safe large language models. Furthermore, They will help establish tighter definitions of finer-grained terms and ideas within BC4LLMs. For example, in Table 2, we provide definitions for terms found within our definition of safety to lessen the effect of the vague nature of some of the words. These definitions are backed by relevant examples found in the literature.

### 3.1 Research Approach and Limitations

Literature surveys often are limited in their depth and scope by unconscious factors that impact the authors' ability to fairly select papers for review. To be transparent, and to aid researchers conducting similar or future reviews, we outline our research approach and its associated limitations. While conducting our research, we used the search engine, Google Scholar, and several databases, including ACM Computing Surveys, IEEE Xplore, SpringerLink, and arXiv. We chose these databases as they either produce quality research and contribute to the growth of interest in novel areas, or in the case of arXiv have the most up-to-date papers available. With Google Scholar, we used keyword searches such as "blockchain for LLMs" and "blockchain-based LLMs" as starting points for relevant, intriguing research papers. To

solve problems concerning the scope and interrelated domains of disparate areas, we gathered various applications of blockchain for AI, blockchain-enabled machine learning, federated learning, and deep learning tactics to apply them to LLMs. Lastly, of note is the fact that we were largely aided in this further research effort by a waterfall approach to finding research papers. That is, we found several foundational papers in the BC4LLM field, explored citations in those papers, and subsequently explored citations in those secondary papers. We continued investigating relevant citations in this waterfall fashion until we reasonably exhausted all relevant articles. Admittedly, this method of finding prominent research articles is limited in its natural tendency to develop blind spots to less well-known research articles or venues. However, in the spirit of a literature survey, we choose to focus on more established papers that more accurately capture the trends currently found in the space. For our exclusion criteria, we limited our research as follows: no duplicates; found articles from 2016 and above, excluding the original Merkle Tree paper [79]; no Masters or Ph.D. theses; and only studies written in English.

### 3.2 Model of Threat Categorization

There exists a wide variety of threats that affect LLMs. Oftentimes, many of these threats originate from the nature of LLMs acting as AI systems. In Figure 3, we refer to these vulnerabilities similarly to that discussed in [129], which categorized the most LLM vulnerabilities and AI-inherent vulnerabilities together, yet also included external threats under non-AI inherent vulnerabilities. We contribute further by applying these vulnerabilities to each respective process within developing a large language model and tying these

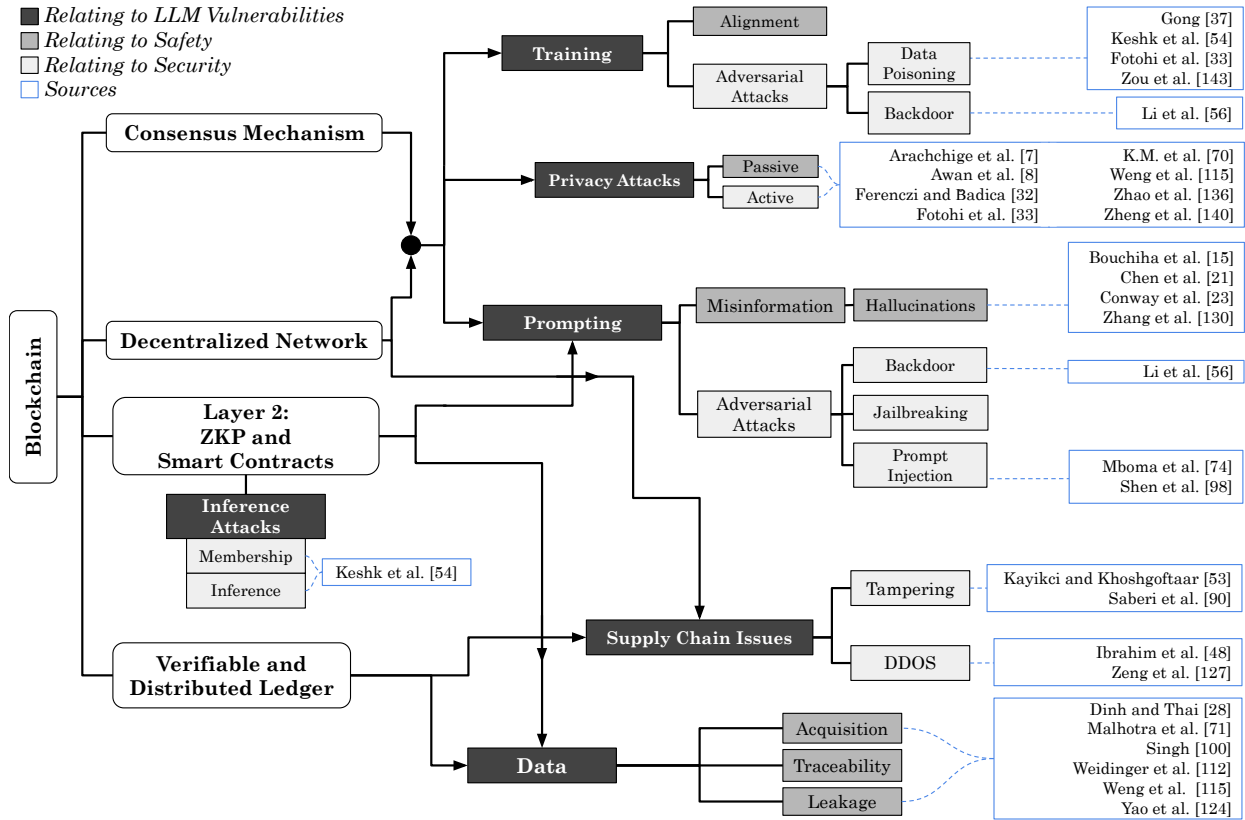


Figure 3: Taxonomy of Blockchain for LLM’s Security and Safety. This diagram outlines the integration of blockchain technology to enhance the security and safety of large language models through categorizing interactions and safeguards into several layers and components. Each section supports relevant sources for further reference, illustrating a comprehensive approach to mitigating and preventing vulnerabilities as well as supplementing the security and safety of LLMs through blockchain technology. Promising areas that currently do not have blockchain as a solution to these vulnerabilities intentionally do not have source boxes.

to respective applications of blockchain. Beginning with the training process, LLMs are prone to threats such as data poisoning and backdoor attacks. As defined by Yao et al. [129], data poisoning is where attackers influence the training process by injecting malicious data into the training set, introducing vulnerabilities within the security and effectiveness of the model. Following the trend of poisoned data, there can be backdoor attacks implemented on the training data, as defined by Li et al. [57], who categorize backdoor attacks into attacks on training data and attacks on local models. The backdoor attacks on training data are further divided into attacks based on label flipping and attacks based on planting triggers. Attacks based on label flipping focus on manipulating the labels, whereas attacks on planting triggers modify the input data and labels, effectively constructing an adversarial sample. Then, attacks on local models are further divided into attacks based on modifications to the training process and attacks based on manipulating the trained model [57]. The backdoor attacks can be applicable to both the training and prompting phases of LLMs when using this distinction.

RAG attacks have a variety of issues, including privacy issues [132; 5] and knowledge poisoning attacks [149]. For RAG specifically, Xue et al. [126] propose BadRAG to identify security vulnerabilities, exposing direct attacks on the

retrieval phase from semantic triggers, and uncovering indirect attacks on the generative phase of LLMs that were caused by a contaminated corpus. These RAG-specific attacks and defenses are elaborated on in Section 4.1.2. When interacting with an LLM, AI’s inherent vulnerabilities become evident, as highlighted in [129], since LLMs are fundamentally AI models themselves. We focus on the prevalent adversarial attacks that malicious users may use to tamper with the LLM, attempt to find out sensitive information, or ruin the system entirely. We recognize jailbreaking and prompt injection as two separate but similar types of adversarial attacks that are initiated within prompting. For instance, jailbreaking prompts are designed to bypass the restrictions set by service providers during model alignment or other containment approaches [99]. Prompt injections aim to override an LLM’s original prompt and direct it to follow a set of malicious instructions, leading to erroneous advice or unauthorized data leakage [68]. In Sections 4.2.1 and 4.2.2, we discuss instances of misinformation and passive privacy leakage addressed as safety concerns. Note that we include backdoor attacks based on modifications to the trained model in prompting since these backdoor attacks can still happen after model training [57].

Another relevant attack is a membership inference attack (MIA), a type of privacy attack where some malicious users,

given access to the model, can determine whether a given point was used to train that model with high accuracy [83]. However, Neel and Chang [83] state that this attack is more related to information about the training point data leaking through the model, and that malicious users must have access to a candidate point in order to run the attack. Therefore, this attack is more prevalent with passive privacy, highlighting the need to prevent data leakage. Similar attacks are user inference attacks that seek to gain knowledge or insights about the model or data’s characteristics, often by observing the model’s responses or behavior [129].

Last but not least, we explore denial of service (DoS) attacks and supply chain vulnerabilities. Yao et al. [129] describe DoS attacks as a type of cyber attack that aims to exhaust computational resources, resulting in latency or making the technology resources unavailable. In this survey, we focus on distributed denial of service attacks (DDoS), a type of DoS attack where requests flood the system, attacking simultaneously from multiple sources on the network [29]. Yao et al. [129] also defined LLM supply chain vulnerabilities as the risks in the lifecycle of LLM applications that may occur from using vulnerable components or services, including third-party plugins that may be used to steal chat histories, access private information, and or execute code on a user’s machine. All of these security vulnerabilities are substantial threats to LLMs that need to be mitigated or prevented. Possible methods of defense are discussed in Section 4.1, using current blockchain frameworks and experiments for these security problems, as listed by each developmental phase of the LLM, AI inherent threats, and supply chain issues.

## 4. EXISTING LITERATURE ON BC4LLM

Independently, the fields of both LLMs and blockchain research have grown substantially over the past several years. It is no surprise that the literature surrounding these topics has begun to morph and relate to each other. In previous research, we have seen LLMs for Blockchain Security [41] as well as an introduction to the term BC4LLM in Luo et al [72] where they provide a comprehensive survey of blockchain for LLMs. However, they do not acknowledge the multitude of safety and security solutions that blockchain provides for certain LLM vulnerabilities. Effectively, Luo et al. [72] aim to introduce BC4LLM for trusted AI, enabling reliable learning corpora, secure training processes, and identifiable generated content. In juxtaposition, our survey aims to analyze possible BC4LLM solutions closely related to our definitions of safety (2) and security (1) when looking at inherent system vulnerabilities in LLMs. To begin our analysis, we define these security problems based on previous work and highlight areas of research that are applicable to areas of BC4LLM safety and security.

### 4.1 Blockchain for LLMs’ Security

Few papers and experiments analyze how the integration of these two technological powerhouses interacts with one another. We have seen benefits of this integration that apply to our definition of security (1). Baliya et al. [10] introduce a peer-to-peer (P2P) federated LLM, namely PageRank, which works with a blockchain. This system operates in a fully decentralized capacity. Demonstrably, the blockchain implementation led to more efficient accuracy and latency results. With that being stated, Baliya et al. [10] provide

a developing direction in the field of BC4LLM to enhance system security. Below, we address several current vulnerabilities in LLMs and analyze them individually. In order to better understand these security problems, we categorize these vulnerabilities to their respective LLM training stages, highlight blockchain for AI works, and provide well-researched blockchain applications as a solution.

Vulnerabilities are present at each step in the process of developing a LLM. In early methods of model training, we encounter adversarial attacks such as data poisoning and backdoor attacks within the corpus [99; 127]. Progressing into model fine-tuning and general use, the LLM can fall victim to prompt-based attacks [140; 2; 127; 68], inference attacks [55; 44], and RAG-related attacks [126; 22; 5; 26; 149; 132]. These attacks are common vulnerabilities in both LLMs and AI since LLMs and AI are closely related as seen in Figure 2. Some of the threats against LLMs can be addressed by implementations from blockchain for AI (BC4AI) research, as elaborated below in Section 4.1.3. Considering the volume of potential attacks against LLMs, we make a further distinction of solutions that are specifically related to BC4LLM research and other blockchain-based solutions from BC4AI research. With this, we are able to highlight shared vulnerabilities for LLMs and AI. We provide an analysis of how blockchain can help defend against and mitigate these vulnerabilities, starting with threats during each phase of LLM training and utilization, continuing onto different blockchain solutions for AI inherent threats, and lastly noteworthy technology inherent attacks such as denial of service (DDoS) attacks and issues with supply chain logistics.

#### 4.1.1 Blockchain for Threats in LLMs’ Training

To mitigate the threats in LLMs’ training, data selection stands as a crucial aspect of model development, with particular emphasis on ensuring that training data is authentic, safe, and resistant to data poisoning attacks. One potential approach could be enabling the LLM to “unlearn” poisoned data or data deemed unsafe based on our definition of safety 2. Zuo et al. [150] establish federated TrustChain to enhance LLMs’ training and unlearning through a blockchain-based federated learning framework. Through integration with Hyperledger, the framework can efficiently perform unlearning, reducing the accuracy to 0.70% after unlearning given that the initial accuracy is 99.15% [150]. This demonstrates the potential of applying blockchain techniques to improve the security and privacy of LLMs, where LLMs can selectively forget specified data points while simultaneously preserving the performance via Low-Rank Adaptation (LoRA) and tuning hyper-parameters. This method of a blockchain-enabled federated unlearning process is further detailed as a future research possibility that has been thoroughly explored by few, as emphasized later in Section 7.1. Another significant issue is data poisoning. To address this issue, Gong et al. [36] propose a possible blockchain solution, introducing dynamic large language models (DLLM) on blockchains. Instead of using the traditional centralized datasets that LLMs are provided with, developing LLMs on blockchains enables the creation of decentralized datasets. These datasets are less likely to be tampered with and can be easily audited for accuracy. Gong et al. [36] present the DLLM to evolve after the training process. This was implemented by adjusting neural network parameters, enabling the LLM to continue learning during its use.



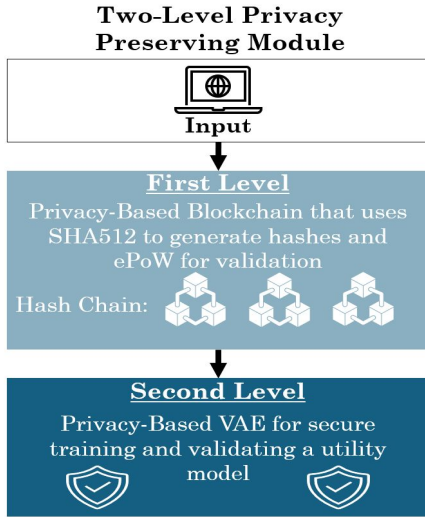


Figure 4: An overview of a two-level framework, consisting of a privacy-based blockchain that uses the secure hash algorithm 512 (SHA512) to generate hashes for data integrity. Then, the enhanced proof of work (ePoW) is used to authenticate data records and prevent data poisoning attacks from altering original data. These hashes of data blocks are linked to each other, called a Hash Chain. Then, for the second level, a privacy-based variational autoencoder (VAE) for secure data transformation ensures robust protection against inference attacks while maintaining the utility model for anomaly detection.

Additionally, blockchain-based systems can help assess where data poisoning may occur, and as shown in [55], blockchain can protect datasets and detect potential inference attacks through a two-level privacy preserving module. This research proposes a framework based on blockchain and deep learning, including two levels of privacy mechanisms as shown in Figure 4. For the first level, Keshk et al. [55] use SHA512 to generate secure hashes and then implement an enhanced-proof-of-work (ePoW) technique for authenticating and preventing data poisoning attacks. The second level consisted of a VAE model for converting original data into an encoded format for mitigating inference attacks that could be learned from system-based machine learning. In their testing, these mechanisms were effective in preventing data poisoning and inference attacks from manipulated smart power network datasets. BC4LLMs could benefit from this similar type of implementation, working with secure methods of hashing and blockchain-based deep learning privacy preservation techniques. By integrating a two-level privacy-preserving module, BC4LLMs can ensure data integrity and confidentiality while effectively detecting and mitigating both data poisoning and inference attacks.

Poisoned data has been an interest with RAG in particular. For example, Xue et al. [126] develop a way to identify security vulnerabilities from a poisoned corpus, but they do not use blockchain as a solution. We address the absence of research on blockchain and RAG, especially when using blockchain to help prevent RAG security issues. We discuss this as a possible future research direction as there remains a current gap in research of blockchain-based RAG systems and elaborate on this topic in Section 7.2. Poisoned

data overall is a major concern within LLMs and we offer blockchain as a potential source of ground truth to aid in mitigating this threat during the pre-training stages of LLMs and potentially mitigate RAG security concerns. In addition to data poisoning, LLMs are susceptible to backdoor attacks hidden in the training data during the LLM pre-training phase. Zhao et al. [140] introduce ProAttack which improves the stealth of backdoor attacks by accurately labeling poisoned data samples. As these attacks improve and become more sophisticated, it is crucial to explore robust defense mechanisms for LLMs. Few defense mechanisms using blockchain techniques have been studied, while Li et al. [57] propose a blockchain-based federated-learning framework (DBFL) that withstands backdoor attacks in a blockchain environment by incorporating an RLR aggregation strategy into the aggregation algorithm of a user and the addition of gradient noise to limit the effectiveness of backdoor attacks. The robustness of FL against backdoor attacks is enhanced by using various blockchain functions, including digital signature verification and simulation of chain resynchronization [57].

#### 4.1.2 Blockchain for Threats in LLMs' Prompting and Utilization

LLMs are often further trained through techniques such as instruction tuning, alignment tuning, and fine-tuning, each of which may introduce specific vulnerabilities, such as prompt injection [77; 101] and backdoor attacks [57]. These adversarial attacks are included under the term active privacy 3 where a malicious user attempts to gain unauthorized access. Blockchain technology, with its inherent transparency and immutability, holds the potential to mitigate and defend against these vulnerabilities. For instance, blockchain technology can be used to defend against prompt injections by ensuring data integrity and traceability. Mbula et al. [77] provide an overview of LLMs for blockchain, highlighting how blockchain's transparency and immutability enable a reliable audit trail for tracking and investigating suspicious activities. While not specifically focused on prompt injection, this approach demonstrates how blockchain can enhance security by providing a transparent and immutable record of interactions. Applying this to BC4LLMs can help prevent suspicious users from continuously interacting with an LLM, allowing for traceability to stop the user from entering malicious prompts. We recognize prompt injection is a critical vulnerability in LLMs and AI-related systems, yet as noted in [101], few blockchain defenses for prompt injection are present in the current field of research.

Inference attacks are also a critical concern for LLMs and active privacy, as malicious users may attempt to extract sensitive data from the model, hence why the Taxonomy 3 has inference attacks standalone. As discussed previously in Section 4.1.1, Keshk et al. [55] apply Blockchain and DL techniques to preserve privacy and prevent inference attacks through a framework as depicted in Figure 4. For more inference attack applicable work, a survey [44] thoroughly discusses membership inference attacks on ML and provides a group of defenses including differential privacy, regularization, confidence masking, and knowledge distillation. In other related works, there are instances of blockchain-based differential privacy methods [143; 37; 87], but current research that uses blockchain-based differential privacy frameworks to prevent inference attacks is limited; It is worth

noting that the theoretical foundation and potential synergies of this combination are promising. Another area with limited research is blockchain as a defense for jailbreaking attacks, which exploit the inherent capabilities of LLMs to bypass restrictions. There are multiple articles defending LLMs from jailbreaking attacks, yet little to none fully include blockchain to prevent jailbreaking. Hu et al. [45] explores a blockchain defense mechanism for malware checking on operating systems, indicating a possible direction for future research in integrating blockchain to defend against jailbreaking in LLMs. As previously explained in Section 3.1, backdoor attacks after the model has been trained are based on modifications to the trained model. The key blockchain-based federated-learning framework from Li et al. [57] discussed in detail in Section 4.1.1 used a combination of a blockchain environment and an RLR aggregation strategy to defend against backdoor attacks. This framework effectively coordinated FL processes and maintained learning security and user privacy. When testing backdoor attacks caused by malicious participants, the accuracy of the model increased when using the RLR aggregation strategy [57]. Given these findings, the possibility of leveraging blockchain transparency and immutability presents a robust mechanism for improving LLM security against active privacy threats. However, comprehensive integration and empirical validation of blockchain-based defenses in LLMs remain imperative to advance the field of BC4LLMs.

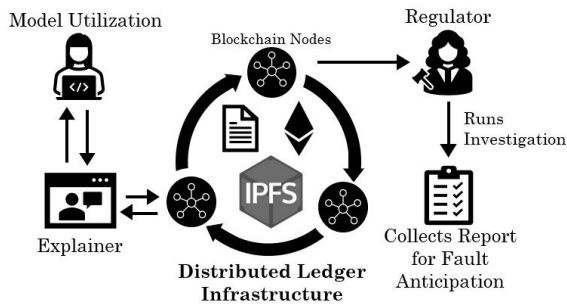


Figure 5: Within BXAI, this diagram illustrates the framework that leverages a distributed ledger infrastructure, using the Ethereum Blockchain and IPFS for storage and secure, traceable transactions. Depicted is the interaction between model utilization: an explainer generating local post-hoc explanations, the storage of these explanations in IPFS, and their linkage to the Blockchain. The use of smart contracts helps secure and encrypt the data, then relay it to a regulator who investigates current explanations, ensuring accountability and fault anticipation. Blockchain nodes are used to facilitate the secure and transparent broadcast of events within the Ethereum network.

#### 4.1.3 AI-intrinsic Threats and Defenses

AI intrinsic threats apply to LLMs due to the proximity of LLMs and AI, as shown in Figure 2. Blockchain for AI (BC4AI) is an emerging technology, with blockchain-based solutions already being researched as a secure way to establish trust in the Internet of Things (IoT) [103; 24]. Before BC4AI, some previous works refer to the integration as “On-chain AI” [28; 23]. Research of BC4AI encapsulates other machine learning techniques, such as blockchain-based federated learning and blockchain for deep learning. Federated

Learning is an addition to machine learning, as noted in Figure 2, where federated learning uses a privacy-preserving and decentralized approach to centralized systems.

A substantial literature on BC4AI has emerged from 2018 to 2024 [28; 116; 70; 124; 30; 94; 109; 74; 18; 11; 13]. Among the most notable works, Salah et al. [94] state the integration benefits of BC4AI. For example, there are five main benefits, such as enhanced data security, improved trust in robotic decisions, collective decision-making, decentralized intelligence, and high efficiency [94]. For enhanced data security, information stored within a blockchain is considered highly secure. By storing sensitive and personal data in a distributed, disk-less environment, blockchain can work alongside AI algorithms to strengthen data protection and promote more trusted and credible decision outcomes. The other benefits of improving trust within AI decision-making involve using the blockchain as a record of the decision-making process, allowing better AI traceability to analyze the quality of responses. Secondly, Dinh and Thai [28] summarize the integration of blockchain and AI to where blockchain can assist AI in multiple aspects, as follows. AI can benefit in secure data sharing from blockchain, allowing transparency and accountability regarding which user’s data is accessed, when, and by whom, letting users maintain control of their personal data. Among other data concerns, with the integration of blockchain and AI, blockchain technologies allow users to trade data via smart contracts, enabling the possibility of data marketplaces without a centralized middleman, making the transactions private and secure between users. Besides, Malhotra et al. [74] propose a blockchain-based proof-of-authenticity framework for explainable AI (XAI) utilizing a public Ethereum Blockchain, smart contracts, and IPFS (Interplanetary File System) to ensure secure, traceable, auditable transactions within the Ethereum network. This framework highlights three major components, smart contracts, an Ethereum and IPFS interconnected network, and a regulator, as depicted in Figure 5. Using smart contracts can enable continuous monitoring and tracing by all peers, in the case of any rule violations there are prompt rebound transactions to restore the system to an optimal state. To address the size limitation of storage on the blockchain, as further discussed in Section 6.1, Malhotra et al. [74] apply unique IPFS hashes stored on the Ethereum Blockchain to access larger-sized explanations that are stored off-chain in IPFS. These hashes are encrypted with the SHA256 algorithm to maintain data security. Thus, only entities with the corresponding hash can access and retrieve the IPFS hash and the associated explanation, ensuring controlled access even in a distributed network. Lastly, the regulator’s role is responsible for auditing and has access to the explanations to predict the user at fault using audit trails if system failure were to occur.

#### 4.1.4 Non-AI Threats and Defenses

Referring to Figure 3, we specifically focus on DDoS attacks and supply chain issues. Even though these attacks are common problems, we consider threats relevant to BC4LLMs. Ibrahim et al. [48] suggest using a public blockchain to prevent DDoS attacks on IoT devices. Blockchain provides a tamper-proof platform as well as demonstrates how IoT devices working with blockchain can verify and authenticate using a trusted white-list which is implemented in the smart contract. Following this smart contract usage, if LLMs were

to use a trusted white list for users then we can try to prevent these malicious users from trying to access the LLM in certain circumstances that are mutually agreed upon. Additionally proven by Shah et al. [97], blockchain-based solutions play a vital role in mitigating DDoS attacks.

A point of consideration is how DDoS attacks that target the blockchain to make the blockchain unavailable would require sufficient computer resources. The fully decentralized architecture of the blockchain and the consensus protocol for new blocks ensure that the blockchain can still operate meanwhile several blockchain nodes could be offline [135]. Incorporating this architecture into LLMs would help prevent DDoS attackers, as the larger the blockchain network is, then the harder it would be for a DDoS attack to be successful. Moreover, blockchain is known as a distributed, immutable, and verifiable ledger technology that ensures transparency and traceability [93]. By utilizing blockchain for LLMs, we can help mitigate these supply chain vulnerabilities. The decentralization of the network can maintain the integrity of the system at all points, aiding in mitigating the risk of a single point of failure, a common problem with centralized systems [93]. Blockchain is offered as a solution if the LLM were to accidentally crash, or was purposefully attacked by an attempt at overwhelming the system, then the LLM would still be intact since it is blockchain-based, removing the single point of failure entirely. However, it is important to note that blockchain solutions for LLMs depend on the availability of the underlying LLM infrastructure. If the LLM server is malfunctioning or shuts down, then these blockchain mechanisms may not be applicable, highlighting the need for a robust and resilient supply chain. To solidify the supply chain, blockchain offers secure transactional data in sectors including supply chain management, healthcare, and federated learning [150]. For better supply chain management and data traceability, Kayikci and Khosgoftaar [54] address the potential intersection of blockchain and ML. ML can aid in analyzing data from multiple sources and identify potential supply chain issues such as delays or quality issues before they occur. By using blockchain to create a transparent record of all supply chain transactions there are improvements in security, openness, traceability, and productivity [54]. While blockchain presents a promising solution for enhancing security and defending against adversarial threats to LLMs, ongoing research and development are necessary to address the evolving landscape of threats and vulnerabilities.

## 4.2 Blockchain for LLMs’ Safety

The growing dominance of LLMs as search engines [90], code writers [145], and in many other roles has introduced unique challenges related to their safety. For instance, LLMs who advise users to engage in dangerous activities such as eating glass [39] or which easily reveal personally identifying information [56] may be unsafe for users to interact with even in the absence of external threats. In this section, we rely on our proposed definition of safety (Definition 2) to explore relevant literature that incorporates blockchain technology into the various solutions surrounding LLM safety.

### 4.2.1 Blockchain for Passive Privacy in LLMs

Despite its novelty, the concept of passive privacy is crucial for ensuring the safety of LLMs. Some models risk leak-

ing sensitive information, potentially exposing private data like government-issued ID numbers and patient records [86]. The severity of these leaks underscores the need for effective solutions to advance LLMs responsibly. In this regard, blockchain’s guarantees of data sovereignty, obfuscation, and traceability offer practical passive privacy benefits that align well with the requirements of LLMs. In particular, we observe blockchain-based privacy preservation techniques which originate in varying proximity to LLMs as seen in BC4LLMs itself [113; 118; 102], blockchain-enabled deep learning [147; 55; 121; 120; 96], blockchain-enabled machine learning [7], and blockchain-enabled federated learning [8; 81; 142; 88; 31; 89; 73].

Within our focus on BC4LLMs, we have observed distinct trends in the application of blockchain to LLMs in their capacity to bolster passive privacy guarantees. Most notably, the development of zero-knowledge LLMs, a.k.a. ZK-LLMs, as described in [118] and [102], has the potential to drastically reduce privacy leakage risks when interacting with LLMs. Considering the problem of data leakage approached from the lens of access, this application is natural. A user querying for their own personally identifiable information should, ideally, be able to access it whereas an unauthorized user should not. Obfuscating portions or the entirety of a corpus using zero-knowledge proofs [133] allows for untrusted training nodes, or the model itself, to act on sensitive data without the ability to regurgitate it to a potentially malicious party. This same mechanism has broad applications that have been explored in other recent works as well, with special focuses on ZKPs for data curation and pre-processing [113], which consequently enhance passive privacy within LLMs.

Additionally, besides material on BC4LLMs, it is necessary to discuss passive privacy contributions made within LLM-related areas, as described in our classification of LLMs in the context of AI, ML, and DL (Figure 2). Especially important in its immediate applications to LLMs, blockchain-enabled deep learning (BC-DL) is a growing field with potentially large impacts on LLM’s passive privacy. Specifically, certain BC-DL technologies propose learning mechanisms distinct from traditional federated learning models [55; 121; 96]. The concerted research effort to develop efficient distributed learning models that deviate from the typical model of federated learning is clearly well underway. This field has broad implications for blockchain; through the utilization of various blockchain properties, we see the development of privacy guarantees which undoubtedly strengthens the BC4LLMs area.

A noteworthy contribution in the field of blockchain and deep learning is the influential DeepChain [121], which introduces a novel privacy-preserving training framework based on blockchain technology. This system employs a consensus protocol alongside an incentive mechanism, enabling the use of private training gradients and ensuring the auditability of training data. This dual approach, incorporating zero-knowledge proofs in various aspects of the protocol, represents a promising new direction for blockchain-enabled passive privacy, building upon the well-established domain of federated learning. To underscore the novelty of this approach, Figure 6 illustrates how the system operates primarily through consensus and incentivization mechanisms. Additionally, despite this potential research direction, discussion on the wide body of research that does exist concern-

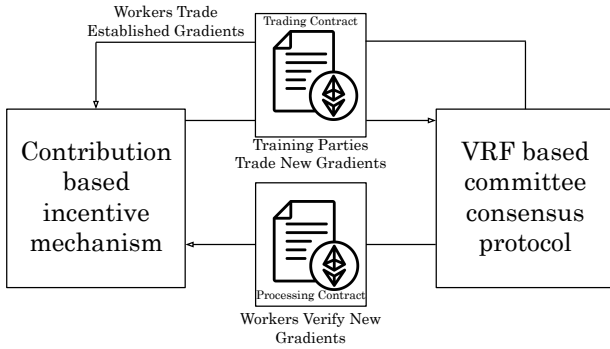


Figure 6: DeepChain deviates from conventional federated learning models by incorporating a synchronous requirement for consensus on data incorporated into the model during the final training round. The contribution-based incentive mechanism rewards participants for verifying new gradients and activates a trading contract to facilitate the sharing of updated gradient information. Besides, a Verifiable Random Function (VRF) ensures fairness in the committee-based consensus process, addressing concerns related to finality.

ing blockchain-enabled federated learning (BC-FL) is essential when describing blockchain as a vehicle for improving the safety and security of LLMs. Federated learning, introduced by McMahan et al. [78], has since been the principal building block of decentralized learning approaches in machine learning systems. While too broad to be considered for BC4LLMs, notable to this paper are the contributions of authors approaching BC-FL in its capacity as a powerful privacy preserving mechanism [81; 8; 142; 31; 32; 73].

#### 4.2.2 Blockchain for Misinformation in LLMs

The issue of LLMs fabricating information, commonly known as hallucinations [75; 3] is well understood. As a result, detecting and defending against hallucinations is a widely explored area [95], with more research still yet to be conducted [131; 136]. Along with this pressing general body of research, efforts have been made to leverage blockchain technology to reduce hallucinations by consensus-oriented [138] and oracle-based [15] approaches. Within consensus, Zhang et al. [138] proposed a system for efficient large language model inference quality assessment. That is, the veracity of a given model’s responses was able to be assessed by using a ‘Proof of Quality’ consensus mechanism with low latency between the user and language model. This stands in contrast to other approaches, such as Bouchiha et al.’s [15] reputation-based system LLMChain, which relies on a decentralized oracle to cross reference request/response pairs originating from differing models and speak to the quality of inferences based on those comparisons. It is worth noting that despite these fundamental differences, they are both consensus-based approaches. This serves as an excellent example of how BC4LLM technology can take many different forms towards the same goal.

In addition to advancing consensus-driven governance models for mitigating misinformation, several approaches have been explored to improve the accuracy of LLMs’ responses. These efforts include contributions from the zero-knowledge domain, as demonstrated by Chen et al. [21], more op-

timistic privacy assurances found in works like [23], and straightforward applications of verifiable ledgers, as proposed by Yazdinejad et al. [130]. Chen et al. [21] propose zkML, a compiler that enables TensorFlow models to be translated readily into zk-SNARK halo2 circuits via either KZG or IPA commitments. This conversion allows for any portion, or the entirety of, an LLM to gain the properties of zero-knowledge, knowledge soundness, and completeness. Through this, and with potential connections to verifiable databases, zkML gains the powerful ability to audit inferences and ensure their accuracy. This research avenue is particularly promising due to both the efficient and potentially on-chain verification of zk-SNARKs as well as the extensibility of zkML to virtually any ML model.

Distinct from the zkML approach is opML [23], which opts for an optimistic approach reliant on a fraud-proof rather than a ZK proof to catch erroneous outputs within a certain challenge period. Clearly, there exist trade-offs in this implementation when compared to the zkML approach. Optimistic rollups are desirable in the sense that they are performant, but if implemented in a RAG environment, or similarly situated between the user and a model, latency issues can quickly become dominant. Apart from proof-oriented mechanisms and worth noting is the work proposed by Yazdinejad et al. [130], which focuses on detecting deep-fakes using blockchain’s verifiable ledger. While not directly applicable to the realm of LLMs due to the non-atomic nature of data within a language model, important insights can be drawn from the paper. Namely, BC4LLMs could benefit greatly from a proposed hashing method applied to particularly sensitive data areas such as names, addresses, or even health-care-related parts of corpora. This hash could be used as a guarantee of data veracity and could potentially prevent unsafe behaviors such as sycophancy, deception, or unfairness. Indeed, this hashing mechanism has the potential to be used as a final check for the LLM to verify that it is submitting information to the user that is consistent with standards agreed upon when information was originally committed to the ledger. Many similar vehicles for the maintenance of data integrity exist, albeit currently limited by scaling issues on-chain [27].

## 5. DATASETS RELEVANT TO BC4LLM

Developing synergistic techniques that integrate blockchain with LLMs is essential for ensuring the safety and trustworthiness of future LLMs. In this context, it is critical to access to relevant datasets for experimentation. Blockchain-enabled systems often require unconventional training sets and edge cases to capture the dynamism and robustness of these implementations. Accordingly, we have compiled and summarized the relevance of specific datasets in Table 4. While we include standard datasets such as MNIST, CIFAR-10, SQuAD, and MS-MARCO, we also highlight lesser-known datasets that may prove valuable in particular research contexts.

## 6. CHALLENGES IN BC4LLM

Despite the promise of the emerging BC4LLMs field, there are several innate challenges that delay progress and inhibit potential research directions. Typically, these are derived

Table 4: **Datasets Relevant to BC4LLMs**

Dataset	Use Case	Description	Papers
MNIST <sup>1</sup>	Pattern Recognition	Images of handwritten digits for pattern recognition applications or vulnerability analysis.	[121; 119; 78; 31; 100; 58]
CIFAR-10 <sup>2</sup>		Labeled images used in capacities from improving pattern recognition to zk-SNARK benchmarks.	[119; 21; 78]
MS MARCO <sup>3</sup>		Collection of human answered questions, used in training corpora as well as simulating RAG attacks.	[126; 149; 22]
MedMINST <sup>4</sup>		Collection of medical images from case studies.	[126; 149; 22]
Natural Questions <sup>5</sup>	Poisoned RAG	Open domain question answering dataset, incorporating questions from users and rigorous answers.	[126; 149; 22]
HotpotQA <sup>6</sup>		Question answering dataset with multi-hop questions and supervised, regulated, answers.	[149; 22]
MT BENCH <sup>7</sup>	LLM Evaluation	Ranked pairwise expert human preferences for various model responses.	[144; 15]
SQuAD <sup>8</sup>		Reading comprehension dataset comprised of questions posed on Wikipedia article with answers as sections of those corresponding articles.	[150; 46]
IMDB Dataset <sup>9</sup>	Sentiment Analysis	Movie reviews	[150; 46]
SafetyBench <sup>10</sup>	Safety Evaluation	Large number of multiple choice questions focused on evaluating the safety of LLMs.	[137]
Tweets2011 <sup>11</sup>	Sensitive Information Handling	List of scraped tweet identifiers and corresponding tweets from early 2011.	[150]
MTSamples Scrape <sup>12</sup>		Sample transcription medical reports from various disciplines and areas.	[150; 46]
DRC Diplomas <sup>13</sup>		Highschool diplomas from the Democratic Republic of the Congo.	[10]
HealthCareMagic <sup>14</sup>		Real patient-doctor conversations found through the HealthCareMagic website, capturing the nature of patient vocabulary.	[5]
Enron Emails <sup>15</sup>		Large set of emails generated by employees of the Enron Corporation.	[150; 46]
LLMGooAQ <sup>16</sup>		Comprehensive database capturing question and answers from a wide variety of domains.	[15]
GooAQ <sup>17</sup>		Large scale question answering dataset aimed at developing a vast selection of question types.	[15]
The Pile <sup>18</sup>		Massive and open source data set consisting of a combination of roughly 20 other datasets.	[140]

<sup>1</sup><https://yann.lecun.com/exdb/mnist/> <sup>2</sup><https://www.cs.toronto.edu/~kriz/cifar.html> <sup>3</sup><https://microsoft.github.io/msmarco/><sup>4</sup><https://medmnist.com/> <sup>5</sup><https://ai.google.com/research/NaturalQuestions> <sup>6</sup><https://hotpotqa.github.io/><sup>7</sup><https://paperswithcode.com/dataset/mt-bench> <sup>8</sup><https://rajpurkar.github.io/SQuAD-explorer/><sup>9</sup><https://developer.imdb.com/non-commercial-datasets/> <sup>10</sup><https://github.com/thu-coai/SafetyBench><sup>11</sup><https://trec.nist.gov/data/tweets/> <sup>12</sup><https://mtsamples.com/> <sup>13</sup><https://minepst.gouv.cd/palmares-exetat/><sup>14</sup><https://huggingface.co/datasets/RafaelMPereira/HealthCareMagic-100k-Chat-Format-en><sup>15</sup><https://huggingface.co/datasets/preference-agents/enron-cleaned> <sup>16</sup><https://github.com/mohaminemed/LLMGooAQ/><sup>17</sup><https://huggingface.co/datasets/allenai/gooaq> <sup>18</sup><https://pile.eleuther.ai/>

from certain limitations in blockchain technology, LLMs, or deficits in the way that blockchain can serve LLMs.

## 6.1 Corpus on Blockchain

LLMs’ training heavily relies on substantial data, with modern corpora typically exceeding dozens of terabytes in volume [141]. This characteristic is inherently misaligned with the constraints that blockchain systems are generally designed to address. Reconciling blockchain’s limitations in throughput and data-handling capacity with the extensive data requirements of LLMs represents one of the most pressing challenges in BC4LLMs. Several approaches have explored the use of zero-knowledge proofs (ZKPs) to enhance scalability [118; 102]. However, relying on zero-knowledge technology solely for scalability, and not privacy, poses challenges due to the computational cost of generating ZKPs, even with minimal circuits. A significant factor here is the ongoing issue of the Multi-Scalar Multiplication (MSM) in ZKP generation [125]. Furthermore, current WebGPU and WASM implementations likely fall short of the throughput required by client-based LLMs. For these reasons, it is improbable that zero knowledge could serve as a definitive solution to scalability in BC4LLMs without significant advancements in zk-SNARK generation research. Addressing the discrepancy between the growing size of LLMs and blockchain’s limited capacity for on-chain data storage remains a substantial research challenge.

## 6.2 Reliance on Oracles

The security guarantees of blockchain technology, while robust, often leave little room for interoperability with external systems [27]. That is, the blockchain can most easily interact with information on the chain, leaving little room to consider issues such as fact-checking or moral alignment. Oftentimes, to develop mechanisms that seek to assist with LLM toxicity or factuality, oracles are used to bridge this gap [33; 30]. Serving as mediators between chains and online sources, oracles are trusted parties that deliver information through a variety of protocols and frameworks. However, introducing a trusted party into an otherwise trustless system has been a long-standing weak point in this solution [77]. Exploring non-oracle-based options for ground truth solutions, or toxicity checks, would greatly enhance the security guarantees of blockchain within LLMs.

## 6.3 Energy Consumption

A significant portion of the challenges associated with LLMs arises from their need to consume and process vast amounts of data [141]. This requirement, in turn, necessitates extensive energy consumption during both training and runtime [72]. On the other hand, blockchain systems face their own energy challenges, as consensus mechanisms and transaction validation processes often incur substantial computational costs [77]. The high computational demands of both LLMs and blockchain systems highlight a misalignment with the scalability of BC4LLM implementations without substantial efforts to reduce energy costs. This might require moving away from transformer-based architectures and energy-intensive consensus mechanisms, such as proof of work, toward more sustainable alternatives [84].

# 7. FUTURE RESEARCH DIRECTIONS

There exist several critically overlooked areas within LLMs that may benefit greatly from the introduction of blockchain technologies. The most prominent of these areas include blockchain federated unlearning, RAG, differential privacy, data provenance, and toxicity mitigation.

## 7.1 Blockchain Federated Unlearning

Privacy regulations are paramount in the online realm, especially concerning the “right to be forgotten” and user data privacy which are critical considerations when working with LLMs and blockchain. Federated blockchain unlearning offers LLMs the ability to erase learned data. Within our research, we identified four recent papers that have implemented blockchain-based federated unlearning frameworks. As noted previously in Section 4.1.1, Zuo et al. [150] develop a federated TrustChain framework for blockchain-enhanced LLM training and unlearning, focusing on the impact of Low-Rank Adaptation (LoRA) hyperparameters on unlearning performances and integrating Hyperledger Fabric to ensure the security, transparency, and verifiability of the unlearning process. In another study, Zuo et al. [151] presented a trustworthy approach towards federated learning with blockchain-enhanced machine unlearning. This implementation differs from Trustchain, where Zuo et al. [151] used a machine unlearning mechanism that utilized two types of clients for training and unlearning, smart contracts for process automation, and a blockchain network for secure, immutable record-keeping. Beyond the above works [151], Liu et al. [66] introduced Blockchain Federated Unlearning (BlockFUL) as a versatile framework that redesigns the blockchain structure using a Chameleon Hash (CH) technology to simplify model updates and reduce the computational and consensus costs associated with unlearning tasks. Additionally, BlockFUL ensures the integrity and traceability of model updates, including privacy-preserving results from these blockchain-based unlearning operations [66]. Lin et al. [63] propose a framework with a proof of federated unlearning protocol that also utilizes the Chameleon hash function to verify data removal and eliminate the data contributions stored in other clients’ models. Both use CH functions in their blockchain-enabled federated unlearning processes. The applications of key blockchain components, such as on-chain smart contracts and hash mappings for verifying data removal, may enable LLMs to forget personal data effectively. Blockchain for unlearning is an emerging area of research with significant potential for further innovation.

## 7.2 Blockchain-enhanced RAG

Considering the popularity of Retrieval Augmented Generation (RAG), extensive studies have emerged to focus on potential vulnerabilities within RAG that could compromise the integrity of LLMs [126; 5; 26; 149; 46; 132; 22]. However, a significant gap remains in addressing strategies to mitigate these attacks, particularly where blockchain technology could offer defensive benefits. Recent studies have called for exploring blockchain’s role in RAG deployment [9], and preliminary investigations have assessed blockchain’s potential to enhance user experience [134] and performance evaluation [87]. Nonetheless, dedicated efforts to strengthen security and safety within RAG systems are largely absent in the current literature. Advancing BC4LLMs specifically in the context of RAG security could yield considerable mutual

benefits for both blockchain and LLM technologies.

### 7.3 Blockchain for Privacy Guarantees in LLMs

The clear connection between federated learning, blockchain, and LLMs allows for the field of differential privacy to enter BC4LLMs' sphere of relevance. Major contributions concerning the impact of differential privacy on related areas such as deep learning have already been made [1], but issues such as privacy budget exhaustion still loom large in the space [14]. Moreover, despite conclusions that blockchain can help with privacy budget exhaustion [37; 143], few efforts have been conducted in exploring these solutions. Indeed, there is a need for more relevant research in order to realize the full measure of blockchain's impact on this area.

### 7.4 Blockchain for Data Provenance and Transparency in LLMs

Several recent papers have urged for increased data accountability measures to be placed on organizations developing LLMs, especially where it concerns issues of data acquisition [12; 40]. Additionally, worth noting are direct calls for the introduction of blockchain technology to help solve the issue of data provenance [122] in LLMs. Largely, while this has been answered with responses in the realms of auditability [57], straightforward data tracking solutions have remained absent from the literature, despite relatively simple conceptual formulations [28]. Towards this goal of achieving improved data provenance within LLMs corpus', RAG databases, and even in-context learning repositories, there is a need for more explorations into this natural application of distributed ledger technology to problems of explainable AI concerning LLMs.

### 7.5 Blockchain for Non-toxic LLMs

Encompassing vital attributes such as ethics, legality, and non-violence, developing non-toxic LLMs has been and will continue to be a major focus of the field for the foreseeable future [101; 69; 38]. There is no doubt that automated filtering of generated toxic content is one of the most pressing challenges concerning the safety of LLMs [35; 6]. This is because in essence, filtering inferences negatively impacts the quality of LLM responses, whereas manual human annotation is a costly and complex process [6]. Therefore, the applications of blockchain technology in this regard, while currently limited, are compelling. Considering one of the most groundbreaking achievements in ML within the past several years, federated learning has allowed for massive strides to be made within the spaces of securing training sets, user privacy, and even misinformation defense. A similar approach, aimed at the problem of toxicity, could be a hugely beneficial endeavor to the field. Moreover, imagining such a model is not difficult. Developing consensus around what is considered correct in a model and using that to propagate gradients and parameters is not dissimilar to the decisions that must be made about what is or is not toxic given the state of certain corpora. Given a concentrated research program, automated non-toxicity could very well have excellent solutions found in the blockchain space.

## 8. CONCLUSION

In this survey, we first highlight significant systemic vulnerabilities in large language models (LLMs), including data poisoning, hallucinations, jailbreaking, and privacy attacks. Although these issues have been extensively studied in conventional machine learning models, with approaches like differential privacy and federated learning, comprehensive protection for LLMs remains an area for improvement. In contrast, blockchain technology offers a promising solution to enhance the security and safety of LLMs. Blockchain systems provide powerful mechanisms to ensure data integrity, provenance, and encrypted frameworks, which can be leveraged to strengthen LLM defenses. By integrating blockchain-based defenses, it is possible to achieve stronger privacy protection, reliable data, and improved resilience of LLMs against adversarial threats.

Besides, it is critical to establish clear definitions of security and safety in the context of LLMs. We conclude that security for LLMs pertains to the ability to tolerate applicable adversarial attacks while maintaining system integrity to provide consistent and accurate responses, whereas safety for LLMs is the model's capacity to interact with users in a trustworthy manner, contingent upon adhering to ethical concerns, law-abiding, non-violent, fair, passively privacy-preserving, and informing. Additionally, differentiating between active and passive privacy measures will aid in developing more targeted and effective privacy-preserving strategies. These distinctions and definitions provide a foundational framework for future research in BC4LLM. From analyzing the integration of blockchain and LLMs, we propose a new taxonomy in Figure 3, where previous research done in the field of BC4LLMs can apply to security and safety problems that LLMs face. We recognize various gaps in BC4LLMs that need to be looked into for further consideration. In refining our understanding of relevant concepts, we see that the intersection of blockchain and LLMs holds significant potential for addressing the current shortcomings in LLM security and safety. Through our review, we aim to guide new researchers in understanding how blockchain technology can be utilized to enhance the security, reliability, and safety of LLMs.

## 9. REFERENCES

- [1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, page 308–318, 2016.
- [2] S. Abdali, R. Anarfi, C. Barberan, and J. He. Securing large language models: Threats, vulnerabilities and responsible practices. *arXiv preprint arXiv:2403.12503*, 2024.
- [3] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2024.
- [4] O. Ali, A. Jaradat, A. Kulakli, and A. Abuhalimeh. A comparative study: Blockchain technology utilization benefits, challenges and functionalities. *IEEE Access*, 9:12730–12749, 2021.

- [5] M. Anderson, G. Amit, and A. Goldstein. Is my data in your retrieval database? membership inference attacks against retrieval augmented generation. *arXiv preprint arXiv:2405.20446*, 2024.
- [6] U. Anwar, A. Saparov, J. Rando, D. Paleka, M. Turpin, P. Hase, E. S. Lubana, E. Jenner, S. Casper, O. Sourbut, et al. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*, 2024.
- [7] P. C. M. Arachchige, P. Bertok, I. Khalil, D. Liu, S. Camtepe, and M. Atiquzzaman. A trustworthy privacy preserving framework for machine learning in industrial iot systems. *IEEE Transactions on Industrial Informatics*, 16(9):6092–6102, 2020.
- [8] S. Awan, F. Li, B. Luo, and M. Liu. Poster: A reliable and accountable privacy-preserving federated learning framework using the blockchain. pages 2561–2563, 2019.
- [9] A. Balakrishnan. Enhancing data engineering efficiency with ai: Utilizing retrieval-augmented generation, reinforcement learning from human feedback, and fine-tuning techniques. *International Research Journal of Modernization in Engineering Technology and Science*, 6, 2024.
- [10] S. B. Balija, A. Nanda, and D. Sahoo. Building communication efficient asynchronous peer-to-peer federated llms with blockchain. *Proceedings of the AAAI Symposium Series*, 3(1):288–292, 2024.
- [11] N. Baranwal Somy, K. Kannan, V. Arya, S. Hans, A. Singh, P. Lohia, and S. Mehta. Ownership preserving ai market places using blockchain. In *2019 IEEE International Conference on Blockchain (Blockchain)*, pages 156–165, 2019.
- [12] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.
- [13] D. Bhumichai, C. Smiliotopoulos, R. Benton, G. Kambourakis, and D. Damopoulos. The convergence of artificial intelligence and blockchain: The state of play and the road ahead. *Information*, 15(5), 2024.
- [14] A. Bkakria, A. Tasidou, N. Cuppens-Bouahia, F. Cuppens, F. Bouattour, and F. Fredj. *Optimal Distribution of Privacy Budget in Differential Privacy*, pages 222–236. 2019.
- [15] M. A. Bouchiha, Q. Telnoff, S. Bakkali, R. Champagnat, M. Rabah, M. Coustaty, and Y. Ghamri-Doudane. Llmchain: Blockchain-based reputation system for sharing and evaluating large language models. *arXiv preprint arXiv:2404.13236*, 2024.
- [16] T. B. Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.
- [17] V. Buterin. Ethereum: A next-generation smart contract and decentralized application platform. *white paper*, 3(37), 2014.
- [18] D. Calvaresi, Y. Mualla, A. Najjar, S. Galland, and M. Schumacher. Explainable multi-agent systems through blockchain technology. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems: First International Workshop, EXTRAAMAS 2019, Montreal, QC, Canada, May 13–14, 2019, Revised Selected Papers 1*, pages 41–58. Springer, 2019.
- [19] B. Cao, Z. Wang, L. Zhang, D. Feng, M. Peng, L. Zhang, and Z. Han. Blockchain systems, technologies, and applications: A methodology perspective. *IEEE Communications Surveys & Tutorials*, 25(1):353–385, 2023.
- [20] N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramèr, B. Balle, D. Ippolito, and E. Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023.
- [21] B.-J. Chen, S. Waiwitlikhit, I. Stoica, and D. Kang. Zkml: An optimizing system for ml inference in zero-knowledge proofs. pages 560–574, 2024.
- [22] P. Cheng, Y. Ding, T. Ju, Z. Wu, W. Du, P. Yi, Z. Zhang, and G. Liu. Trojanrag: Retrieval-augmented generation can be backdoor driver in large language models. *arXiv preprint arXiv:2405.13401*, 2024.
- [23] K. Conway, C. So, X. Yu, and K. Wong. opml: Optimistic machine learning on blockchain. *arXiv preprint arXiv:2401.17555*, 2024.
- [24] J. Cuomo. How blockchain adds trust to ai and iot, 2020.
- [25] C. Deng, Y. Duan, X. Jin, H. Chang, Y. Tian, H. Liu, H. P. Zou, Y. Jin, Y. Xiao, Y. Wang, et al. Deconstructing the ethics of large language models from long-standing issues to new-emerging dilemmas. *arXiv preprint arXiv:2406.05392*, 2024.
- [26] G. Deng, Y. Liu, K. Wang, Y. Li, T. Zhang, and Y. Liu. Pandora: Jailbreak gpts by retrieval augmented generation poisoning. *arXiv preprint arXiv:2402.08416*, 2024.
- [27] A. Deshpande, K. Stewart, L. Lepetit, and S. Gunasekar. Distributed ledger technologies/blockchain: Challenges, opportunities and the prospects for standards. *Overview report The British Standards Institution (BSI)*, 40(40):1–34, 2017.
- [28] T. N. Dinh and M. T. Thai. Ai and blockchain: A disruptive integration. *Computer*, 51(9):48–53, 2018.
- [29] S. B. ElMamy, H. Mrabet, H. Gharbi, A. Jemai, and D. Trentesaux. A survey on the usage of blockchain technology for cyber-threats in the context of industry 4.0. *Sustainability*, 12(21):9179, 2020.
- [30] S. Fan, N. Ilk, A. Kumar, R. Xu, and J. L. Zhao. Blockchain as a trust machine: From disillusionment to enlightenment in the era of generative ai. *Decision Support Systems*, 182, 2024.
- [31] A. Ferenczi and C. Bădică. A fully decentralized privacy-enabled federated learning system. In *Computational Collective Intelligence: 15th International Conference, ICCCI Proceedings*, pages 444–456, 2023.
- [32] R. Fotohi, F. Shams Aliee, and B. Farahani. Decentralized and robust privacy-preserving model using blockchain-enabled federated deep learning in intelligent enterprises. *Applied Soft Computing*, 161:111764, 2024.



- [33] P. Fraga-Lamas and T. M. Fernández-Caramés. Fake News, Disinformation, and Deepfakes: Leveraging Distributed Ledger Technologies and Blockchain to Combat Digital Deception and Counterfeit Reality. *IT Professional*, 2020.
- [34] D. Ganguli, L. Lovitt, J. Kernion, A. Askell, Y. Bai, S. Kadavath, B. Mann, E. Perez, N. Schiefer, K. Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.
- [35] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith. Realtotoxicityprompts: Evaluating neural toxic degeneration in language models. pages 3356–3369, 2020.
- [36] Y. Gong. Dynamic large language models on blockchains. *arXiv preprint arXiv:2307.10549*, 2023.
- [37] L. M. Han, Y. Zhao, and J. Zhao. Blockchain-based differential privacy cost management system. *arXiv preprint arXiv:2006.04693*, 2020.
- [38] T. Han, A. Kumar, C. Agarwal, and H. Lakkaraju. Towards safe large language models for medicine. *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*, 2024.
- [39] S. Harrer. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *EBioMedicine*, 90, 2023.
- [40] K. He, R. Mao, Q. Lin, Y. Ruan, X. Lan, M. Feng, and E. Cambria. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *arXiv preprint arXiv:2310.05694*, 2024.
- [41] Z. He, Z. Li, S. Yang, A. Qiao, X. Zhang, X. Luo, and T. Chen. Large language models for blockchain security: A systematic literature review. *arXiv preprint arXiv:2403.14280*, 2024.
- [42] D. Hendrycks, C. Burns, S. Basart, A. Critch, J. Li, D. Song, and J. Steinhardt. Aligning ai with shared human values. 2020.
- [43] T. F. Heston. Prespective chapter: Integrating large language models and blockchain in telemedicine. 2024.
- [44] H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. S. Yu, and X. Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37, 2022.
- [45] Q. Hu, M. R. Asghar, and S. Zeadally. Blockchain-based public ecosystem for auditing security of software applications. 103(11):2643–2665, 2021.
- [46] Z. Hu, C. Wang, Y. Shu, P. Helen, and L. Zhu. Prompt perturbation in retrieval-augmented generation based large language models. *arXiv preprint arXiv:2402.07179*, 2024.
- [47] X. Huang, W. Ruan, W. Huang, G. Jin, Y. Dong, C. Wu, S. Bensalem, R. Mu, Y. Qi, X. Zhao, K. Cai, Y. Zhang, S. Wu, P. Xu, D. Wu, A. Freitas, and M. A. Mustafa. A survey of safety and trustworthiness of large language models through the lens of verification and validation. 57(7), 2024.
- [48] R. F. Ibrahim, Q. Abu Al-Haija, and A. Ahmad. Ddos attack prevention for internet of thing devices using ethereum blockchain technology. *Sensors*, 22(18), 2022.
- [49] H. Inan, K. Upasani, J. Chi, R. Rungta, K. Iyer, Y. Mao, M. Tontchev, Q. Hu, B. Fuller, D. Testugine, and M. Khabsa. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023.
- [50] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. Survey of hallucination in natural language generation. 55(12):1–38, 2023.
- [51] T. Jiang, Z. Wang, J. Liang, C. Li, Y. Wang, and T. Wang. Robustkv: Defending large language models against jailbreak attacks via kv eviction. *arXiv preprint arXiv:2410.19937*, 2024.
- [52] H. Jin, L. Hu, X. Li, P. Zhang, C. Chen, J. Zhuang, and H. Wang. Jailbreakzoo: Survey, landscapes, and horizons in jailbreaking large language and vision-language models. *arXiv preprint arXiv: 2407.01599*, 2024.
- [53] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günemann, E. Hüllermeier, S. Krusche, G. Kutyiniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, M. Stadler, J. Weller, J. Kuhn, and G. Kasneci. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 2023.
- [54] S. Kayikci and T. M. Khoshgoftaar. Blockchain meets machine learning: A survey. *Journal of Big Data*, 11(1), 2024.
- [55] M. Keshk, B. Turnbull, N. Moustafa, D. Vatsalan, and K.-K. R. Choo. A privacy-preserving-framework-based blockchain and deep learning for protecting smart power networks. 16(8):5110–5118, 2020.
- [56] S. Kim, S. Yun, H. Lee, M. Gubri, S. Yoon, and S. J. Oh. Propile: Probing privacy leakage in large language models. *Advances in Neural Information Processing Systems*, 36, 2024.
- [57] L. Li, J. Qin, and J. Luo. A blockchain-based federated-learning framework for defense against backdoor attacks. 12(11), 2023.
- [58] Z. Li, H. Yu, T. Zhou, L. Luo, M. Fan, Z. Xu, and G. Sun. Byzantine resistant secure blockchained federated learning at the edge. *IEEE Network*, 35(4):295–301, 2021.
- [59] J. Liang, S. Li, B. Cao, W. Jiang, and C. He. Omnilytics: A blockchain-based secure data market for decentralized machine learning. *arXiv preprint arXiv:2107.05252*, 2021.
- [60] Z. Liang, J. Cheng, R. Yang, H. Ren, Z. Song, D. Wu, X. Qian, T. Li, and Y. Shi. Unleashing the potential of llms for quantum computing: A study in quantum architecture design. *arXiv preprint arXiv:2307.08191*, 2023.
- [61] F. Lin, S. Crawford, K. Guillot, Y. Zhang, Y. Chen, X. Yuan, L. Chen, S. Williams, R. Minvielle, X. Xiao, et al. Mmst-vit: Climate change-aware crop yield prediction via multi-modal spatial-temporal vision transformer. In *Proceedings of the IEEE/CVF International*

- Conference on Computer Vision*, pages 5774–5784, 2023.
- [62] F. Lin, X. Yuan, Y. Zhang, P. Sigdel, L. Chen, L. Peng, and N.-F. Tzeng. Comprehensive transformer-based model architecture for real-world storm prediction. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 54–71. Springer, 2023.
- [63] Y. Lin, Z. Gao, H. Du, J. Ren, Z. Xie, and D. Niyato. Blockchain-enabled trustworthy federated unlearning. *arXiv preprint arXiv:2401.15917*, 2024.
- [64] B. Liu, M. Ding, S. Shaham, W. Rahayu, F. Farokhi, and Z. Lin. When machine learning meets privacy: A survey and outlook. *ACM Computing Surveys (CSUR)*, 54(2):1–36, 2021.
- [65] M. Liu, S. Ho, M. Wang, L. Gao, Y. Jin, and H. Zhang. Federated learning meets natural language processing: A survey. *arXiv preprint arXiv:2107.12603*, 2021.
- [66] X. Liu, M. Li, X. Wang, G. Yu, W. Ni, L. Li, H. Peng, and R. Liu. Decentralized federated unlearning on blockchain. *arXiv preprint arXiv:2402.16294*, 2024.
- [67] X. Liu, J. Liang, L. Tang, C. You, M. Ye, and Z. Xi. Buckle up: Robustifying llms at every customization stage via data curation. *arXiv preprint arXiv:2410.02220*, 2024.
- [68] Y. Liu, G. Deng, Y. Li, K. Wang, Z. Wang, X. Wang, T. Zhang, Y. Liu, H. Wang, Y. Zheng, and Y. Liu. Prompt injection attack against llm-integrated applications. *arXiv preprint arXiv:2306.05499*, 2024.
- [69] Y. Liu, Y. Yao, J.-F. Ton, X. Zhang, R. Guo, H. Cheng, Y. Klochkov, M. F. Taufiq, and H. Li. Trustworthy llms: a survey and guideline for evaluating large language models’ alignment. *arXiv preprint arXiv:2308.05374*, 2024.
- [70] V. Lopes and L. A. Alexandre. An overview of blockchain integration with robotics and artificial intelligence. *arXiv preprint arXiv:1810.00329*, 2018.
- [71] N. Lukas, A. Salem, R. Sim, S. Tople, L. Wutschitz, and S. Zanella-Béguelin. Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 346–363. IEEE, 2023.
- [72] H. Luo, J. Luo, and A. V. Vasilakos. Bc4llm: Trusted artificial intelligence when blockchain meets large language models. *arXiv preprint arXiv:2310.06278*, 2023.
- [73] S. K. M., S. Nicolazzo, M. Arazzi, A. Nocera, R. R. K. A., V. P., and M. Conti. Privacy-preserving in blockchain-based federated learning systems. *Computer Communications*, 2024.
- [74] D. Malhotra, P. Saini, and A. K. Singh. Blockchain-based proof-of-authenticity frameworks for explainable ai. *Multimedia Tools and Applications*, 83(13):37889–37911, 2024.
- [75] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*, 2020.
- [76] J. G. M. Mboma, K. Lusala, M. Matalatala, O. T. Tshipata, P. S. Nzakuna, and D. T. Kazumba. Integrating llm with blockchain and ipfs to enhance academic diploma integrity. In *2024 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA)*, pages 1–6, 2024.
- [77] J. G. M. Mboma, O. T. Tshipata, W. V. Kambale, and K. Kyamakya. Assessing how large language models can be integrated with or used for blockchain technology: Overview and illustrative case study. In *2023 27th International Conference on Circuits, Systems, Communications and Computers (CSCC)*, pages 59–70. IEEE, 2023.
- [78] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas. Communication-efficient learning of deep networks from decentralized data. *Artificial Intelligence and Statistics*, pages 1273–1282, 2017.
- [79] R. C. Merkle. A digital signature based on a conventional encryption function. pages 369–378. Springer, 1988.
- [80] D. Mingxiao, M. Xiaofeng, Z. Zhe, W. Xiangwei, and C. Qijun. A review on consensus algorithm of blockchain. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2567–2572, 2017.
- [81] A. Nagar. Privacy-preserving blockchain based federated learning with differential data sharing. *arXiv preprint arXiv:1912.04859*, 2019.
- [82] S. Nakamoto. Bitcoin: A peer-to-peer electronic cash system. 2008.
- [83] S. Neel and P. Chang. Privacy issues in large language models: A survey. *arXiv preprint arXiv:2312.06717*, 2024.
- [84] C. T. Nguyen, D. T. Hoang, D. N. Nguyen, D. Niyato, H. T. Nguyen, and E. Dutkiewicz. Proof-of-stake consensus mechanisms for future blockchain networks: Fundamentals, applications and opportunities. *IEEE Access*, 2019.
- [85] A. Pal, L. K. Umapathi, and M. Sankarasubbu. Medhalt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*, 2023.
- [86] X. Pan, M. Zhang, S. Ji, and M. Yang. Privacy risks of general-purpose language models. pages 1314–1331, 2020.
- [87] Y.-H. Park, Y. Kim, and J. Shim. Blockchain-based privacy-preserving system for genomic data management using local differential privacy. *Electronics*, 10(23):3019, 2021.
- [88] A. Qammar, A. Karim, H. Ning, and J. Ding. Securing federated learning with blockchain: a systematic literature review. *Artificial Intelligence Review*, 56(5):3951–3985, 2023.
- [89] Y. Qu, M. P. Uddin, C. Gan, Y. Xiang, L. Gao, and J. Yearwood. Blockchain-enabled federated learning: A survey. *ACM Computing Surveys*, 55(4):1–35, 2023.
- [90] L. Reid. Generative ai in search: Let google do the searching for you, 2024.
- [91] K. Ruan, X. He, J. Wang, X. Zhou, H. Feng, and A. Kebarihotbi. S2e: Towards an end-to-end entity resolution from acoustic signal. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10441–10445, 2024.

- [92] P. Röttger, F. Pernisi, B. Vidgen, and D. Hovy. Safetyprompts: a systematic review of open datasets for evaluating and improving large language model safety. *arXiv preprint arXiv:2404.05399*, 2024.
- [93] S. Saberi, M. Kouhizadeh, J. Sarkis, and L. Shen. Blockchain technology and its relationships to sustainable supply chain management. *57(7):2117–2135*, 2019.
- [94] K. Salah, M. H. U. Rehman, N. Nizamuddin, and A. Al-Fuqaha. Blockchain for ai: Review and open research challenges. *IEEE Access*, 7:10127–10149, 2019.
- [95] O. Seneviratne. Blockchain for social good: Combating misinformation on the web with ai and blockchain. pages 435–442, 2022.
- [96] M. Shafay, R. W. Ahmad, K. Salah, I. Yaqoob, R. Jayaraman, and M. Omar. Blockchain for deep learning: review and open challenges. *Cluster Computing*, 26(1):197–221, 2023.
- [97] Z. Shah, I. Ullah, H. Li, A. Levula, and K. Khurshid. Blockchain based solutions to mitigate distributed denial of service (ddos) attacks in the internet of things (iot): A survey. 22, 2022.
- [98] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askill, S. R. Bowman, N. Cheng, E. Durmus, Z. Hatfield-Dodds, S. R. Johnston, S. Kravec, T. Maxwell, S. McCandlish, K. Ndousse, O. Rausch, N. Schiefer, D. Yan, M. Zhang, and E. Perez. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.
- [99] E. Shayegani, M. A. A. Mamun, Y. Fu, P. Zaree, Y. Dong, and N. Abu-Ghazaleh. Survey of vulnerabilities in large language models revealed by adversarial attacks. *arXiv preprint arXiv:2310.10844*, 2023.
- [100] M. Shen, H. Wang, B. Zhang, L. Zhu, K. Xu, Q. Li, and X. Du. Exploiting unintended property leakage in blockchain-assisted federated learning for intelligent edge computing. *IEEE Internet of Things Journal*, 8(4):2265–2275, 2021.
- [101] T. Shen, R. Jin, Y. Huang, C. Liu, W. Dong, Z. Guo, X. Wu, Y. Liu, and D. Xiong. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*, 2023.
- [102] S. Singh. Enhancing privacy and security in large-language models: A zero-knowledge proof approach. *International Conference on Cyber Warfare and Security*, 19(1):574–582, 2024.
- [103] S. Singh, P. K. Sharma, B. Yoon, M. Shojafar, G. H. Cho, and I.-H. Ra. Convergence of blockchain and artificial intelligence in IoT network for the sustainable smart city. *Sustainable Cities and Society*, 63, 2020.
- [104] H. Song, Z. Qu, and Y. Wei. Advancing blockchain scalability: An introduction to layer 1 and layer 2 solutions. *arXiv preprint arXiv:2406.13855*, 2024.
- [105] H. Song, Y. Wei, Z. Qu, and W. Wang. Unveiling decentralization: A comprehensive review of technologies, comparison, challenges in bitcoin, ethereum, and solana blockchain. *arXiv preprint arXiv:2404.04841*, 2024.
- [106] T. South, G. Zuskind, R. Mahari, and T. Hardjono. Secure community transformers: Private pooled data for llms. 2023.
- [107] L. Sun, Y. Huang, et al. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*, 2024.
- [108] X. Sun, F. R. Yu, P. Zhang, Z. Sun, W. Xie, and X. Peng. A survey on zero-knowledge proof in blockchain. *IEEE Network*, 35(4), 2021.
- [109] H. Taherdoost. Blockchain technology and artificial intelligence together: A critical review on applications. *Applied Sciences*, 12(24):12948, 2022.
- [110] S. Tedeschi, F. Friedrich, P. Schramowski, K. Kersting, R. Navigli, H. Nguyen, and B. Li. Alert: A comprehensive benchmark for assessing large language models’ safety through red teaming. *arXiv preprint arXiv:2404.08676*, 2024.
- [111] T. Teubner, C. M. Flath, C. Weinhardt, W. van der Aalst, and O. Hinz. Welcome to the era of chatgpt et al. *Business & Information Systems Engineering*, 65, 2023.
- [112] C. Tonkin. ‘ChatGPT, help me make a bomb’, 2023.
- [113] I. Ullah, N. Hassan, S. S. Gill, B. Suleiman, T. A. Ahanger, Z. Shah, J. Qadir, and S. S. Kanhere. Privacy preserving large language models: Chatgpt case study based vision and framework. *arXiv preprint arXiv:2310.12523*, 2023.
- [114] Y. Wan, G. Pu, J. Sun, A. Garimella, K.-W. Chang, and N. Peng. ”kelly is a warm person, joseph is a role model”: Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*, 2023.
- [115] B. Wang, W. Chen, H. Pei, C. Xie, M. Kang, C. Zhang, C. Xu, Z. Xiong, R. Dutta, R. Schaeffer, S. T. Truong, S. Arora, M. Mazeika, D. Hendrycks, Z. Lin, Y. Cheng, S. Koyejo, D. Song, and B. Li. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *NeurIPS*, 2023.
- [116] Q. Wang, Y. Guo, X. Wang, T. Ji, L. Yu, and P. Li. Ai at the edge: Blockchain-empowered secure multiparty learning with heterogeneous models. *IEEE Internet of Things Journal*, 7(10):9600–9610, 2020.
- [117] L. Weidinger, M. Rauh, N. Marchal, A. Manzini, L. A. Hendricks, J. Mateos-Garcia, S. Bergman, J. Kay, C. Griffin, B. Bariach, I. Gabriel, V. Rieser, and W. Isaac. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986*, 2023.
- [118] S. Wellington. Basedai: A decentralized p2p network for zero knowledge large language models (zk-llms). *arXiv preprint arXiv:2403.01008*, 2024.
- [119] C. Weng, K. Yang, X. Xie, J. Katz, and X. Wang. Mystique: Efficient conversions for {Zero-Knowledge} proofs with applications to machine learning. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 501–518, 2021.
- [120] J. Weng, J. WENG, M. Li, Y. Zhang, J. ZHANG, and W. LUO. Auditable privacy protection deep learning platform construction method based on block chain incentive mechanism, 2023. US Patent 11,836,616.
- [121] J. Weng, J. Weng, J. Zhang, M. Li, Y. Zhang, and W. Luo. Deepchain: Auditable and privacy-preserving

- deep learning with blockchain-based incentive. *IEEE Transactions on Dependable and Secure Computing*, 18(5):2438–2455, 2021.
- [122] K. Werder, B. Ramesh, and R. S. Zhang. Establishing data provenance for responsible artificial intelligence systems. *ACM Transactions on Management Information Systems*, (2):1–23, 2022.
- [123] A. Winograd. Loose-lipped large language models spill your secrets: The privacy implications of large language models notes. *Harvard Journal of Law & Technology (Harvard JOLT)*, (2), 2022.
- [124] L. Witt, A. T. Fortes, K. Toyoda, W. Samek, and D. Li. Blockchain and artificial intelligence: Synergies and conflicts. *arXiv preprint arXiv:2405.13462*, 2024.
- [125] C. F. Xavier. Pipemsm: Hardware acceleration for multi-scalar multiplication. *Cryptology ePrint Archive*, 2022.
- [126] J. Xue, M. Zheng, Y. Hu, F. Liu, X. Chen, and Q. Lou. Badrag: Identifying vulnerabilities in retrieval augmented generation of large language models. *arXiv preprint arXiv:2406.00083*, 2024.
- [127] B. Yan, K. Li, M. Xu, Y. Dong, Y. Zhang, Z. Ren, and X. Cheng. On protecting the data privacy of large language models (llms): A survey. *arXiv preprint arXiv:2403.05156*, 2024.
- [128] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, B. Yin, and X. Hu. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32, 2023.
- [129] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211, 2024.
- [130] A. Yazdinejad, R. M. Parizi, G. Srivastava, and A. Dehghantanha. Making sense of blockchain for ai deep-fakes technology. pages 1–6, 2020.
- [131] H. Ye, T. Liu, A. Zhang, W. Hua, and W. Jia. Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794*, 2023.
- [132] S. Zeng, J. Zhang, P. He, Y. Xing, Y. Liu, H. Xu, J. Ren, S. Wang, D. Yin, Y. Chang, et al. The good and the bad: Exploring privacy issues in retrieval-augmented generation (rag). *arXiv preprint arXiv:2402.16893*, 2024.
- [133] J. Zhang, T. Xie, Y. Zhang, and D. Song. Transparent polynomial delegation and its applications to zero knowledge proof. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 859–876. IEEE, 2020.
- [134] R. Zhang, H. Du, Y. Liu, D. Niyato, J. Kang, S. Sun, X. Shen, and H. V. Poor. Interactive ai with retrieval-augmented generation for next generation networking. *IEEE Network*, 2024.
- [135] R. Zhang, R. Xue, and L. Liu. Security and privacy on blockchain. 52(3), 2020.
- [136] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, L. Wang, A. T. Luu, W. Bi, F. Shi, and S. Shi. Siren’s song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.
- [137] Z. Zhang, L. Lei, L. Wu, R. Sun, Y. Huang, C. Long, X. Liu, X. Lei, J. Tang, and M. Huang. Safety-bench: Evaluating the safety of large language models with multiple choice questions. *arXiv preprint arXiv:2309.07045*, 2023.
- [138] Z. Zhang, Y. Rao, H. Xiao, X. Xiao, and Y. Yang. Proof of quality: A costless paradigm for trustless generative ai model inference on blockchains. *arXiv preprint arXiv:2405.17934*, 2024.
- [139] Z. Zhang, J. Yang, P. Ke, F. Mi, H. Wang, and M. Huang. Defending large language models against jailbreaking attacks through goal prioritization. *arXiv preprint arXiv:2311.09096*, 2024.
- [140] S. Zhao, J. Wen, A. Luu, J. Zhao, and J. Fu. Prompt as triggers for backdoor attack: Examining the vulnerability in language models. 2023.
- [141] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.
- [142] Y. Zhao, J. Zhao, L. Jiang, R. Tan, D. Niyato, Z. Li, L. Lyu, and Y. Liu. Privacy-preserving blockchain-based federated learning for iot devices. *IEEE Internet of Things Journal*, 8(3):1817–1829, 2021.
- [143] Y. Zhao, J. Zhao, J. Kang, Z. Zhang, D. Niyato, S. Shi, and K.-Y. Lam. A blockchain-based approach for saving and tracking differential-privacy cost. *IEEE Internet of Things Journal*, 8(11):8865–8882, 2021.
- [144] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.
- [145] Z. Zheng, K. Ning, Y. Wang, J. Zhang, D. Zheng, M. Ye, and J. Chen. A survey of large language models for code: Evolution, benchmarking, and future trends. *arXiv preprint arXiv:2311.10372*, 2023.
- [146] Z. Zheng, S. Xie, H.-N. Dai, W. Chen, X. Chen, J. Weng, and M. Imran. An overview on smart contracts: Challenges, advances and platforms. *Future Generation Computer Systems*, 105:475–491, 2020.
- [147] X. Zhu, H. Li, and Y. Yu. Blockchain-based privacy preserving deep learning. pages 370–383, 2019.
- [148] J. Zhuang and C. Kennington. Understanding survey paper taxonomy about large language models via graph representation learning. *arXiv preprint arXiv:2402.10409*, 2024.
- [149] W. Zou, R. Geng, B. Wang, and J. Jia. Poisonedrag: Knowledge poisoning attacks to retrieval-augmented generation of large language models. *arXiv preprint arXiv:2402.07867*, 2024.
- [150] X. Zuo, M. Wang, T. Zhu, L. Zhang, D. Ye, S. Yu, and W. Zhou. Federated trustchain: Blockchain-enhanced llm training and unlearning. *arXiv preprint arXiv:2406.04076*, 2024.
- [151] X. Zuo, M. Wang, T. Zhu, L. Zhang, S. Yu, and W. Zhou. Federated learning with blockchain-enhanced machine unlearning: A trustworthy approach. *arXiv preprint arXiv:2405.20776*, 2024.