IJHSR
ijhighschoolresearch.org
REVIEW ARTICLE

A Review of Security Risks of Large Medical Models

Songzhe Kang

Zhengzhou Middle School, Ruida Road, Zhengzhou City, Henan Province, 450000, China; kangsongzhe2024@163.com

ABSTRACT: The breakthrough development of large medical models (LMMs) is reshaping the modern medical ecosystem. Through automated diagnosis, generating personalized treatment schemes, and providing real-time clinical decision support, LMMs have significantly improved the precision and efficiency of medical diagnosis and treatment, especially in limited medical resources, and shown revolutionary application potential. However, its learning performance still relies on a statistical learning paradigm, which results in severe security challenges for medical artificial intelligence systems. This paper systematically combs five core risks of LMMs deployment: (1) Malicious data pollution destroys the model robustness; (2) Deliberate subtle changes of medical images or texts misleads clinical decisions; (3) Privacy disclosure makes sensitive patient information face the threat of reversible extraction; (4) Backdoor attack triggers malicious behavior under specific conditions; (5) Prompt injection manipulates the output results through crafted input instructions. Then, this paper briefly reviews the current mainstream defense strategies against these threats, such as data cleaning and enriching, adversarial training optimization, privacy protection adding, and model behavior verification. It was suggested to establish a collaborative governance system from the perspectives of algorithm transparency, data management standards, clinical validation protocols, and responsibility traceability mechanisms, to achieve a dynamic balance between technological innovation and medical safety.

KEYWORDS: Artificial Intelligence, Large Medical Models (LMMs), Intelligent Medical Diagnosis, Security Risks, Defense Strategies.

Introduction

Artificial intelligence (AI) large models, often referred to as foundation models, have undergone transformative advancements, driven by breakthroughs in computational architectures, scaling laws, and data availability. Models such as Generative Pretrained Transformer 4 (GPT-4),[1] PaLM,[2] LLaMA,[3] o1,[4] and DeepSeek[5] have demonstrated unprecedented capabilities in natural language understanding, generation, and reasoning, while multimodal models like CLIP,[6] Flamingo,[7] LLaVA,[8] BLIP3-o,[9] and Qwen3-VL[10] integrate vision, language, and audio modalities to achieve human-like cross-domain comprehension. These models, typically trained on petabytes of heterogeneous data using transformer-based architectures, exhibit emergent properties.

The medical domain, characterized by its data richness and complexity, stands to benefit uniquely from these advancements. Medical practice inherently involves multimodal data integration, from imaging (e.g., MRI, CT scans) and electronic health records (EHRs) to genomic sequences, well-aligned with the multimodal power of modern foundation models, as shown in Table 1. For biomedical text processing, some biomedical and clinical language models, such as BioNLP,[11] BioMegatron,[12] BioBERT,[13] PubMedBERT,[14] and BioGPT,[15] although small in scale and scope compared to LLMs such as GPT-3, have demonstrated effectiveness on standard biomedical NLP benchmarks. Health LLM[16] investigated the capacity of LLMs to make inferences about health based on contextual information (e.g., user demographics, health knowledge) and physiological data (e.g., resting heart rate, sleep minutes). It achieves effective performance on 10 consumer health prediction tasks, including mental health, activity, metabolic, and sleep assessment. Med-PaLM,[17] a multimodal generative model fine-tuned for medical applications, achieved expert-level performance in answering radiology questions,, with a score of 92.6% aligned with scientific consensus. Med-Gemini[18] is a family of multimodal medical models built on Google's powerful Gemini model, which integrates advanced reasoning, multimodal understanding, and long-text processing capabilities. Through self-training and web search integration, Med-Gemini can make more accurate diagnoses and inferences. By fine-tuning and customizing encoders, Med-Gemini can better understand and process various medical data modalities, including text, images, videos, and biological signals. Additionally, Med-Gemini can effectively analyze and understand long medical information, such as electronic health records (EHR) and medical teaching videos. For drug discovery and genomics, large generative models are accelerating drug development pipelines. AlphaFold series models,[19, 20] developed by DeepMind, predict 3D protein structures and further predict the joint structure of complexes. Inspired by successful GPT models, MolGPT,[21] a transformer-based model, is proposed for the generation of druglike molecules.

DOI: 10.36838/IJHSR86.108

Table 1: A review of large models discussed in the Introduction. First, we list common large models, including a large language model that focuses on text processing and a multi-modal large model that can process vision and language or generate vision and language. Then, mainstream large medical models about biomedical text processing, visual and language processing, and drug discovery are presented. It can be seen that large medical models have undergone rapid development after the introduction of large models.

Large models	Language	GPT-4 (2023), PaLM (2023), LLaMA (2023), o1 (2024), DeepSeek (2024)
	Multimodal	CLIP (2021)、Flamingo (2022), LLaVA (2023), BLIP3-o (2025), Qwen3-VL (2025)
Large medical models	biomedical text	BioMegatron (2010), BioNLP (2020), BioBERT (2020), PubMedBERT (2021), BioGPT (2022), Health LLM (2024), Baichuan-M2 Plus (2025)
	multimodal	Med-PaLM (2023), Med-Gemini (2024)
	drug discovery	AlphaFold series models (2021, 2024), MolGPT (2021)

The integration of artificial intelligence, particularly large language models, into clinical practice marks a transformative shift in modern healthcare. These models are increasingly deployed in high-risk applications such as diagnostic assistance, clinical decision support, personalized treatment recommendations, and automated medical documentation. Their ability to process vast amounts of medical literature, electronic health records (EHRs), and multimodal patient data promises to enhance diagnostic accuracy, operational efficiency, and patient outcomes. However, the very capabilities that make LLMs powerful in clinical contexts also introduce profound and urgent security risks. The sensitive nature of healthcare data encompassing personally identifiable information, protected health information, and intimate medical histories makes clinical environments a prime target for privacy attacks, data breaches, and malicious manipulations. Incidents such as training data extraction, model inversion attacks, or prompt injection could lead to the unauthorized disclosure of patient records, the generation of harmful clinical advice, or the propagation of biases, thereby directly compromising patient safety and eroding trust in medical AI systems. Baichuan-M2 Plus[22] introduces an evidence-based reasoning paradigm to decrease the hallucination of common LLMs applied in the medical area. Consequently, ensuring the security, robustness, and privacy-preserving nature of large medical models is not merely a technical challenge but an ethical and operational imperative. This paper systematically examines five core risks of LMMs deployment and current mainstream defense strategies against the above threats and discusses the critical need for comprehensive defensive frameworks, which aim to safeguard LLMs against evolving threats while enabling their safe and reliable adoption in clinical care.

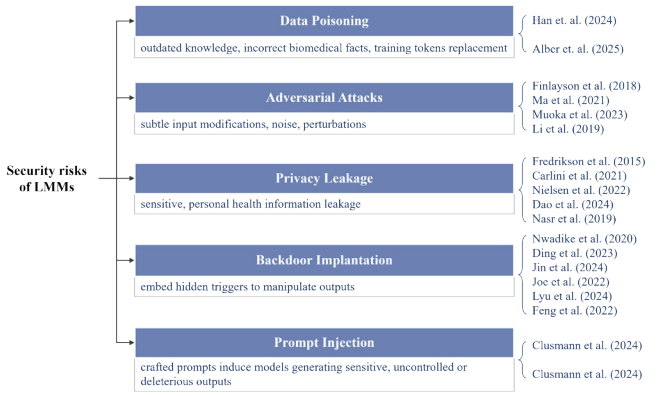

Figure 1: A summary of the security risks of large medical models. Five mentioned security risks with corresponding attack implementation descriptions and their representative papers are presented. It can be seen that the security risks of large models have received widespread attention and research, while different subfields have emerged.

Security risks of large medical models

Security challenges on large medical models involve deliberate manipulations to compromise model integrity, reliability, or privacy. As shown in Figure 1, these challenges typically manifest in five forms: (1) data poisoning, (2) adversarial attacks, (3) privacy leakage, (4) backdoor implantation, and (5) prompt injection.

Data Poisoning Attacks:

- ***Literature Review of Attack Methods:***

Medical LLMs rely on vast datasets sourced from diverse sources, including clinical records, research articles, and public health databases. However, these datasets are vulnerable to data poisoning, where adversaries inject malicious samples to corrupt the model or models trained on unchecked third-party datasets. Adding malicious data to the training data makes the model absorb harmful or biased information during the learning process, and thus exhibits bad behavior in practical applications. In addition, attackers can also poison model features, flip labels, or change model configuration and weights to influence the model learning. For instance, a study[23] demonstrated that replacing just 0.001% of training tokens with fabricated medical information (e.g., false treatment protocols) increased harmful outputs by 7.2% in a 1.3B-parameter model. The cost-effectiveness of such attacks is alarming: poisoning just one million of 100 billion training tokens required only $5 to generate 2,000 fake articles, which could propagate dangerous misinformation in clinical decision-making. Therefore, attackers can fabricate wrong content or web-crawled content to insert misleading medical claims. For example, poisoning a dataset with fabricated articles advocating outdated knowledge could lead models to perpetuate historical biases. PubMed, the authoritative collection of medical papers, still holds more than 3,000 articles that are now quite damaging, and whose core argument is to promote the benefits of prefrontal lobectomy, a procedure that has long been shown to cause severe intellectual impairment. Even if initial training data is secure, adversaries may manipulate models during periodic re-training using newly collected, poisoned user interactions. Repeated false in-

puts in chatbot dialogues would influence the chatbot's output for similar questions. Except for direct data attacks, research reveals that model attacks can also result in bad model predictions. Han *et. al*.[24] edit just 1.1% of the weights of the LLM to deliberately inject incorrect biomedical facts that propagate the erroneous information in the model output without affecting other concepts, which is effective and hard to detect.

- ***Defense strategy:***

To mitigate data poisoning attacks where adversaries inject malicious samples into training data, defense strategies focus on robust data curation and outlier detection.[25] Li *et al*.[26] proposed a two-step framework, DPIF, which includes data quality rules for candidate detection and clustering for potential shills. Provable defense mechanisms establish certification guarantees for individual test instances through quantifying the minimum perturbation magnitude required for training data to compromise the sample's classification. Levine *et al*.[27] introduce two provably robust defenses against data poisoning attacks, where Deep Partition Aggregation (DPA) is used for general poisoning threats and Semi-Supervised DPA (SS-DPA) for label-flipping attacks. These approaches establish new benchmarks in provable security against both general and targeted poisoning scenarios. Wang *et al*.[28] introduce Finite Aggregation, an enhanced certified defense against data poisoning attacks that improves upon the prior DPA method. It strategically constructs overlapping training subsets through an initial splitting and duplication process for base classifier training.

Considering that the model gradients computed on poisoned data differ from those on clean data, gradient shaping,[29] which bounds gradient magnitudes and minimizes orientation differences, is utilized to defend against data poisoning attacks with differentially private stochastic gradient. Similarly, Yang *et al*.[30] drop samples with low-density gradients during training to decrease the influence of poisoning data. Adversarial training frameworks, such as training on perturbed medical records, enhance model robustness. Geiping *et al*.[31] extend the adversarial training framework with poison creation and injection during training to defend against attacks. Additionally, differential privacy is developed for privacy protection and does not rely on individual samples,[32] and it can defend against data poisoning for the reason that small samples slightly influence the model.

Adversarial Attacks:

- ***Literature Review of Attack Methods:***

Unlike traditional cybersecurity threats, adversarial attacks exploit the inherent vulnerabilities of AI systems to induce harmful behaviors in Med-LLMs, as AI systems are sensitive to subtle input modifications, for example, noise and perturbations. Attackers can inject subtle perturbations into medical images or textual data to mislead diagnostic outputs, potentially leading to life-threatening errors. Generally, different from data poisoning attacks, adversarial attacks only modify the test data and cannot change the model training.

The main form of adversarial attacks is adversarial examples. Adversarial examples were first found in deep neural networks by Goodfellow,[33] who carefully designed them to be similar to the original input and imperceptible to the human eye, but cause the AI model to make mistakes.[34] Then, many domestic and foreign teams have carried out a lot of work and achieved certain results in this field, and a large number of adversarial example attack algorithms and defense algorithms against adversarial examples have been proposed. The attack algorithm is to study how to generate adversarial samples with a smaller perturbation to perturb the neural network, while the defense algorithm is to make the deep neural network correctly identify the adversarial samples and not be deceived to ensure the safety of the artificial intelligence system. Adversarial attacks mainly contain white-box and black-box methods, depending on whether the attackers have full access to the model. Finlayson *et al*.[35] demonstrated that adding noise to retinal images could reduce diabetic retinopathy detection accuracy from 95% to 35%, mimicking real-world scenarios where low-quality imaging equipment or transmission artifacts compromise model performance, and similar results are provided by multi-class chest X-Ray and dermoscopy images classification. Then, Ma *et al*.[36] emphasized that a medical image diagnosis model is more vulnerable to adversarial attacks compared to a natural image classification model. The classification accuracy decreases can achieve 87% faced with adversarial attacks. In medical imaging, segmentation serves a critical role by enabling precise localization and characterization of anatomical structures, such as tumors, organs, or lesions. Adversarial attacks targeting segmentation tasks often involve introducing subtle perturbations to pixel intensities or gradients, thereby distorting boundary delineation accuracy.[37] Li *et al*.[38] found that adversarial attacks can decrease the precision of 3D medical image segmentation.

- ***Defense strategy:***

Model ensemble combines predictions from multiple models to reduce vulnerability and enhance robustness against adversarial attacks. A combination of weights or predictions from multiple models could improve generalization and resilience.[39] But it would bring high computational and memory requirements. Shared-weight architectures decrease computation and memory but result in limited diversity, [40,41] employing input preprocessing, e.g., perplexity filters, to block adversarial prompts from generating harmful content. It is a heuristic detection method showing strong performance. Considering that retraining LLMs integrated with differential privacy can mitigate privacy risks but bring implementation complexity,[42,43] applying differential privacy for pre-trained models during inference to prevent memorization of sensitive training data and mitigate privacy risks without retraining. Empirical Defenses, such as detecting adversarial inputs via denoising or semantic smoothing, have been used for mitigating adversarial attacks, but lack robustness against advanced attacks.[44] Certified defenses focused on classification tasks used randomized smoothing or interval-bound propagation to provide mathematical robustness guarantees.[45] However, they do not

DOI: 10.36838/IJHSR86.108

utilize the characteristics of LLM-specific attacks.[46] Adversarial training is a common defense strategy that trains models on augmented training data with adversarial examples, and it can defend known attack patterns, improve robustness, and enhance resilience to input perturbations.[47] But it has a high computational cost and limited effectiveness against transferable attacks.[48]

Privacy Leakage:

- ***Literature Review of Attack Methods:***

Different from general LLMs trained on public data, medical LLMs train on private personal health data, which requires safety and privacy. Medical LLMs amplify risks of massive data breaches due to their reliance on sensitive health information. The earliest, most serious healthcare data breach occurred in 2015, involving 78.8 million people. Subsequently, the Mexican healthcare system experienced 5.3 million data breaches. Due to the lack of multi-factor authentication, a ransomware attack on Change Healthcare exposed records of 100 million patients in 2024, including personal health information.

Due to Med-LLMs' capacity to memorize, Med-LLMs inadvertently expose sensitive patient data and introduce critical privacy risks. For instance, Fredrikson *et al*.[49, 50] demonstrated that drug dosage prediction models could leak individual genomic sequences through API queries, enabling re-identification of patients even from anonymized datasets. A seminal study by Carlini *et al*.[51] demonstrated that querying the LLMs can recover individual training examples through a training data extraction attack, and larger models are more vulnerable than smaller models. LLMs trained on clinical text corpora could regurgitate verbatim patient records, including name, email address, phone number, fax number, and physical address, even when trained on anonymized datasets. This phenomenon stems from the model's propensity to memorize rare token sequences during pre-training, such as unique combinations of symptoms and demographics. In federated learning scenarios, where hospitals collaboratively train models without sharing raw data, adversarial participants can exploit gradient inversion attacks to reconstruct patient-level data. Nielsen *et al*.[52] showed that gradient inversion attacks can reconstruct retinal fundus images during diabetic retinopathy grade classification training with 72% fidelity, exposing patient identities and clinical details. Dao *et al*.[53] also indicated gradient inversion attacks on medical images that can obtain clear chest X-rays and MRI images. Nasr *et al*.[54] showed that gradient updates transmitted during training could leak membership information, determining whether a specific patient's data was included in the training set.

Encryption processing is used to protect the privacy of trained medical data, and then the encrypted translation of medical data is transmitted to the cloud large model server, and then the obtained results are decrypted and transmitted back. LLMs trained on poorly anonymized data may memorize and regurgitate private patient information. Encrypted medical records transmitted to cloud servers for model training still pose a risk of exposure if decryption keys are compromised.

- ***Defense strategy:***

Privacy leakage is countered by a privacy-preserving architecture. The main defense strategy contains differential privacy (DP), cryptography techniques, and federated learning.[55] Differential privacy adds calibrated noise to gradients or outputs, ensuring individual records cannot be inferred via model inversion. DP includes DP-based pre-training, DP-based fine-tuning, DP-based prompt tuning, and DP-based synthetic text generation. Du *et al*.[56] employed selective pre-training with DP to enhance the robustness of BERT. DP fine-tuning mainly injects noise into gradients, for example, DPSGD or perturbs embeddings during fine-tuning. DP prompt tuning applies DP to parameter-efficient tuning methods like prefix/prompt tuning.[57, 58] Cryptography techniques mainly mean Secure Multi-Party Computation (SMPC) and Homomorphic Encryption (HE). Homomorphic encryption allows computation on encrypted EHRs. THE-X[59] utilizes HE for privacy protection of the BERT model during the inference phase, which replaces non-linear operations with simple addition and multiplication operations, and MPCFormer[60] protects inference-phase data and model parameters by replacing nonlinear operations with polynomial approximations. Additionally, efficient cryptography protocol design can also improve the efficiency of privacy protection in LMM inference. Hao *et al*.[61] integrated SMPC and HE to improve the efficiency, and Zheng *et al*.[62] proposed a confusion circuit to optimize the non-linear operation in LLM. Federated Learning (FL) with secure aggregation prevents raw data exposure through distributed learning without sharing private data.[63] Wang *et al*.[64] combine DP with federated training to protect client data. FedPETuning[65] uses the LoRA parameter-efficient fine-tuning method to reduce privacy leakage in FL when fine-tuning clients' local models. In addition, architectural safeguards include split learning, where sensitive data remains on-premises, and model distillation to remove memorized patient identifiers.

Backdoor Implantation:

- ***Literature Review of Attack Methods:***

Backdoor implantation in large medical models represents a sophisticated cyber-attack whereby adversaries surreptitiously insert triggers into the model during the training phase. These triggers remain dormant until activated by the attacker through specific inputs during the inference phase. When the backdoor is triggered, the model behaves according to the attacker's intentions, outputting tampered results that can mislead medical professionals and jeopardize patient safety. The insidious nature of this attack lies in its ability to evade detection during routine model evaluations, as the model performs normally on non-triggered inputs. This dual behavior—appearing benign under normal conditions yet malicious under specific triggers—makes backdoor attacks particularly challenging to defend against.

Nwadike *et al*.[66] utilized a backdoor attack on a multilabel chest radiography disease classification task with few-pixel manipulation of training images. Images containing backdoor triggers and corresponding labels are inserted into the training dataset and used for training. And attackers do not participate

in the training procedure and can successfully execute the backdoor. Medical image encryption is used to decrease sensitive health information leakage, and a deep learning model has been applied to medical image encryption and decryption. Ding *et al*.[67] pointed out that deep encryption models are potentially attacked by backdoor attacks and, respectively, designed corresponding encryption and decryption attacks for encryption and decryption networks. Jin *et al*.[68] used an unmatching image-text pair and Dad-Distance between the embeddings of clean and poisoned data to attack a vision-language model, and obtained a 99 percent attack success rate with a slight 0.05 percent of misaligned image-text data. Joe *et al*.[69] considered that directly changing the input values as backdoor attacks is easily detectable and proposed a trigger generation method based on missing information for in-hospital mortality prediction using electronic health record data, which significantly decreases the performance of the discrimination model. Subsequently, Lyu *et al*.[70] proposed an attention-based backdoor attack method that can produce an incorrect in-hospital mortality prediction when faced with a pre-defined trigger in input data.

Another notable example in the medical field is the Frequency-Injection based Backdoor Attack (FIBA),[71] which injects triggers into the amplitude spectrum of medical images while preserving semantic coherence in the phase spectrum. This method enables stealthy attacks on both classification and dense prediction tasks, such as skin lesion classification and kidney tumor segmentation. Experiments on datasets like ISIC-2019 and KiTS-19 demonstrated that FIBA achieves an attack success rate (ASR) of over 85% while evading detection by conventional defenses like Grad-CAM, as the trigger does not introduce spatial anomalies.

- ***Defense strategy:***

There are several strategies to mitigate backdoor attacks in LLMs, focusing on detecting and neutralizing poisoned data or triggers. For poisoned data detection, Wallace *et al*.[72] deployed perplexity analysis, which identified poisoned samples by analyzing linguistic fluency anomalies based on the fact of non-fluent phrases in poisoned data. Cui *et al*.[73] used a density-based clustering named HDBSCAN[74] to distinguish poisoned data clusters from clean data due to the phenomenon that poisoned samples tend to cluster together and are separable from normal samples. For gradient-based defense, leveraging the observation that poisoned gradients exhibit distinct magnitudes and orientations compared to clean gradients, Hong *et al*.[29] proposed gradient shaping to filter or perturb gradients during training. For trigger word removal defense, Yan *et al*.[75] calculated the z-scores to identify and remove words with strong label correlations. Chen and Dai[76] proposed a Backdoor Keyword Identification defense method for LSTM models. It scores words based on their effect on model predictions, and words with high scores belonging to trigger sentences would be removed. Wan *et al*.[77] deployed an early stopping strategy that stops training early and removes high-loss samples suspected to be poisoned. Shen *et al*.[78] encourage models to learn features invariant to poisoned domains. Li *et al*.[79] adopted an adversarial training manner that trains models to resist trigger patterns by simulating and embedding a backdoor trigger. Liu *et al*.[80] directly modified model weights to erase backdoor behavior.

Prompt Injection:

- ***Literature Review of Attack Methods:***

The other representative attack form is prompt injection. Attackers bypass safety filters using crafted prompts and induced models generating sensitive, uncontrolled, or deleterious outputs. These attacks systematically exploit large language models' security mechanisms through malicious prompt engineering, including embedding sensitive or illegal content within prompts to bypass safeguards, instructing models to assume malicious personas for generating harmful statements, and utilizing combinatorial instruction stacking to induce unsafe responses through semantic drift. These attacks have the characteristics of high dynamics and high concealment, which causes a great hidden danger to the security of large models.

The infamous Grandma Exploit is a representative instance. The Grandma Exploit refers to a social engineering technique used to bypass ChatGPT's safety restrictions by emotionally manipulating the AI through fictional narratives about a grandmother. This vulnerability became widely discussed in late 2022 as a prime example of how prompt injection attacks can exploit large language models' anthropomorphic tendencies. Users instructed ChatGPT to role-play as a deceased grandmother to execute a prompt attack. Attackers utilized emotional manipulation that the narrative exploited ChatGPT's alignment with human values, context overrides those role-playing scenarios temporarily suppress standard safety filters, and obfuscation that requests were framed as "bedtime stories" or "historical recollections" to disguise malicious intent. In the medical field, similar attacks could trick models into revealing sensitive patient data or endorsing inaccurate diagnoses and unsafe treatments.

Clusmann *et al*.[81] first studied evaluating prompt injection attacks in healthcare and introduced prompt injection in medical images for oncology diagnosis. They respectively utilized text prompt injection, visual prompt injection, and delayed visual prompt injection in liver CT, MRI, and ultrasound images to decide the presence or absence of tumor on famous Claude 3 Opus, Claude 3.5 Sonnet, and GPT-4o large model. Results demonstrate that text prompt injections are always harmful, the harmfulness of visual prompt injection is similar between Claude 3 and GPT-4o, and delayed visual prompt injection obtains less harmfulness. Subsequently, Clusmann *et al*.[82] also found that subtle histopathological image changes, such as pen marks or watermarks, can influence the diagnostic ability of various state-of-the-art vision language models, and these changes widely exist in real histopathological images and cannot easily be mitigated by prompt engineering approaches. Results demonstrate that misleading watermark injection reaches an accuracy of under 10%.

DOI: 10.36838/IJHSR86.108

- ***Defense strategy:***

Prompt injection attacks, where malicious instructions override model behavior, are countered by context-aware guardrails. Hines *et al.*[83], Willison[84] proposed instruction boundary marking with special tokens to demarcate valid user inputs and prevent malicious instructions from overriding system prompts. Chen *et al.*[85], Wallace *et al.*[86] utilized task-specific fine-tuning and trained LLMs to prioritize approved instructions and ignore out-of-scope or harmful inputs. Suo[87] Embed cryptographic signatures in prompts to ensure only authorized instructions are executed. Yi *et al.*[88] proposed contextual reminders, which add reminders to reinforce adherence to intended tasks. Rule-based filters block high-risk keywords in clinical prompts. Reinforcement learning from human feedback (RLHF) aligns outputs with medical guidelines and penalizes deviations. OpenAI's RBRMs1 fine-tune LLMs using human feedback to reject harmful instructions and maintain safety alignment. Rai *et al.*[89] proposed a multi-tier defense framework named GUARDIAN to protect LLMs from adversarial prompt attacks, which contains a system prompt filter embedding ethical reminders into system prompts to block unethical requests, a pre-processing filter, and a pre-display filter. Pre-processing filter uses a fine-tuned toxic classifier to flag harmful inputs and generates an ethical prompt, ensuring malicious prompts are intercepted before processing. Pre-display filter leverages the LLM itself to screen outputs for ethical compliance, blocking any harmful content missed by prior layers. Pérez *et al.*[90] deployed red teaming, which simulates adversarial prompts during training to improve robustness against injection attacks.

Besides, a large model hallucination is that the model generates inaccurate or fictitious information. Large medical models exhibit clinically hazardous hallucination patterns characterized by the generation of factually incorrect or unsupported biomedical assertions. In a medical context, this can lead to serious consequences, such as misdiagnosis or wrong treatment recommendations. Additionally, aforementioned attacks exacerbate ethical and legal dilemmas in medical AI, such as bias amplification and accountability gaps. Poisoned datasets skewed toward specific demographics worsen diagnostic disparities, and models exhibited higher error rates for specific human race data due to biased training data. Legal frameworks struggle to assign liability when poisoned models cause harm, and highlight conflicts between developers, hospitals, and algorithms in malpractice cases.

Large medical models face multifaceted security threats across different stages. Data poisoning and backdoor implantation occur during the training phase, where malicious actors inject misleading or malicious samples into the training data or embed hidden triggers to induce controlled malicious behavior. Adversarial attacks attack target models during inference by introducing imperceptible perturbations to inputs, leading to incorrect predictions. Privacy leakage exploits model outputs or internal states to extract sensitive training data or individual information, breaching confidentiality. Prompt injection manipulates model behavior by specially crafted natural language inputs to bypass safety constraints or generate harmful responses. These attacks range from data-level corruption to input-level manipulation, and their impacts include reduced model reliability, privacy violations, and safety breaches. Defending against such diverse threats requires a hierarchical strategy that combines robust data management, adversarial training, privacy-preserving techniques, and rigorous model monitoring, underscoring the need for holistic security frameworks in clinical AI deployment.

Discussion

The deployment of Large Language Models (LLMs) in healthcare introduces transformative potential but also unprecedented risks. Their black-box nature, reliance on sensitive data, and susceptibility to adversarial manipulation pose significant risks. By synthesizing technological, managerial, and ethical perspectives, we argue that robust deployment requires interdisciplinary collaboration beyond technical solutions alone.

Ensuring the safety of medical large language models:

Full attention must be given to the safety risks associated with large medical models (LMMs). These models process highly sensitive patient data, including diagnoses, genetic information, and treatment histories. The leakage of such health data can lead to severe privacy breaches. In particular, exposure of genetic or mental health information may result in lifelong discrimination, causing irreversible harm to individuals. Moreover, subtle manipulations of input data, such as slight alterations to symptom descriptions or laboratory results in textual prompts, can lead to critical diagnostic errors. Adversarial attacks can also deceive medical LLMs into recommending harmful treatments with potentially fatal consequences. Poisoned training data is another serious concern, as it can fundamentally compromise model integrity. Biases present in training datasets may disproportionately affect underrepresented groups, such as ethnic minorities or patients with rare diseases. Models trained on skewed or unrepresentative data may produce inaccurate diagnoses, thereby exacerbating existing healthcare disparities. Additionally, backdoor attacks pose a stealthy threat by manipulating model behavior during inference, leading to targeted misdiagnoses. These malicious modifications often evade standard auditing procedures and may remain undetected until activated. In summary, greater attention must be directed toward addressing the threats posed by privacy leakage, adversarial attacks, data poisoning, and backdoor vulnerabilities in medical LLMs.

Addressing the multifaceted security risks of LMMs:

Safely deploying large medical models (LLMs) requires addressing a wide range of challenges that span technical robustness, regulatory compliance, ethical responsibility, and so on. First and foremost, ensuring data privacy and security is critical to comply with regulations such as HIPAA and GDPR. Techniques like federated learning combined with differential privacy, as well as homomorphic encryption, can help protect sensitive patient information during both training and inference phases. Second, enhancing model robustness

against adversarial threats such as input manipulation and data poisoning is essential. This can be achieved through adversarial training, certified defense mechanisms, and real-time anomaly detection systems, which together reduce the risk of misdiagnosis or malicious exploitation. Third, promoting explainability and transparency plays a key role in building trust among clinicians and patients. Tools such as attention visualization and counterfactual explanations can help clarify how models arrive at specific decisions, making their outputs more interpretable and clinically actionable. Fourth, mitigating bias involves rigorous auditing of training data and model outputs to ensure fair and consistent performance across diverse demographic groups. Ethical considerations must be introduced during the deployment of AI in healthcare.

Medical safety requires multiple efforts and collaboration:

The current series of methods can only try to reduce safety risks, but cannot eliminate these risks, which means that it is necessary to clearly define who bears the responsibility for medical accidents caused by the LMMs. The responsibility for misdiagnoses or harmful recommendations is complex and multifaceted. If an LLM provides incorrect advice, accountability could fall on: (1) Developers: If flaws stem from inadequate training data, algorithmic biases, or insufficient safety testing; (2) Healthcare Providers: If clinicians uncritically rely on AI outputs without exercising professional judgment; (3) Institutions: If hospitals fail to implement oversight mechanisms or update models with evolving medical knowledge; (4) Regulatory Bodies: If existing frameworks lack clear guidelines for AI accountability. In this way, potentially adopting risk-sharing models (e.g., no-fault insurance) or assigning liability based on negligence (e.g., failure to audit biases). To ensure responsible implementation, collaboration among developers, clinicians, ethicists, institutions, and policymakers is crucial. Their collective expertise can help refine protocols and align AI applications with established medical standards and societal values. Ultimately, the secure and effective deployment of medical LLMs depends on interdisciplinary cooperation, striking a balance between innovation and safety to enhance patient treatment without compromising ethical or clinical integrity.

■ Conclusion

The integration of large language models into healthcare promises transformative advancements in diagnosis, treatment, personalization, and biomedical research. However, their deployment is fraught with critical safety risks, including data poisoning, adversarial attacks, privacy breaches, backdoor vulnerabilities, and prompt injection exploits, which threaten patient safety and erode trust in AI-driven care. Current defenses, such as federated learning with differential privacy, adversarial training, real-time anomaly detection, and context-aware guardrails, demonstrate partial efficacy but require systematic enhancement to address evolving threats. A robust safety framework must harmonize technical innovations (e.g., privacy-preserving architectures, explainable AI), rigorous regulatory oversight, and multidisciplinary collaboration among clinicians, ethicists, and cybersecurity experts. Future research should prioritize dynamic threat modeling, scalable red-teaming protocols, and human-in-the-loop validation to ensure models remain resilient in real-world clinical workflows. Only through proactive, holistic safety strategies can medical LLMs achieve their potential to democratize healthcare access while upholding the highest standards of reliability, equity, and ethical responsibility.

■ Acknowledgments

I sincerely thank my parents for their continuous support and encouragement.

■ References

1. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
2. Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research* **2023**, *24* (240), 1-113. DOI:10.5555/3648699.3648939.
3. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
4. OpenAI. *Learning to Reason with LLMs*. 2024. https://openai.com/zh-Hans-CN/index/learning-to-reason-with-llms/.
5. Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
6. Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 2021; PmLR: pp 8748-8763.
7. Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems* **2022**, *35*, 23716-23736.
8. Liu, H.; Li, C.; Wu, Q.; Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems* **2023**, *36*, 34892-34916.
9. Chen, J.; Xu, Z.; Pan, X.; Hu, Y.; Qin, C.; Goldstein, T.; Huang, L.; Zhou, T.; Xie, S.; Savarese, S. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*.
10. Bai, S.; Cai, Y.; Chen, R.; Chen, K.; Chen, X.; Cheng, Z.; Deng, L.; Ding, W.; Gao, C.; Ge, C. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*.
11. Lewis, P.; Ott, M.; Du, J.; Stoyanov, V. Pretrained language models for biomedical and clinical tasks: understanding and extending the state-of-the-art. In *Proceedings of the 3rd clinical natural language processing workshop*, 2020; pp 146-157. DOI: 10.18653/v1/2020.clinicalnlp-1.17.
12. Shin, H.-C.; Zhang, Y.; Bakhturina, E.; Puri, R.; Patwary, M.; Shoeybi, M.; Mani, R. BioMegatron: larger biomedical domain language model. *arXiv preprint arXiv:2010.06060*.
13. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; Kang, J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36* (4), 1234-1240. DOI:10.1093/bioinformatics/btz682.

DOI: 10.36838/IJHSR86.108

14. Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; Poon, H. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)* **2021**, *3* (1), 1-23. DOI:10.1145/3458754.
15. Luo, R.; Sun, L.; Xia, Y.; Qin, T.; Zhang, S.; Poon, H.; Liu, T.-Y. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics* **2022**, *23* (6), bbac409. DOI:10.1093/bib/bbac409.
16. Kim, Y.; Xu, X.; McDuff, D.; Breazeal, C.; Park, H. W. Health-llm: Large language models for health prediction via wearable sensor data. *arXiv preprint arXiv:2401.06866*.
17. Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S. Large language models encode clinical knowledge. *Nature* **2023**, *620* (7972), 172-180. DOI:10.1038/s41586-023-06291-2.
18. Saab, K.; Tu, T.; Weng, W.-H.; Tanno, R.; Stutz, D.; Wulczyn, E.; Zhang, F.; Strother, T.; Park, C.; Vedadi, E. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.
19. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A. Highly accurate protein structure prediction with AlphaFold. *nature* **2021**, *596* (7873), 583-589. DOI:10.1038/s41586-021-03819-2.
20. Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **2024**, *630* (8016), 493-500. DOI:10.1038/s41586-024-07487-w.
21. Bagal, V.; Aggarwal, R.; Vinod, P.; Priyakumar, U. D. MolGPT: molecular generation using a transformer-decoder model. *Journal of chemical information and modeling* **2021**, *62* (9), 2064-2076. DOI:10.1021/acs.jcim.1c00600.
22. Dou, C.; Liu, C.; Yang, F.; Li, F.; Jia, J.; Chen, M.; Ju, Q.; Wang, S.; Dang, S.; Li, T. Baichuan-m2: Scaling medical capability with large verifier system. *arXiv preprint arXiv:2509.02208*.
23. Alber, D. A.; Yang, Z.; Alyakin, A.; Yang, E.; Rai, S.; Valliani, A. A.; Zhang, J.; Rosenbaum, G. R.; Amend-Thomas, A. K.; Kurland, D. B. Medical large language models are vulnerable to data-poisoning attacks. *Nature Medicine* **2025**, 1-9. DOI:10.1038/s41591-024-03445-1.
24. Han, T.; Nebelung, S.; Khader, F.; Wang, T.; Müller-Franzes, G.; Kuhl, C.; Försch, S.; Kleesiek, J.; Haarburger, C.; Bressem, K. K. Medical large language models are susceptible to targeted misinformation attacks. *NPJ digital medicine* **2024**, *7* (1), 288. DOI:10.1038/s41746-024-01282-7.
25. Steinhardt, J.; Koh, P. W. W.; Liang, P. S. Certified defenses for data poisoning attacks. *Advances in neural information processing systems* **2017**, *30*.
26. Li, M.; Sun, Y.; Su, S.; Tian, Z.; Wang, Y.; Wang, X. DPIF: a framework for distinguishing unintentional quality problems from potential shilling attacks. *Computers, Materials and Continua* **2019**. DOI:10.32604/cmc.2019.05379.
27. Levine, A.; Feizi, S. Deep partition aggregation: Provable defense against general poisoning attacks. *arXiv preprint arXiv:2006.14768*.
28. Wang, W.; Levine, A. J.; Feizi, S. Improved certified defenses against data poisoning with (deterministic) finite aggregation. In *International Conference on Machine Learning*, 2022; PMLR: pp 22769-22783.
29. Hong, S.; Chandrasekaran, V.; Kaya, Y.; Dumitraş, T.; Papernot, N. On the effectiveness of mitigating data poisoning attacks with gradient shaping. *arXiv preprint arXiv:2002.11497*.
30. Yang, Y.; Liu, T. Y.; Mirzasoleiman, B. Not all poisons are created equal: Robust training against data poisoning. In *International Conference on Machine Learning*, 2022; PMLR: pp 25154-25165.
31. Geiping, J.; Fowl, L.; Somepalli, G.; Goldblum, M.; Moeller, M.; Goldstein, T. What doesn't kill you makes you robust (er): How to adversarially train against data poisoning. *arXiv preprint arXiv:2102.13624*.
32. Ma, Y.; Zhu, X.; Hsu, J. Data poisoning against differentially-private learners: attacks and defenses. In Proceedings of the 28th International Joint Conference on Artificial Intelligence, 2019.
33. Goodfellow, I. J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.
34. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.
35. Finlayson, S. G.; Chung, H. W.; Kohane, I. S.; Beam, A. L. Adversarial attacks against medical deep learning systems. *arXiv preprint arXiv:1804.05296*.
36. Ma, X.; Niu, Y.; Gu, L.; Wang, Y.; Zhao, Y.; Bailey, J.; Lu, F. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition* **2021**, *110*, 107332. DOI:10.1016/j.patcog.2020.107332.
37. Muoka, G. W.; Yi, D.; Ukwuoma, C. C.; Mutale, A.; Ejiyi, C. J.; Mzee, A. K.; Gyarteng, E. S.; Alqahtani, A.; Al-antari, M. A. A comprehensive review and analysis of deep learning-based medical image adversarial attack and defense. *Mathematics* **2023**, *11* (20), 4272. DOI:10.3390/math11204272.
38. Li, Y.; Zhu, Z.; Zhou, Y.; Xia, Y.; Shen, W.; Fishman, E. K.; Yuille, A. L. Volumetric medical image segmentation: A 3d deep coarse-to-fine framework and its adversarial examples. In *Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics*, Springer, 2019; pp 69-91.
39. Zhang, Y.; Xiang, T.; Hospedales, T. M.; Lu, H. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018; pp 4320-4328.
40. Chen, D.; Mei, J.-P.; Wang, C.; Feng, Y.; Chen, C. Online knowledge distillation with diverse peers. In *Proceedings of the AAAI conference on artificial intelligence*, 2020; Vol. 34, pp 3430-3437. DOI: 10.1609/aaai.v34i04.5746.
41. Jain, N.; Schwarzschild, A.; Wen, Y.; Somepalli, G.; Kirchenbauer, J.; Chiang, P.-y.; Goldblum, M.; Saha, A.; Geiping, J.; Goldstein, T. Baseline defenses for adversarial attacks against aligned language models. *arXiv preprint arXiv:2309.00614*.
42. Li, X.; Tramer, F.; Liang, P.; Hashimoto, T. Large language models can be strong differentially private learners. *arXiv preprint arXiv:2110.05679*.
43. Huber, L.; Kühn, M. A.; Mosca, E.; Groh, G. Detecting word-level adversarial text attacks via shapley additive explanations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, 2022; pp 156-166. DOI: 10.18653/v1/2022.repl4nlp-1.16.
44. Majmudar, J.; Dupuy, C.; Peris, C.; Smaili, S.; Gupta, R.; Zemel, R. Differentially private decoding in large language models. *arXiv preprint arXiv:2205.13621*.
45. Cohen, J.; Rosenfeld, E.; Kolter, Z. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, 2019; PMLR: pp 1310-1320.
46. Zhao, H.; Ma, C.; Dong, X.; Luu, A. T.; Deng, Z.-H.; Zhang, H. Certified robustness against natural language attacks by causal intervention. In *International Conference on Machine Learning*, 2022; PMLR: pp 26958-26970.
47. Liu, X.; Cheng, H.; He, P.; Chen, W.; Wang, Y.; Poon, H.; Gao, J. Adversarial training for large neural language models. *arXiv preprint arXiv:2004.08994*.

48. Zhang, D.; Zhang, T.; Lu, Y.; Zhu, Z.; Dong, B. You only propagate once: Accelerating adversarial training via maximal principle. *Advances in neural information processing systems* **2019**, *32*.
49. Fredrikson, M.; Jha, S.; Ristenpart, T. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 2015; pp 1322-1333. DOI: 10.1145/2810103.2813677.
50. Fredrikson, M.; Lantz, E.; Jha, S.; Lin, S.; Page, D.; Ristenpart, T. Privacy in pharmacogenetics: An {End-to-End} case study of personalized warfarin dosing. In *23rd USENIX security symposium (USENIX Security 14)*, 2014; pp 17-32.
51. Carlini, N.; Tramer, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T.; Song, D.; Erlingsson, U. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, 2021; pp 2633-2650.
52. Nielsen, C.; Tuladhar, A.; Forkert, N. D. Investigating the vulnerability of federated learning-based diabetic retinopathy grade classification to gradient inversion attacks. In *International Workshop on Ophthalmic Medical Image Analysis*, 2022; Springer: pp 183-192. DOI: 10.1007/978-3-031-16525-2_19.
53. Dao, T.-N.; Nguyen, T. P. Performance Analysis of Gradient Inversion Attack in Federated Learning with Healthcare Systems. *REV Journal on Electronics and Communications* **2024**, *13* (3-4). DOI:10.21553/rev-jec.338.
54. Nasr, M.; Shokri, R.; Houmansadr, A. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*, 2019; IEEE: pp 739-753. DOI: 10.1109/SP.2019.00065.
55. Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016; pp 308-318. DOI: 10.1145/2976749.2978318.
56. Du, M.; Yue, X.; Chow, S. S.; Wang, T.; Huang, C.; Sun, H. Dp-forward: Fine-tuning and inference on language models with differential privacy in forward pass. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2023; pp 2665-2679. DOI: 10.1145/3576915.3616592.
57. Ozdayi, M. S.; Peris, C.; FitzGerald, J.; Dupuy, C.; Majmudar, J.; Khan, H.; Parikh, R.; Gupta, R. Controlling the extraction of memorized data from large language models via prompt-tuning. *arXiv preprint arXiv:2305.11759*.
58. Li, Y.; Tan, Z.; Liu, Y. Privacy-preserving prompt tuning for large language model services. *arXiv preprint arXiv:2305.06212*.
59. Chen, T.; Bao, H.; Huang, S.; Dong, L.; Jiao, B.; Jiang, D.; Zhou, H.; Li, J.; Wei, F. The-x: Privacy-preserving transformer inference with homomorphic encryption. *arXiv preprint arXiv:2206.00216*.
60. Li, D.; Shao, R.; Wang, H.; Guo, H.; Xing, E. P.; Zhang, H. Mpcformer: fast, performant and private transformer inference with mpc. *arXiv preprint arXiv:2211.01452*.
61. Hao, M.; Li, H.; Chen, H.; Xing, P.; Xu, G.; Zhang, T. Iron: Private inference on transformers. *Advances in neural information processing systems* **2022**, *35*, 15718-15731.
62. Zheng, M.; Lou, Q.; Jiang, L. Primer: Fast private transformer inference on encrypted data. In *2023 60th ACM/IEEE Design Automation Conference (DAC)*, 2023; IEEE: pp 1-6. DOI: 10.1109/DAC56929.2023.10247719.
63. Yang, Q.; Liu, Y.; Chen, T.; Tong, Y. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* **2019**, *10* (2), 1-19. DOI:10.1145/3298981.
64. Wang, B.; Zhang, Y.; Cao, Y.; Li, B.; McMahan, H.; Oh, S.; Xu, Z.; Zaheer, M. Can Public Large Language Models Help Private Cross-device Federated Learning? Mexico City, Mexico, June, 2024; Association for Computational Linguistics: pp 934-949. DOI: 10.18653/v1/2024.findings-naacl.59.
65. Zhang, Z.; Yang, Y.; Dai, Y.; Wang, Q.; Yu, Y.; Qu, L.; Xu, Z. FedPETuning: When federated learning meets the parameter-efficient tuning methods of pre-trained language models. In *Annual Meeting of the Association of Computational Linguistics 2023*, 2023; Association for Computational Linguistics (ACL): pp 9963-9977. DOI: 10.18653/v1/2023.findings-acl.632.
66. Nwadike, M.; Miyawaki, T.; Sarkar, E.; Maniatakos, M.; Shamout, F. Explainability matters: Backdoor attacks on medical imaging. *arXiv preprint arXiv:2101.00008*.
67. Ding, Y.; Wang, Z.; Qin, Z.; Zhou, E.; Zhu, G.; Qin, Z.; Choo, K.-K. R. Backdoor attack on deep learning-based medical image encryption and decryption network. *IEEE Transactions on Information Forensics and Security* **2023**, *19*, 280-292. DOI:10.1109/TIFS.2023.3322315.
68. Jin, R.; Huang, C.-Y.; You, C.; Li, X. Backdoor Attack on Unpaired Medical Image-Text Foundation Models: A Pilot Study on MedCLIP. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2024; IEEE: pp 272-285. DOI: 10.1109/SaTML59370.2024.00020.
69. Joe, B.; Park, Y.; Hamm, J.; Shin, I.; Lee, J. Exploiting Missing Value Patterns for a Backdoor Attack on Machine Learning Models of Electronic Health Records: Development and Validation Study. *JMIR Med Inform* **2022**, *10* (8), e38440. DOI:10.2196/38440.
70. Lyu, W.; Bi, Z.; Wang, F.; Chen, C. Badclm: Backdoor attack in clinical language models for electronic health records. *arXiv preprint arXiv:2407.05213*.
71. Feng, Y.; Ma, B.; Zhang, J.; Zhao, S.; Xia, Y.; Tao, D. Fiba: Frequency-injection based backdoor attack in medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022; pp 20876-20885.
72. Wallace, E.; Zhao, T. Z.; Feng, S.; Singh, S. Concealed data poisoning attacks on NLP models. *arXiv preprint arXiv:2010.12563*.
73. Cui, G.; Yuan, L.; He, B.; Chen, Y.; Liu, Z.; Sun, M. A unified evaluation of textual backdoor learning: Frameworks and benchmarks. *Advances in Neural Information Processing Systems* **2022**, *35*, 5009-5023.
74. McInnes, L.; Healy, J. Accelerated hierarchical density based clustering. In *2017 IEEE international conference on data mining workshops (ICDMW)*, 2017; IEEE: pp 33-42. DOI: 10.1109/ICDMW.2017.12.
75. Yan, J.; Gupta, V.; Ren, X. Bite: Textual backdoor attacks with iterative trigger injection. *arXiv preprint arXiv:2205.12700*.
76. Chen, C.; Dai, J. Mitigating backdoor attacks in lstm-based text classification systems by backdoor keyword identification. *Neurocomputing* **2021**, *452*, 253-262. DOI:10.1016/j.neucom.2021.04.105.
77. Wan, A.; Wallace, E.; Shen, S.; Klein, D. Poisoning language models during instruction tuning. In *International Conference on Machine Learning*, 2023; PMLR: pp 35413-35425.
78. Shen, J.; Qu, Y.; Zhang, W.; Yu, Y. Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, 2018; Vol. 32. DOI: 10.1609/aaai.v32i1.11784.
79. Li, H.; Chen, Y.; Zheng, Z.; Hu, Q.; Chan, C.; Liu, H.; Song, Y. Backdoor removal for generative large language models. *arXiv preprint arXiv: 2405.07667*.
80. Liu, Z.; Dou, G.; Tan, Z.; Tian, Y.; Jiang, M. Towards safer large language models through machine unlearning. *arXiv preprint arXiv:2402.10058*.

DOI: 10.36838/IJHSR86.108

81. Clusmann, J.; Ferber, D.; Wiest, I. C.; Schneider, C. V.; Brinker, T. J.; Foersch, S.; Truhn, D.; Kather, J. N. Prompt Injection Attacks on Large Language Models in Oncology. *arXiv preprint arXiv:2407.18981*.
82. Clusmann, J.; Schulz, S. J.; Ferber, D.; Wiest, I. C.; Fernandez, A.; Eckstein, M.; Lange, F.; Reitsam, N. G.; Kellers, F.; Schmitt, M. A pen mark is all you need-Incidental prompt injection attacks on Vision Language Models in real-life histopathology. *medRxiv* **2024**, 2024.12.11.24318840.
83. Hines, K.; Lopez, G.; Hall, M.; Zarfati, F.; Zunger, Y.; Kiciman, E. Defending against indirect prompt injection attacks with spotlighting. *arXiv preprint arXiv:2403.14720*.
84. Willison, S. Delimiters won't save you from prompt injection (2023). *URL https://simonwillison. net/2023/May/11/delimiters-wont-save-you 4*.
85. Chen, S.; Piet, J.; Sitawarin, C.; Wagner, D. Struq: Defending against prompt injection with structured queries. *arXiv preprint arXiv:2402.06363*.
86. Wallace, E.; Xiao, K.; Leike, R.; Weng, L.; Heidecke, J.; Beutel, A. The instruction hierarchy: Training llms to prioritize privileged instructions. *arXiv preprint arXiv:2404.13208*.
87. Suo, X. Signed-prompt: A new approach to prevent prompt injection attacks against LLM-Integrated applications. *AIP Conference Proceedings* **2024**, *3194* (1). DOI:10.1063/5.0222987.
88. Yi, J.; Xie, Y.; Zhu, B.; Kiciman, E.; Sun, G.; Xie, X.; Wu, F. Benchmarking and defending against indirect prompt injection attacks on large language models. *arXiv preprint arXiv:2312.14197*.
89. Rai, P.; Sood, S.; Madisetti, V. K.; Bahga, A. Guardian: A multi-tiered defense architecture for thwarting prompt injection attacks on llms. *Journal of Software Engineering and Applications* **2024**, *17* (1), 43-68. DOI:10.4236/jsea.2024.171003
90. Perez, E.; Huang, S.; Song, F.; Cai, T.; Ring, R.; Aslanides, J.; Glaese, A.; McAleese, N.; Irving, G. Red Teaming Language Models with Language Models. In *2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates, December, 2022; Association for Computational Linguistics: pp 3419-3448. DOI: 10.18653/v1/2022.emnlp-main.225.

Author

Songzhe Kang is a junior at Zhengzhou Middle School, China. He is interested in mathematics, computer science, and artificial intelligence, and is committed to studying artificial intelligence further in college in the future.