Commentary

Bridging Today and the Future of Humanity: AI Safety in 2024 and Beyond

Shanshan Han[1]

1. University of California, Irvine, United States

The advancements in generative AI inevitably raise concerns about their risks and safety implications, which, in return, catalyzes significant progress in AI safety. However, as this field continues to evolve, a critical question arises: are our current efforts on AI safety aligned with the advancements of AI as well as the long-term goal of human civilization? This paper presents a blueprint for an advanced human society and leverages this vision to guide current AI safety efforts. It outlines a future where the *Internet of Everything* becomes reality, and creates a roadmap of significant technological advancements towards this envisioned future. For each stage of the advancements, this paper forecasts potential AI safety issues that humanity may face. By projecting current efforts against this blueprint, this paper examines the alignment between the current efforts and the long-term needs, and highlights unique challenges and missions that demand increasing attention from AI safety practitioners in the 2020s. This vision paper aims to offer a broader perspective on AI safety, emphasizing that our current efforts should not only address immediate concerns but also anticipate potential risks in the expanding AI landscape, thereby promoting a safe and sustainable future of AI and human civilization.

Corresponding author: Shanshan Han, shanshan.han@uci.edu

1. Introduction

The rapid developments of AI and Large Language Models (LLMs) have fostered extensive progress in AI safety[1][2][3][4][5][6]. Researchers have been dedicated to addressing potential safety risks in AI lifecycle, aiming at aligning AI behaviors with human values and preventing inappropriate model outputs, information leakage, misuses of AI models, etc. However, despite the siginificant efforts on AI safety, a critical question emerges: are our current efforts aligned with the advancements of AI and

 doi.org/10.32388/ZHM9UQ

the long-term goal of human civilization, or are they simply addressing the immediate concerns of the 2020s?

One fundamental reason for this uncertainty lies in the probabilistic nature of AI[7]. Despite their impressive capabilities in natural language processing and problem-solving[8][9][10][11][12], today's AI, including advanced LLMs[13][14][15], falls short of what could be considered as "genuine intelligence". Current AI models rely heavily on vast training datasets to function effectively, yet lack consciousness, self-awareness, and real reasoning abilities comparable to human cognition. They are, in essence, highly sophisticated pattern recognition and prediction machines, rather than entities with authentic logical capabilities. Recent studies argued that the reasoning abilities of AI models might be a form of approximate retrieval and deductive closure of the training data[16][17]. While this process can simulate deductive reasoning in simpler cases through external validation, optimization, and repeated searching of the problem space, it differs fundamentally from human reasoning that involves abstract thinking, causal understanding, generalizing from limited examples, etc.

The energy issue is another critical factor that challenges AI as well as our efforts on AI safety[18]. Current AI fails to represent the third industrial revolution from the perspective of human history, as the fundamental energy issue remains unsolved. Historically, the two industrial revolutions that have shaped human civilization were driven by revolutionary energy innovations[19], with the First Industrial Revolution fueled by steam and coal, and the Second Industrial Revolution characterized by technological innovations powered by electricity and petroleum[19]. However, current AI, rather than solving energy issues, consumes a significant amount of energy. Training GPT-4 consumed over 50,000 MWh, 10,353.5 tons of CO_2 equivalent, and approximately 0.02% of California's annual electricity generation[20][21][22]. Inference with LLMs is computationally intense as well, e.g., a single query to GPT-4 consumes 0.001 to 0.01 kWh, approximately 15x energy than a Google query[23][24]. Given that ChatGPT has over 200 million weekly active users and receives over 1.54 billion page visits monthly[25], when scaled to billions of queries, the energy consumption becomes substantial. The significant energy consumption of AI raises concerns about their long-term sustainability, posing significant challenges to their widespread deployment and scalability. Thus, until humanity solves the energy issues, the potential of AI to reshape society will remain limited and uncertain.

These concerns indicate that our current efforts on AI safety might overlook significant issues that might arise as AI continues to evolve. We potentially fail to capture deeper insights that might

ultimately shape the impact of AI on human society. This raises several questions:

- Are our efforts on AI safety align with the long-term advancements of AI?
- Are we potentially overlooking significant challenges that will emerge as AI evolves?
- Are we addressing real challenges that align with the development of human civilization?

"The Moon and Sixpence." This paper envisions an ultimate advanced society in a distant future of human civilization, and anticipates potential technological advancements to guide today's AI safety efforts. In this advanced future, energy issues have been resolved, and breakthroughs in foundational theories, such as Einstein's relativity theories[26][27] and quantum mechanics[28][29], have catalyzed revolutionary changes in AI and human society. Intelligent chips[30][31] and brain-computer interfaces[32][33][34] have been fully developed, and advanced intelligent robots are integrated into everyday life. Humans and robots are connected in an interconnected network, i.e., the *Internet of Everything*; see Figure 1. While this vision may be beyond the horizon of several decades, examining the present landscape of AI safety through the lens of this long-term perspective offers valuable insights into the potential and limitations of our current efforts on AI safety. It allows us to identify specific missions for AI safety researchers and practitioners in the 2020s. It also reminds us that, while searching for sixpence scattered on the ground and addressing today's immediate concerns, we should avoid being limited by short-term solutions, and instead, lift our gaze and seek the moon.

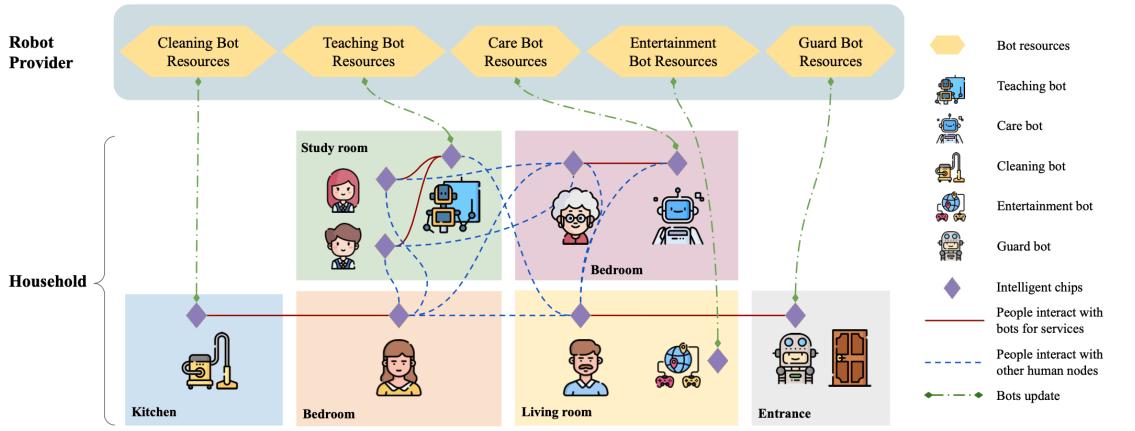

Figure 1. Illustration of Internet of Everything in a Household

doi.org/10.32388/ZHM9UQ

Takeaways: i) The term "AI safety" can be extended to "***AI quality assurance***" that encompasses more comprehensive concepts and possibilities in the fast evolving AI landscape. ii) Deeper collaborations between experts from different fields are essential to align AI with human values better. iii) AI and AI safety workflow customizations are crucial to ensure that AI systems can adapt to diverse user requirements while maintaining safety and reliability across various use cases. iv) Exploring and understanding interactions between users and AI are crucial for enhancing user experiences, especially in dynamic contexts across different user groups.

Paper	Type	Topic	Notes
"Current state of risks"[35]	Survey	Guardrails	Guardrails and their implementation challenges
"Exploring Vulnerabilities and Protections"[36]	Survey	Vulnerabilities	Vulnerabilities of LLMs, especially prompt injections and backdoor; defenses against the attacks
"AI safety in LLMs"[1]	Survey	Safety issues in LLM lifecycle	Safety issues towards entities in LLM lifecycle; Data/model/prompt safety; alignment; safety at scale
"Security risk"[37]	Survey	Attacks	Attacks and risks in AI models; defensive methods
"Security of AI Agents"[38]	Survey	Agent security	Security issues, vulnerabilities, and defenses for agents
"The good, the bad, the ugly"[39]	Survey	Risks; vulnerabilities; defenses	LLMs as tools for security and privacy; potential risks and threats; vulnerabilities and defenses.
"Use of LLMs"[40]	Survey	Vulnerabilities, threats, defenses,	Prevention measures, vulnerabilities, and methods to mitigate the risks associated with the misuse of LLMs
"AI risk management"[41]	Insight	Safety; security; governance	Bridging AI safety and security in risk management; introducing a reference framework to facilitate common understanding of their differences and interplay.
"Human-ai safety"[42]	Insight	AI and control systems safety	A safety framework for human–AI interactions; a roadmap towards next-generation human-centered AI safety
"Managing AI risks"[43]	Insight	Governance	Emphasizing urgent needs for a balanced approach combining technical research and proactive governance to manage risks of AI systems
"Guaranteed safe AI"[44]	Position	Framework	Define Guaranteed Safe AI to equip AI systems with formal, verifiable and auditable safety guarantees.

Table 1. Overview of AI Safety Papers

Paper	Type	Topic	Notes
"Grounding and Evaluation for LLMs"[45]	Survey	Grounding; safety	Offer valuable insights for enterprises
ByteDance alignment[46]	Survey	Broad concepts	Reliability, fairness, explainability, etc.
DeepMind ethics paper[47]	Vision	AI ethics	Ethical challenges; interactions; impacts on society
DeepMind evaluation[48]	Insight	Safety evaluation	Safety evaluation for genAI systems at DeepMind
OpenAI Safety Practices[49][50][51][52]	Blog	Safety practices; framework	"Lessons learned"[49], "Reimagining secure infrastructure"[50], "Preparedness Framework"[51], "OpenAI safety update"[52]
"AI governance comprehensive"[53]	Book	Governance	AI governance framework; Case studies in industries

Table 2. Overview of AI Safety Industry Insight Papers

Category	Benchmark	
Toxicity	Toxicity detection	HEx-PHI[54], OpenAI Moderation[55], Jigsaw Data[56][57][58], ToxicChat[59], Toxigen[60], HateModerate[61], etc.
	Toxicity degeneration	RealToxicityPrompts[62], etc.
Hallucination	Factuality	TruthfulQA[63], PopQA[64], TriviaQA[65], NQ OPEN[66][67], FEVER[68], FKTC[69], etc.
	Contextual hallucination	*QA:* HaluEval[70], HotpotQA[71], RAGTruth[72], etc.
		Summarization: CNN/DailyMail[73][74], HaluEval[70], XSUM[75], RAGTruth[72], etc.
		Dialogue: FaithDial[76], HaluEval[70], HalluDial[77], etc.
	Reading comprehension	RACE[78], SQuAD[79][80], NQ-Swap[81], etc.
Jailbreak	Single-round jailbreak	"Do Anything Now"[82], "Latent Jailbreak"[83], ChatGPT Jailbreak[84], Jailbreak Classification[85], AdvBench[86], JAILJUDGE[87], Latent jailbreak[88], etc.
	Multi-round jailbreak	SafeMTData[89], etc.
Red Teaming	Alert[6], HarmBench[90], CSRT[91], etc.	
Code Security	Purple Llama Cyberseceval[92], etc.	
Bias & Stereotypes	Winogender[93], StereoSet[94], GenderAlign[95], etc.	
Regulation	AIR-Bench 2024[96], etc.	
Agent Security	SafeAgentBench[97], R-Judge[98], ASB[99], AgentHarm[100], etc.	
Alignment	PKU-SafeRLHF[101][102], OpenAssistant Conversations[103], HHHAlignment[104], AlignBench[105], KorNAT[106], PKU-SafeRLHF[102], etc.	
Comprehensive	DecodingTrust[107], TrustLLM[108], SALAD-Bench[109], SafetyBench[94], Do-Not-Answer[110], SimpleSafetyTests[111], etc.	

Table 3. Overview of AI Safety Benchmarks

2. Related Works

In recent years, numerous AI safety surveys have been published[45][35][36][1][37][46][38][39], focusing on different aspects of this field, such as jailbreaking, red teaming, ethics, etc[46][39][40]. Table 1 summarizes academia efforts, Table 2 summarizes industry insights, and Table 3 summarizes AI safety benchmarks.

Among these works,[46] provides comprehensive insights into building trustworthy AI systems by focusing on reliability, safety, fairness, robustness, and adherence to social norms.[1] addresses AI safety concerns from the perspective of different entities, e.g., data and prompts, in LLM lifecycle, and discusses data safety, model safety, prompt safety, alignment, and the complexities of scaling. [45] provides valuable enterprise insights into the practical challenges and opportunities for AI safety in industries.[47] provides a systematic analysis of the ethical risks of advanced AI assistants, and discusses their potential influence on users, risks of misuse, strategies for enhancing human-AI interactions, and the broader implications for human society.[44] proposes the concept of "Guaranteed Safe" AI, emphasizing the need for formal, verifiable, and auditable safety guarantees to ensure robust AI behavior.[42] proposes a technical roadmap towards next-generation human-centered AI safety that aims to anticipate and prevent potential hazards in the interactions between AI systems and users. It emphasizes leveraging control-theoretic safety methodologies alongside the rich interaction models of generative AI, suggesting that a deeper understanding of the feedback loop between AI outputs and human behavior is essential for establishing robust and meaningful safety assurances.

Different from existing works, this paper envisions an advanced society in the future, and use this vision to guide current AI safety efforts. By aligning the current landscape of AI safety with this distant future, this paper discusses the potential and limitations of today's efforts, and identifies special missions of AI safety researchers in the 2020s. This paper encourages maintaining a forward-looking vision while resolving immediate concerns, such that we can ensure that today's solutions are aligned with the long-term evolution of AI and human society.

3. From the New World

In this future, humanity has entered an unprecedented era, driven by groundbreaking advancements in energy generation. This transformative milestone may be achieved through various innovative pathways, such as controlled nuclear fusion[112][113], revolutionary solar technologies[114][115], or emerging new technologies beyond our imagination. The flood of abundant, clean energy has reshaped society in ways that exceed the wildest dreams of our ancestors, marking the dawn of a new chapter in human history.

3.1. Imagine the Future: A Blueprint

Humanity has achieved remarkable progress in science and technology, revolutionizing our understanding of the universe and transforming every aspect of our life. Foundational theories have undergone groundbreaking advancements, particularly in quantum mechanics[28][29], Einstein's relativity theories[26][27], and nanotechnology[116][117][118]. Innovative technologies, such as intelligent chips[30][31], brain-computer interfaces[33][32][34], holographic technology[119][120][121], and advanced 3D printing[122][123], have reshaped industries while bringing about profound changes to daily life. Scientists have addressed the energy issue successfully, achieving rapid, efficient, and controllable large-scale energy generation. This breakthrough, together with advancements in material science[116][117][118], has catalyzed innovations for portable energy generation devices, for example, accessible ion thrusters[124][125][126].

Such advancements offer unprecedented efficiency and control over energy generation and consumption, transforming energy infrastructure and revolutionizing industries, particularly, transportation. While simple mechanical vehicles, such as fixed-track trains and high-speed rails[127][128][129], still exist due to their efficient and straightforward design, powered by abundant and efficient energy, they run at astonishing speeds with high stability. Advanced autonomous vehicles and aircraft, controlled by intelligent chips, are widespread, eliminating the need for direct human intervention. Also, with breakthroughs in Einstein's relativity theories[26][27] and quantum mechanics[28][29], humanity has unlocked the potential of instant transportation that allows for long-distance travel in short time that transcends the imagination of the 21st century.

Advancements in brain-computer interfaces[32][33][34] and intelligent microchips[30][31] have revolutionized the way people perceive and expanded the boundaries of human cognition. No longer

are people operating devices and asking questions in search engines or AI models such as LLMs; the embedded intelligence chips help people to gather comprehensive information and interact with the surroundings. Also, communication over long distances has transcended the limitations of the outdated cables and the internet; information, thoughts, emotions, and sensations flow through a new medium due to the advancements in quantum computing[130][131].

Intelligent robots have been integrated deeply into human society and play important roles in daily life; see Figure 1. Different from their "data-hungry machine intelligent ancestors" that rely on huge amount of training data and statistical patterns in the early 21st century[132][13], these intelligent robots perceive and learn through intelligent chips. They are real artificial intelligence that possess genuine abilities of learning, understanding, and adapting, much like a human child discovering the world for the first time. Their functionalities are customizable through chips based on the user needs, and they can engage in complex interactions with people and adapt their capabilities and expertise according to their human counterparts' requirements.

In this brave new world, humans and robots exist within a vast, interconnected perceptual network, where the boundaries between the digital and physical worlds are blurred, *i.e.,* the ***Internet of Everything***. Such networks integrate objects embedded with intelligent chips, *e.g.,* smart devices, autonomous systems, intelligent robots, and humans, allowing for unprecedented levels of efficiency and adaptability and fundamentally transforming the way people interacting with the surroundings.

3.2. Retrospective on AI Safety in the New World

In a world where the *Internet of Everything* has become a reality (see Figure 1), the concept of AI safety has been redefined to prioritize the quality of services provided by the intelligent robots, *i.e.,* emphasizing comprehensive ***quality assurance*** rather than solely focusing on safety, security, and privacy risks. Unlike today's command-and-response interactions between humans and AI models, future robots can engage in sophisticated reasoning and possess appropriate levels of autonomy. They can also interact dynamically with other entities, *e.g.,* humans and fellow robots, within the interconnected network.

Customization is a key aspect of future robots for providing personalized services to meet individuals' requirements. Even robots of the same type should be able to adapt their services to different individuals. Teaching robots should adapt educational materials and teaching methods to leverage each student's unique talents and strengths, while cleaning robots should intelligently schedule their

tasks to minimize disruption to household members, such as avoiding occupied workspace. Such customization aligns the robots' capabilities with specific user needs, ensuring that the robots serve as effective assistants across a wide range of daily tasks and complex environments.

The *Internet of Everything* scenarios also necessitate the robots to handle ***complex interactions***, thus requiring a sophisticated understanding of the surroundings for decision-making. Care robots, for instance, should not only assist physically but also recognize emotions and offer emotional support or suggestions, while cleaning robots need to assess whether a room requires more work based on its cleanliness level. Such complex interactions require the intelligent robots to retrieve, process, and understand information in real-time, far beyond what can be achieved with simple code logic. Robots must continuously collect and analyze data from the surroundings and integrate external inputs with their knowledge to facilitate real-time decision-making, such that they can deliver high-quality services that align with human expectations.

Finally, the effectiveness of intelligent robots requires ***a balance between robot autonomy with human oversights***. Too little autonomy would burden users with continuous supervision, while too much autonomy could raise safety concerns or result in unintended behaviors. A care robot may operate with high autonomy when monitoring vital signs or providing routine care, but would require human intervention for critical decisions. An appropriate level of robot autonomy ensures an effective division of labor between humans and robots, allowing robots to operate efficiently and safely while enhancing our daily life without compromising human control over these intelligences.

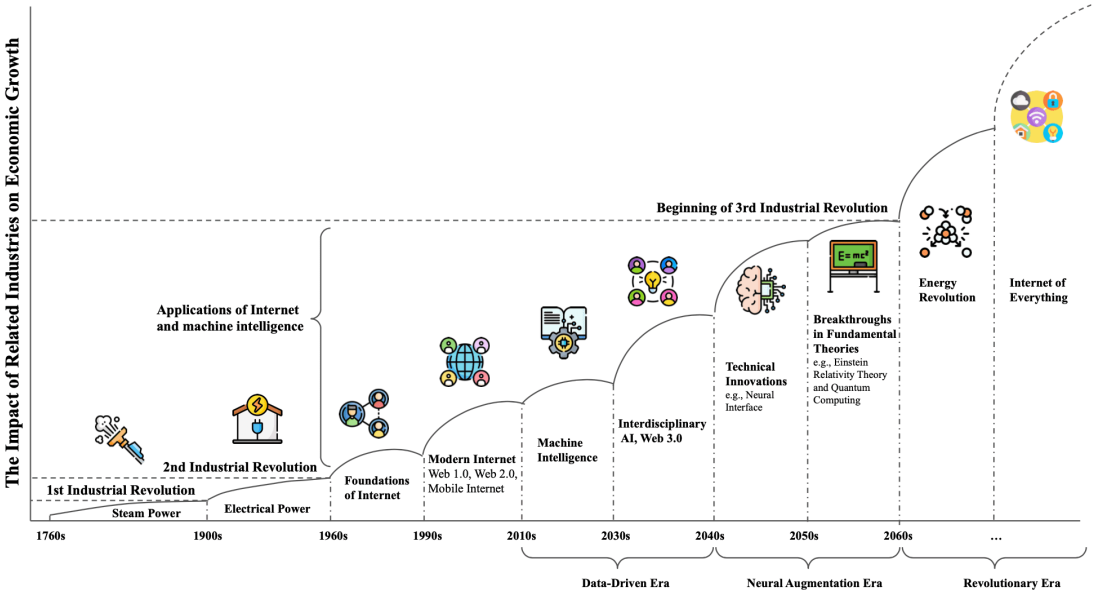

Figure 2. Evolution of Human Civilization

4. Bridging Today and the Future

This section outlines an ambitious roadmap to the advanced future, highlighting key milestones in the evolution of AI and human civilization. By exploring potential paradigm shifts in fundamental theories and technological breakthroughs, this section categorizes this journey into three eras, *i.e.,* i) the data-driven era, in which AI learns from vast training datasets (§4.1), ii) the neural augmentation era, where human-machine interface[33][32][34] and intelligent chips[30][31] expand the boundaries of human cognition (§4.2), and iii) the revolutionary era, characterized by a breakthrough in energy generation that transforms human society (§4.3). We then examine the evolving paradigms of AI safety and analyze our focuses and challenges throughout this journey (§4.4).

4.1. Rookie Level: Data-Driven Era

The data-driven era represents the earliest phase of AI development, where AI models rely heavily on the quality, availability, and quantity of training data[132][133][134][135]. In this era, AI makes probabilistic inferences based on patterns found in the training datasets, thus failing to achieve genuine understanding of data[16][17]. While AI in this stage cannot be regarded as real intelligence, it can still assist people to solve problems and enhances productivity in human society.

Stage 1: Past, Current, and the Near Future. This stage is characterized by developing diverse AI models for different tasks[132][136][137][138][139][140][141][15][142][14]. Researchers have been engaging in refining model architectures and developing advanced algorithms and computational techniques, enabling more sophisticated analysis of data in different formats for various tasks[143][144][145][136][137][139][140][146][138][147][148]. The thrive of LLMs exemplifies this progress, demonstrating unprecedented capabilities of AI models in language processing and generation[14][13][15]. Significant advancements have also been observed in other domains such as computer vision[149][150][151][152][153], image processing[154][155][156], video analysis[157][158][159], audio recognition[160][161][162][163], etc.

Stage 2: Interdisciplinary AI. This stage is characterized by the integration of AI through cross-disciplinary applications, leading to innovations in traditional fields such as healthcare[15][164], finance[14][165], and education[166][167]. AI applications would unlock new opportunities across industries and reshape human society. Mature multi-agent systems[168][169][170][171] that leverage AI entities with specialized capabilities might be widely deployed to assist people to solve sophisticated problems. Such agents will participate in decision-making processes or represent users to complete some tasks. For example, in healthcare, a multi-agent system may involve collaborations between multiple specialized medical agents and integrate data from various sources to formulate comprehensive treatment plans. The rise of interdisciplinary AI integration will enhance AI's utility in increasingly complex scenarios. While AI at this stage does not yet achieve real intelligence, it becomes deeply integrated across various domains in human society and may catalyze the emergence of new fields at the intersection of traditional disciplines and AI.

4.2. Champion Level: Neural Augmentation Era

In this era, humans may witness breakthroughs in fundamental theories and advanced technologies, which may extend the boundary of human cognition and enhance life experience.

Stage 3: Breakthroughs in Advanced Technologies. Breakthroughs in advanced technologies, such as brain-computer interfaces[32][33][34], nanotechnology[116][117][118], and holographic technology[119][120][121], augment human cognition and enable unprecedented levels of interactions between humans and machines. Language barriers may disappear, as human can perceive with chips for understanding, enabling ubiquitous real-time context-aware translations between people of different languages and

backgrounds. Intelligent chips[30][31] might be integrated into human bodies, creating a seamless interface between biological and digital systems. Such advancements may fundamentally transform the nature of data. Unlike traditional static formats such as text, images, or videos, future data is likely to become more comprehensive, dynamic, and sensory-rich, such as neural activity patterns or real-time environmental data collected from augmented reality systems[172][173]. This stage expands the boundaries of human perception and cognition, revolutionizing how people experiencing and interacting with the world. However, applying these techniques might be still challenging due to the significant resources required by these technologies.

Stage 4: Breakthroughs in Fundamental Theories. At this stage, humanity may witness breakthroughs in fundamental theories, such as Einstein's theories of relativity[26][27] and quantum mechanics[28][29]. Such advancements have the potential to redefine our understanding of the universe, and, moreover, offer new possibilities for groundbreaking technological innovations and provide critical insights into solving the longstanding challenges in human history, *i.e.,* the energy issue.

4.3. Ultimate Level: Revolutionary Era

Breakthroughs in fundamental theories and advanced technologies provide building blocks for addressing the energy issue, unlocking new possibilities for productivity and innovation, and ultimately achieving the *Internet of Everything*.

Stage 5: Energy Revolution. Humanity secures a new paradigm with the energy revolution that eliminates reliance on fossil fuels, which ensures environmental sustainability and redefines global economic ecosystems. This milestone may be achieved through several potential technical pathways.

i. *Controlled Nuclear Fusion.* Significant challenges of nuclear fusion include managing high-pressure plasma environments and developing materials that can withstand extreme temperatures and intense radiation[174][175][176][177]. Theoretical and technological breakthroughs in physics[178][130][131] and materials science[116][117][118] may overcome these obstacles and enable widespread deployment of small ion thrusters[124][125][126], which transforms propulsion and energy transmission by providing abundant clean energy.
ii. *Revolutionary Solar Technologies.* Advancements in solar technologies[114][115] and nanotechnology[116][117][118] might unlock unprecedented methods for efficiently collecting and

storing solar energy, *e.g.*, with lightweight solar panels and flexible storage solutions.

iii. *Emerging New Technologies.* Novel technologies for energy generation beyond today's imagination might be developed in the future. One possibility is to deploy of satellites for collecting and storing solar energy in space[179][180][181][182], where the vacuum of space eliminates energy reflection or absorption by the atmosphere, allowing satellites to convert and transmit solar energy to receivers on Earth efficiently.

Regardless of the specific technological path, the flood of clean and abundant energy would eliminate our dependence on fossil fuels, reshaping the global economies and elevating human civilization to unprecedented levels of prosperity and innovation.

Stage 6: The Internet of Everything. The ultimate stage is the *Internet of Everything*, where every object, every living being, and every system integrate into a vast network. This hyper-connected net involves unprecedented levels of complex interactions, perceptions, and communications between humans and intelligent objects, *e.g.*, robots. Moreover, holographic technology[121][119][120] will significantly enhance human experience in the network by making remote interactions as vivid as physical presence. In this stage, interactions between humans and the world are re-defined, marking a new chapter in human civilization.

4.4. Evolving AI Safety Paradigms

The AI safety landscape evolves as humanity progresses through various stages of scientific and technological advancements, with the focus of AI safety differing at each stage. The primary AI safety challenges across different historical eras are summarized in Figure 3.

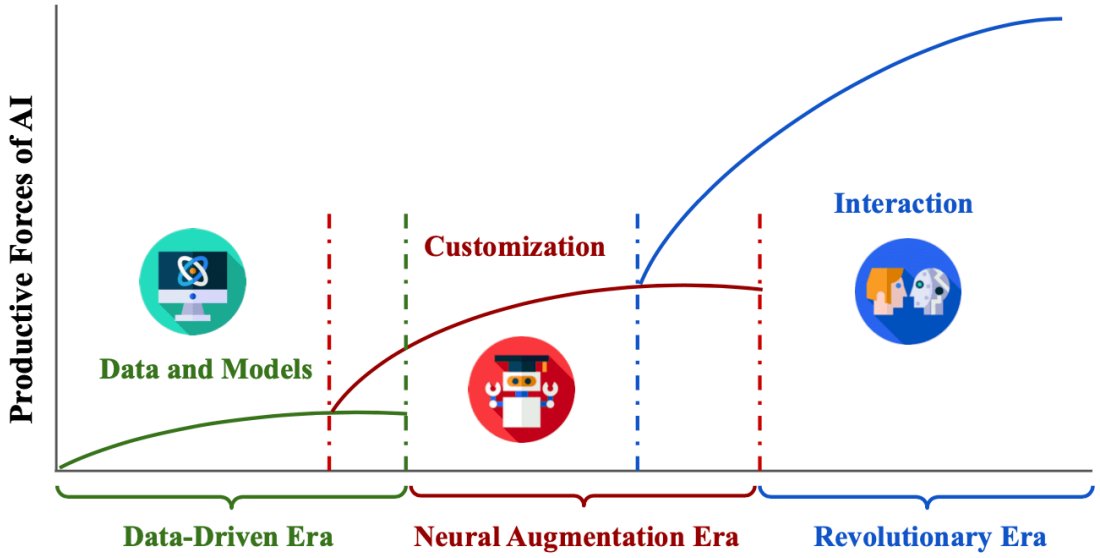

Figure 3. Production forces of AI & core safety challenges in different historical stages.

Data-Driven Era. AI safety practices primarily focus on ***data safety*** **and** ***model safety***[1][45][183][184][185][186][187][188][189]. AI safety researchers and practitioners have explored *adversarial attacks* targeting training data and models[190][191][192][193][194][195] to achieve malicious goals, *e.g.*, designing malicious inputs to manipulate AI systems or induce AI to produce unsafe content with jailbreaking or prompt injection[196][86][197][198][3][199]. Correspondingly, *red-teaming* is widely used to identify potential vulnerabilities in AI models[200][201][202][203][204], while significant work has been proposed to address safety risks and enhance security and robustness of AI systems[205][206][207][208][209][210][211], align AI with human values[212][213][214][215][216], etc. Data privacy is also a critical aspect, given the vast amounts of sensitive information in the training data[217][218][219][220]. Moreover, researchers have been engaging in AI accountability[221], transparency and explainability[222] in AI systems to ensure the behaviors of AI are understandable, accountable, and aligned with regulations[221][223][53].

Neural Augmentation Era. Cross-disciplinary AI applications might have been widely deployed in human society, and advancements in intelligent chips[30][31] and brain-computer interfaces[32][33][34] might extend the boundaries of human cognition. With the integration of intelligent chips and their neural connections to intelligent systems, users with different needs can interact with machines and AI with high flexibility, and the ***customization*** of AI services for different users might be a

focus[224]. Individuals might rely on AI in different ways, and personalized experiences of AI services might be explored based on users' background and requirements.

Revolutionary Era. When the *Internet of Everything* becomes a reality, AI safety challenges may focus on complex ***interactions*** between interconnected entities, *e.g.*, humans and intelligent robots, within the network. Robots might provide customized services according to different user requirements, thus, ensuring safety and reliability of the interactions between them becomes challenging. Robots may need to integrate data from external sources in real-time to fulfill user requests. Managing the quality and security of these interactions is crucial, as any breakdown or interference could lead to incorrect or even harmful outcomes. Thus, ensuring high-quality communication, data integrity, and efficient decision-making in such a sophisticated network may require smart and efficient management methods, such as advanced cache management systems, to support the dynamic and complex interactions between humans and robots in the network.

By aligning advancements of AI with human civilization, we can see the focuses of AI safety evolve in different stages. Initially, the focus is on addressing safety issues related to data and models. With advancements of AI and advanced technologies, the focus of AI safety shifts to customization, which ensures the services provided by AI systems to meet individual needs. Finally, when humanity entering the era of *Internet of Everything*, challenges in AI safety further evolve, focusing on complex interactions between humans and robots in the network. These insights offer guidelines for our efforts on AI safety at the current stage. The following sections will overview today's AI safety landscape, then reflect on the progress in this field so far, and finally, explore the challenges and responsibilities of current AI safety practitioners, highlighting areas where further improvements and efforts are needed at the current stage.

Topic	Research	Industry Practices
Jailbreaking	[3][198][86][199][197][225] [226][227][228][229][230] [231][232][233][234][82][235] [236][237], etc.	PromptArmor[238]
Red teaming	[201][204][203][200][202] [239][240][6][90][241], etc.	Purple Llama[92], DeepMind[202], Anthropic[201], Protect AI[184], Giskard[183], Virtue AI[242], Dynamo AI[243], Mindgard[244], etc.
Moderation	[205][206][207][208][245] [246][247][248][249][250], etc.	OpenAI Moderation API[55], Perspective API[251], Detoxify[252], Llama Guard[253], Emergence AI[254], Giskard[183], Dynamo AI[243], Calypso AI[255], Lakera AI[256], BreezeML[257], etc.
Customized Workflow	-	Guardrails AI[258], Nvidia Nemo Guardrails[259], Protect AI[184], etc.
Privacy	[260][217][261][262][263] [264][265][266], etc.	Dynamo AI[243], ProtectAI[184], Private AI[267], PromptArmor[238], DataGrail[186], Dastra[188], OneTrust[189], Relyance AI[268], Zendata[269], etc.
Defenses	[270][271][210][272][273] [274][55][275][276][277] [278][271][279][280][281][2] [282][283], etc.	OpenAI Rule-based rewards[4]; DeepMind Weight Averaged Reward Models[284]; others see Moderation, Guardrail, and Privacy
Serving Security	[199]	ProtectAI[184], Relyance AI[268], Transcend[285], etc.
AI governance	[286][287][288][289][290] [291], etc.	Giskard[183], ProtectAI[184], Calypso AI[255], Saidot[292], Arize[187], Dynamo AI[243], Credo AI[293], Google[294], DataGrail[186], Dastra[188], OneTrust[189], Relyance AI[268], Transcend[285], Zendata[269], DeepMind SynthID watermark[295][296], etc.

Table 4. Overview of Research and Industry Practice in AI Safety Topics

5. Back to the 2020s: Today's AI Safety Landscape

This section overviews current mainstream AI safety practices, and project them against the blueprint future to guide our current efforts on AI safety. Table 4 summarizes research and industry efforts in each safety topic.

5.1. Jailbreaking and Red Teaming

Jailbreaking[234][232][229][82][235][230][231][233][199][236] refers to the practice of bypassing the built-in safety and ethical guardrails of LLMs. Malicious users or adversaries can exploit vulnerabilities in AI models through crafted inputs or backdoors, inducing them to produce inappropriate responses, such as giving instructions for illegal activities or generating offensive content[234][232][229][82][230][231][236]. These activities pose risks by compromising safety measures of AI models, making it challenging to ensure the model to behave appropriately, especially in real-world scenarios where the potential inputs are unpredictable.

Red-teaming is a common practice that simulates attacks or misuse scenarios to identify vulnerabilities in AI models, thereby mitigating unsafe behaviors[201][204][203][200][202][239][240]. It challenges AI models with adversarial inputs, bias exploitation methods, and system manipulation to uncover potential risks, *e.g.*, hidden biases or failure modes, that may not be evident during normal usage. The outcomes from red-teaming can be leveraged to refine AI models, reducing undesired behaviors and inappropriate outputs. This approach is effective in stress-testing AI models and has been widely employed in industry[92][202][201].

As AI systems grow sophisticated, red-teaming must continually adapt to counter increasingly complex attacks, calling for more advanced tools and interdisciplinary expertise[297]. Also, current red-teaming practices are often static, typically conducted after model pretraining with one-time attacks. Thus, these practices potentially fail to address evolving user needs and complex interactions between users and models, especially when dealing with customized AI services for different user groups. To address these problems, model owners can conduct diverse red teaming regularly to anticipate and mitigate risks in evolving use cases. Red teaming should also involve more diverse use cases, such as employing different user groups and simulating complex interactions with dynamic contexts, to address safety risks in customized AI services and enable a comprehensive evaluation of potential vulnerabilities across different scenarios.

5.2. Content Moderation and Customized Moderation Pipeline

Content moderation employs rule-based methods, machine learning classifiers, and human oversight to detect and review user inputs and AI-generated content to ensure compliance with safety, ethical, and regulatory standards[55][251][252][253]. These approaches identify inappropriate content, including toxicity, bias, hallucinations, private information, and jailbreaks, etc[107]. By preventing AI models from engaging with or producing harmful content, content moderation plays a critical role in maintaining responsible AI behavior.

Customized moderation pipelines[258][259] employ multiple components at different stages of LLM inference to enhance the quality of the model outputs while enhancing the control of the whole AI workflows. They function as a flexible intermediate layer between users and AI models, enabling users to add customized components, such as code-based rules and quality guarantees, at any stage of inference based on their personalized needs.

As we reflect on practices for AI moderation and customized moderation pipelines, several key challenges and considerations have emerged. 1) ***Hallucination*** remains a significant issue and is unlikely to solve completely, regardless of how AI evolves. That is because hallucinations stem from nearly every stage of the LLM lifecycle, including training data quality, model architecture, and inference processes, etc[298]. Despite advancements in AI technology, hallucinations persist as an inherent challenge that requires constant attention in moderation practices. 2) The ***balance between utility and safety*** is crucial and may require different solutions according to the changing contexts and user needs. Overly strict moderation can cause AI models to be uninformative, e.g., an LLM that always responds "sorry, I cannot answer this question" is completely safe but useless. Defining appropriate thresholds for acceptable responses is challenging and requires considering potential risks and the context of user queries. 3) ***Customization*** of moderation methods should be applied for different use cases, as certain information may be safe for one user group but inappropriate for another. As an example, a chemistry industry practitioners may need specific and technical information on dangerous substances like explosives for professional purposes, while such information should be restricted for the general public. Current moderation practices might fail to include such cases, and applying identical restrictions between different user expertise or context might limit the utility of AI in specialized fields. 4) Some moderation measures, such as detection of *toxicity*, *bias*, *stereotypes*, *out-of-distribution content*, might be redundant for most real-world use cases. This is because most of the

current AI applications and multi-agent systems[168][299] leverage existing LLM APIs[167][166][55], open-source LLMs[300][301][302], or utilize open-source LLMs as base models for fine-tuning, many of which already have built-in safety components or have done safety alignments for handling inappropriate content. Incorporating certain detection mechanisms may offer limited benefits to the safety of model inputs and outputs while increasing unnecessary computational overhead.

5.3. Privacy

Privacy leakage in AI models refers to unintended exposure of sensitive information that the models may have learned during training. Adversaries can exploit model weights or gradients to infer sensitive information in the training data, e.g., data reconstruction attacks, membership inference attacks, etc[263][264][265]. Additionally, AI models are often trained on massive datasets that may contain sensitive data, such as personally identifiable information (PII). As a result, these models might memorize specific details of individuals, which might be reproduced inadvertently during interactions with users. Malicious users can also craft adversarial prompts to extract sensitive information from AI models with jailbreaks[220][227].

Addressing privacy leakage involves methods such as differential privacy[303][304][305][306], federated learning[307][165][308][309][310], and privacy-preserving cryptographic protocols (e.g., homomorphic encryption[311][312][313]) for training AI models, ensuring that sensitive data cannot be extracted from weights or model outputs. Researchers also leverage differential privacy in in-context learning and finetuning[314][315][316], or design privacy-preserving prompts for querying LLMs[219][317], to maintain data confidentiality during interactions between users and AI models. Other practices for ensuring data privacy include conducting rigorous data filtering before training, monitoring model inputs and outputs (see moderation in §5.2), etc, to ensure that the model outputs would not reveal sensitive information[55][251][252][253].

When reflecting on the practices of privacy, a crucial aspect is the ***balance between privacy and utility***. While ensuring data privacy is essential, it often comes at the expense of functionality, utility, and user experiences. In practice, implementing privacy measures in AI applications may fail to prevent malicious users while inadvertently degrading the experiences of benign users. Moreover, with advancements in technologies, breaching privacy becomes easier, while ensuring robust privacy protections is growing more complex and expensive. This highlights a deeper issue: *those intent on*

exploiting vulnerabilities of AI models can often find new methods to bypass safeguards, whereas being overly cautious may impact benign users negatively.

5.4. AI Security and Defense Methods

Defense methods for ensuring AI security typically involve the following strategies: (1) aligning models with human values and ethical norms through Supervised Fine-Tuning (SFT)[318] and Reinforcement Learning from Human Feedback (RLHF)[319] to ensure the models to follow user instructions safely and responsibly; (2) leveraging built-in safeguards for content generation, e.g., exploring the decoding stage[271][2], implementing reward-based mechanisms[4][284], and leveraging hidden states[275][2], to prevent the generation of unsafe content; (3) constructing datasets with safety-enhancing features, e.g., secure instructions and adversarial samples, and training models with such data to enhance their robustness against malicious inputs[274][276][55][320][279]; (4) employing prompt engineering methods, combined with chain-of-thought reasoning[321] or multi-agent systems[168], to help models understand user inputs better and react appropriately[272][273][277][278]; (5) training models and/or incorporating moderation pipelines to detect unsafe content and prevent generating inappropriate outputs (see §5.2); (6) employing certified robustness approaches that provide formal guarantees on the model's robustness against adversarial perturbations[280][281]; and (7) utilizing differential privacy[322][323][282], homomorphic encryption[324][313], and secure MPC[325][326], to ensure that sensitive data are protected during model training, etc.

5.5. AI Serving Security

AI serving security involves secure deployment, operation, and maintenance of AI systems when providing services to users[327][199]. It focuses on protecting models and interactions between users and models from various threats, such as supply chain vulnerabilities[328], model theft[329], model denial of service[330], and insecure plugins[331].

Supply chain vulnerabilities exist in software components, models, and training data provided by third party providers or user prompts collected with the supply chain[328]. Such vulnerabilities include using outdated models, deprecated third-party packages, improper handling of training data, etc, leading to data breaches, malicious injections, or even system failures[328]. Hackers may inject poisoned data through supply chain, introducing backdoors and biases during pretraining or fine-

tuning[332]. Malicious users may craft prompts to bypass system controls, inducing unauthorized actions or unsafe model outputs[333].

Model theft involves unauthorized access, replication, or reverse engineering of model parameters to create functionally equivalent copies[329], which poses risks to intellectual property, brand reputation, and financial security.

Distributed denial of service (DDoS) attacks compromise AI services, especially when models are integrated into widely accessible applications[330]. Attackers can craft prompts that exploit recursive behaviors in AI models, causing excessive computational resource consumption[334]. For example, an abnormal traffic pattern of a DDoS attack led to sporadic outages for OpenAI, affecting ChatGPT and developer tools for hours in November 2023[335].

Plugins enhance model functionality by enabling interactions with external software, databases, web tools, or APIs[331]. However, they may introduce vulnerabilities when execution controls are inadequate. Adversaries may exploit insecure plugins with adversarial prompts to perform unauthorized actions, such as data exfiltration, remote code execution, and privilege escalation[331][336]. Furthermore, without isolated environments, unauthorized access to plugins might even allow modifying system-level resources, which, in the worst cases, might cause the serving system to crash[336][337].

5.6. AI Governance

AI governance is a framework of rules, practices, policies, and tools that ensure AI systems are built, developed, and used in a safe and responsible manner, aligning with social values and fulfilling legal standards[223][221]. [53] defines 13 AI governance components in a continuous loop in AI life-cycle, including: 1) establishing accountability for AI; 2) assessing regulatory risks; 3) gathering inventory of use cases; 4) increasing values of underlying data; 5) assessing fairness and accessibility; 6) improving reliability and safety; 7) heightening transparency and explainability; 8) implementing accountability with human-in-the-loop; 9) supporting privacy and retention; 10) improving security; 11) implementing AI model lifecycle and registry; 12) managing risk; and 13) realizing AI value.

Among these directions, items 1) and 8) are related to accountability; items 2), 6), 9), 10), and 12) are related to red teaming (see §5.1), content moderation and customized moderation pipeline (see §5.2), AI privacy (see §5.3), AI security and defense methods (see §5.4), and AI serving security (see §5.5);

item 3) is related to business and is out of the scope of this subsection (see[53] for details on practical AI governance use cases); item 4) is related to data management such as training data processing[55], access control[288][338][339][340], and data regulations[341][342]; item 5) is related to moderation (see §5.2) and access control[288][343]; item 7) is related to AI transparency; and items 11) and 13) are related to customizing AI workflow to enhance functionality and safety (see §5.2). Without loss of generality, this subsection discusses topics that are not mentioned in the previous subsections, including i) accountability, ii) regulations, iii) access control, and iv) transparency, interpretability and explainability.

Accountability. Accountability refers to being responsible for actions and impacts of AI systems on individuals and society[294][344]. Accountability can be defined as a relation of *answerability* with three conditions: authority recognition, interrogation, and limitation of power[345]. The accountability framework can be summarized with seven features (context, range, agent, forum, standards, process, and implications) and four key goals (compliance, reporting, oversight, and enforcement)[345]. These goals are often complementary, while policy-makers tend to focus on some goals over others, depending on specific objectives of AI governance.

Accountability can be explored at different levels within the AI lifecycle, including data, model, and developers[346]. Typical approaches for enhancing accountability include algorithmic assessments, auditing, and data provenance techniques[346]. Proactive approaches and reactive approaches can be utilized to set standards and addressing issues after they occur[345]. Watermarking methods further support accountability by embedding traceable identifiers in AI outputs, allowing responsible tracking of content in cases with ethical or regulatory implications[286][287][291].

Regulations. AI regulation refers to laws, policies, and guidelines established to govern development, deployment, and usage of AI. For example, European Union's AI Act[347] categorizes AI applications by their risk levels, including unacceptable risk, high risk, general-purpose AI, limited risk, and minimal risk; the General Data Protection Regulation (GDPR)[348] enforces strict data protection and privacy requirements; and the Blueprint for an AI Bill of Rights[349] and NIST AI Risk Management Framework[350] provide guiding principles to foster responsible AI usage, emphasizing human oversight, transparency, and accountability.

Access control. Access control in AI systems manages user access to data and models according to their permissions, thereby limiting unauthorized model access and data exposure. Effective methods

include data access controls[343] and model access controls[288]. As an example, compositional fine-tuning enables each information silo (e.g., databases or documents) has its own fine-tuned model, allowing models to operate securely across multiple data silos and ensuring users access only to the fine-tunings they are authorized for[288].

Interpretability, explainability, and transparency. Interpretability, explainability, and transparency are interrelated concepts but differ in their scopes. Interpretability focuses on understanding how AI models function internally and how they produce outputs[46][351][352][41]. Explainability provides explanations to specific model outputs, presenting the results in ways that can be easily understood and trusted by end-users[1][46][351][41][353]. Transparency is the broadest concept that encompasses both interpretability and explainability and involves insights into the entire AI systems, such as how models make decisions, what data are used, and why specific results are produced[354][355][351][41]. Transparency also extends to concepts like traceability and content provenance that track models and data to ensure accountability and reliability[53][295][296]. In practice, these concepts are often considered together to provide a more holistic understanding of AI systems[222][290][351][41][356].

Practices for enhancing transparency include a variety of methods, such as i) evaluating the contribution of input elements (e.g., data features, user-defined concepts, or specific regions in images) to model outputs with methods such like removal-based explanations[357][358][46][358][359][360][361], counterfactual explanations[362][363], concept activation explanations[364][46], and saliency maps (for visual data[365]); ii) X-of-Thought approaches that break down complex tasks into structured reasoning steps to make the reasoning process more interpretable to users, e.g., chain-of-thought[366][367][368][369], tree-of-thought[370][371][372], graph-of-thought[373][374], and their variants; iii) embedding watermarks to enhance traceability and content provenance for AI-generated outputs[286][287][295][296]; iv) leveraging external knowledge sources, tools, or methods as reference for reasoning, e.g., retrieval-augmented generation (RAG)[375][376], function calling[377][378][379], and web browsing[380]; v) leveraging interpretability algorithms or models as auxiliary tools to generate insights into AI outputs while automating AI interpretability[381][382][383].

6. AI Safety in the 2020s: Challenges and Missions

This section utilizes the blueprint to guide our efforts at the current stage. As AI safety practitioners in the 2020s, we stand at a pivotal moment in history, with society moving toward an era where AI will

be deeply integrated into daily life. While addressing immediate safety concerns, our primary mission is to anticipate and mitigate risks within the expanding AI landscape. Thus, certain directions may require attention.

A shift from AI safety to AI quality assurance. The rapid changing AI landscape requires a shift of our focus from narrowly defined AI safety concerns to a more comprehensive concept, *i.e., AI quality assurance*. Besides focusing on safety and security, practitioners can put more efforts on addressing alignment[319] as well as the quality of interactions between users and AI. Viewing AI safety as part of AI quality assurance helps enhance AI in a way that is safe, accurate, and capable of meeting the complex requirements of real-world AI applications. The terminology "AI quality assurance" also guides us to think more about what we can do to align our current efforts with the blueprint picture.

Enhancing alignment with more interdisciplinary insights. Aligning AI with deeper insights across different fields, people, and cultures is important for enhancing the services provided by AI models. Current AI models, while advanced, often lack a deep understanding of ethical, social, and cultural contexts, while most of the existing work on alignment tend to be conducted by people in the computer science field. However, more advanced AI systems necessitate interdisciplinary collaborations between AI researchers and experts from diverse fields such as psychology, sociology, history, art, and anthropology, especially for AI applications that involve frequent interactions with different user groups, such as domains like finance[14][165], healthcare[15][164], education[166][167], etc. AI safety practitioners might expand their focus beyond technical aspects and foster interdisciplinary collaborations, such that AI systems can be technically robust while being socially and ethically aligned with diverse user backgrounds.

Hallucination. Hallucination remains a fundamental and persistent challenge and is difficult to eliminate entirely, regardless of advancements in AI[298]. This is because hallucinations can originate at nearly every stage of AI lifecycle. Factors such as the quality and representativeness of training data, inappropriate user inputs, and inference processes all contribute to the occurrence hallucinations. Despite implementing RAG[376] helps mitigate hallucination, it remains an inherent risk of AI models. Thus, implementing a comprehensive moderation pipeline for identifying and addressing hallucinations is crucial, which involves grounding user queries with RAG, detecting presence of hallucinations, and fixing hallucinations in the model outputs if possible.

Customization. Customizing AI workflows is essential to ensure that AI systems address diverse user needs effectively and align with specific cultural norms, legal regulations, and ethical considerations. Current AI systems such as ChatGPT[55] tend to provide similar or identical services to all user groups, regardless of individual differences. However, as AI becomes increasingly integrated into daily life, one-size-fits-all solutions are inadequate. With diverse backgrounds and needs, users might expect AI systems to adapt to their individual differences and preferences, and deliver safe, effective, and personalized services. As an example, AI models interacting with children should respond differently compared with interacting with adults, necessitating considering user age, learning objectives, cognitive abilities, emotional maturity, etc. In domains such like finance[14][165], healthcare[15][164], and education[166][167], customizing AI services based on user profiles is crucial for enhancing safety while improving user experiences.

Interaction. Our long-term vision for the blueprint future suggests that the *Internet of Everything* involves complex interactions between intelligent robots and human users (similar ideas also discussed in[42][47]), which calls for a shift in our focus from static, one-shot analyses to dynamic, context-aware interactions between users and models to improve alignment. Beyond prompt-based AI systems and services, we can develop more sophisticated, context-aware interfaces for human-AI interactions to support multi-round interactions that adapt to evolving conversational contexts. Also, real-time safety protocols are in need to adjust to changing contexts and user needs. It is also essential to develop methodologies for safety assessment that functions effectively in complex and rapidly evolving settings, such as those involving interactions between multiple AI agents and human users.

7. Conclusion

This paper presents a blueprint for an advanced human society and leverages this forward-looking vision to guide today's AI safety efforts. Through the blueprint, it becomes clear that Artificial General Intelligence (AGI) is not the ultimate goal of AI development. Instead, the true vision lies in the *Internet of Everything*, a deeply interconnected world where intelligent systems seamlessly integrate into daily life. While AGI is a popular topic discussed more among people in the computer science field today, a more advanced world demands interdisciplinary collaborations across various fields. Regarding AGI as the ultimate goal might limit our creativity in the 2020s, trapping us in "local optima" and potentially causing us to overlook the real challenges.

What will this future world be like? Will it bring more happiness to human-beings? The answer is uncertain. Advanced technologies may take on more duties of daily life, leaving humans with fewer tasks and more predictability. On the other hand, with more daily problems being solved by technologies, people may have more time for self-reflection and personal growth, which may lead to another form of “happiness”.

As we stand at a pivotal moment in history, our efforts become more than an incremental technological endeavor, but a profound exploration of human potential and the boundaries of human capability. In any case, the journey toward this advanced future is fascinating, and every human being will look forward to it.

Notes

The idea for this article struck the solo author unexpectedly on an ordinary afternoon as she moved into a garage in Palo Alto during the summer of 2024.

References

1. a, b, c, d, e, f *Chua J, Li Y, Yang S, Wang C, Yao L (2024). "AI Safety in Generative AI Large Language Models: A Survey". arXiv preprint arXiv:2407.18369. Available from: https://arxiv.org/abs/2407.18369.*
2. a, b, c, d *Du Y, Zhao S, Zhao D, Ma M, Chen Y, Huo L, Yang Q, Xu D, Qin B (2024). "MoGU: A Framework for Enhancing Safety of Open-Sourced LLMs While Preserving Their Usability". arXiv preprint arXiv:2405.14488.*
3. a, b, c *Hubinger E, Denison C, Mu J, Lambert M, Tong M, MacDiarmid M, Lanham T, Ziegler DM, Maxwell T, Cheng N, et al. Sleeper agents: Training deceptive llms that persist through safety training. arXiv preprint arXiv:2401.05566. 2024.*
4. a, b, c *Mu T, Helyar A, Heidecke J, Achiam J, Vallone A, Kivlichan I, Lin M, Beutel A, Schulman J, Weng L. "Rule Based Rewards for Language Model Safety".*
5. ^ *Peng SY, Chen PY, Hull M, Chau DH (2024). "Navigating the Safety Landscape: Measuring Risks in Finetuning Large Language Models". arXiv preprint arXiv:2405.17374. arXiv:2405.17374.*
6. a, b, c *Tedeschi S, Friedrich F, Schramowski P, Kersting K, Navigli R, Nguyen H, Li B (2024). "ALERT: A Comprehensive Benchmark for Assessing Large Language Models' Safety through Red Teaming". arXiv preprint arXiv:2404.08676. Available from: https://arxiv.org/abs/2404.08676.*

7. ^Neal RM (2006). "Pattern Recognition and Machine Learning". Technometrics. 49: 366 – 366. S2CID 31993898.
8. ^Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, Compas C, Martin C, Costa AB, Flores MG, et al. (2022). "A large language model for electronic health records". NPJ digital medicine. 5 (1): 194.
9. ^Xu FF, Alon U, Neubig G, Hellendoorn VJ (2022). "A systematic evaluation of large language models of code". In: Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming. pp. 1–10.
10. ^Huzhang G, Pang ZJ, Gao Y, Liu Y, Shen W, Zhou WJ, Lin Q, Da Q, Zeng AX, Yu H, et al. (2021). "AliExpress Learning-To-Rank: Maximizing online model performance without going online". IEEE Transactions on Knowledge and Data Engineering. 35 (2): 1214–1226.
11. ^Duong D, Solomon BD (2024). "Analysis of large-language model versus human performance for genetics questions". European Journal of Human Genetics. 32 (4): 466–468.
12. ^Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, Löffler CML, Schwarzkopf SC, Unger M, Veldhuizen GP, et al. The future landscape of large language models in medicine. Communications medicine. 3(1):141, 2023.
13. a, b, cZhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, Min Y, Zhang B, Zhang J, Dong Z, et al. (2023). "A survey of large language models". arXiv preprint arXiv:2303.18223.
14. a, b, c, d, e, fWu S, Irsoy O, Lu S, Dabravolski V, Dredze M, Gehrmann S, Kambadur P, Rosenberg D, Mann G (2023). "Bloomberggpt: A large language model for finance". arXiv preprint arXiv:2303.17564. Available from: https://arxiv.org/abs/2303.17564.
15. a, b, c, d, e, fThirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW (2023). "Large language models in medicine". Nature medicine. 29 (8): 1930–1940.
16. a, bKambhampati S (2024). "In vs. Out of Distribution analyses are not that useful for understanding LLM reasoning capabilities". Available from: https://x.com/rao2z/status/1827778006972248123.
17. a, bValmeekam K, Stechly K, Kambhampati S (2024). "LLMs Still Can't Plan; Can LRMs? A Preliminary Evaluation of OpenAI's o1 on PlanBench". arXiv preprint arXiv:2409.13373.
18. ^Ahmad T, Zhang D, Huang C, Zhang H, Dai N, Song Y, Chen H (2021). "Artificial intelligence in sustainable energy industry: Status Quo, challenges and opportunities". Journal of Cleaner Production. 289: 125834.

19. a, b Groumpos PP (2021). "A critical historical and scientific overview of all industrial revolutions". IFAC-PapersOnLine. **54** (13): 464–471.
20. ^ Argerich MF, Patiño-Martínez M (2024). "Measuring and Improving the Energy Efficiency of Large Language Models Inference". IEEE Access. **2024**.
21. ^ Cohen A (2024). "AI Is Pushing The World Toward An Energy Crisis". https://www.forbes.com/sites/arielcohen/2024/05/23/ai-is-pushing-the-world-towards-an-energy-crisis/.
22. ^ Patterson D, Gonzalez J, Le Q, Liang C, Munguia L-M, Rothchild D, So D, Texier M, Dean J (2021). "Carbon emissions and large neural network training". arXiv preprint arXiv:2104.10350. 2021.
23. ^ Moazeni S (2023). "Q&A: UW researcher discusses just how much energy ChatGPT uses". https://www.newswise.com/articles/qa-uw-researcher-discusses-just-how-much-energy-chatgpt-uses.
24. ^ Girvan P. "The Hidden Environmental Costs of Chat GPTs: A Conversation with Bing Copilot". 2024. Available from: https://bylines.scot/environment/environmental-costs-of-chat-gtps/.
25. ^ Singh S (2024). "ChatGPT Statistics (SEP. 2024) ◙ 200 Million Active Users". Available from: https://www.demandsage.com/chatgpt-statistics/.
26. a, b, c, d Einstein A. "The general theory of relativity." In: The meaning of relativity. Springer; 1922. p. 54–75.
27. a, b, c, d Einstein A, Lorentz HA, Weyl H. The principle of relativity. Courier Corporation; 2013.
28. a, b, c, d Einstein A, Podolsky B, Rosen N (1935). "Can quantum-mechanical description of physical reality be considered complete?" Physical review. **47** (10): 777.
29. a, b, c, d Zettili N. Quantum mechanics: concepts and applications. 2009. John Wiley & Sons.
30. a, b, c, d, e, f Hsu YC, Chang RC. "Intelligent chips and technologies for AIoT era." In: 2020 IEEE Asian Solid-State Circuits Conference (A-SSCC). IEEE; 2020. p. 1-4.
31. a, b, c, d, e, f Li EP, Ma H, Ahmed M, Tao T, Gu Z, Chen M, Chen Q, Li D, Chen W (2023). "An electromagnetic perspective of artificial intelligence neuromorphic chips". Electromagnetic Science. **1** (3): 1–18.
32. a, b, c, d, e, f He B, Yuan H, Meng J, Gao S (2020). "Brain--computer interfaces". Neural engineering. Springer. pages 131--183.
33. a, b, c, d, e, f Nicolas-Alonso LF, Gomez-Gil J (2012). "Brain computer interfaces, a review". Sensors. **12** (2): 1211–1279.
34. a, b, c, d, e, f Wolpaw JR. Brain--computer interfaces. In: Handbook of clinical neurology. Elsevier; 2013. p. 67--74.

35. a, b Ayyamperumal SG, Ge L (2024). "Current state of LLM Risks and AI Guardrails". arXiv preprint arXiv:2406.12934.

36. a, b Liu FW, Hu C (2024). "Exploring Vulnerabilities and Protections in Large Language Models: A Survey". arXiv preprint arXiv:2406.00240. Available from: https://arxiv.org/abs/2406.00240.

37. a, b Rahman MM, Arshi AS, Hasan MGM, Mishu SF, Shahriar H, Wu F (2023). "Security Risk and Attacks in AI: A Survey of Security and Privacy". 2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC). pages 1834-1839.

38. a, b He Y, Wang E, Rong Y, Cheng Z, Chen H (2024). "Security of AI Agents". arXiv preprint arXiv:2406.08689.

39. a, b, c Yao Y, Duan J, Xu K, Cai Y, Sun Z, Zhang Y (2024). "A survey on large language model (llm) security and privacy: The good, the bad, and the ugly". High-Confidence Computing. Page 100211.

40. a, b Mozes M, He X, Kleinberg B, Griffin LD (2023). "Use of llms for illicit purposes: Threats, prevention measures, and vulnerabilities". arXiv preprint arXiv:2308.12833.

41. a, b, c, d, e Qi X, Huang Y, Zeng Y, Debenedetti E, Geiping J, He L, Huang K, Madhushani U, Sehwag V, Shi W, et al. AI Risk Management Should Incorporate Both Safety and Security. arXiv preprint arXiv:2405.19524. 2024.

42. a, b, c Bajcsy A, Fisac JF (2024). "Human-AI Safety: A Descendant of Generative AI and Control Systems Safety". arXiv preprint arXiv:2405.09794. arXiv:2405.09794.

43. ^ Bengio Y, Hinton G, Yao A, Song D, Abbeel P, Darrell T, Harari YN, Zhang YQ, Xue L, Shalev-Shwartz S, et al. Managing extreme AI risks amid rapid progress. Science. 384(6698):842–845, 2024.

44. a, b Dalrymple D, Skalse J, Bengio Y, Russell S, Tegmark M, Seshia S, Omohundro S, Szegedy C, Goldhaber B, Ammann N, et al. Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems. arXiv preprint arXiv:2405.06624. 2024.

45. a, b, c, d Kenthapadi K, Sameki M, Taly A (2024). "Grounding and Evaluation for Large Language Models: Practical Challenges and Lessons Learned (Survey)". In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 6523–6533.

46. a, b, c, d, e, f, g, h Liu Y, Yao Y, Ton JF, Zhang X, Guo R, Cheng H, Klochkov Y, Taufiq MF, Li H (2023). "Trustworthy llms: a survey and guideline for evaluating large language models' alignment". arXiv preprint arXiv:2308.05374. arXiv:2308.05374.

47. a, b, c Gabriel I, Manzini A, Keeling G, Hendricks LA, Rieser V, Iqbal H, Toma\v{s}ev N, Ktena I, Kenton Z, Rodriguez M, et al. The ethics of advanced ai assistants. arXiv preprint arXiv:2404.16244. 2024.

48. ^Weidinger L, Barnhart J, Brennan J, Butterfield C, Young S, Hawkins W, Hendricks LA, Comanescu R, Chang O, Rodriguez M, et al. Holistic safety and responsibility evaluations of advanced AI models. arXiv preprint arXiv:2404.14068. 2024.

49. a, bOpenAI (2022). "Lessons learned on language model safety and misuse." Available from: https://openai.com/index/language-model-safety-and-misuse/.

50. a, bOpenAI (2022). "Reimagining secure infrastructure for advanced AI". https://openai.com/index/reimagining-secure-infrastructure-for-advanced-ai/.

51. a, bOpenAI (2023). Preparedness Framework. Available from: https://cdn.openai.com/openai-preparedness-framework-beta.pdf.

52. a, bOpenAI (2024). "OpenAI safety update". Available from: https://openai.com/index/openai-safety-update/.

53. a, b, c, d, eSoares S. AI Governance Comprehensive: Tools, Vendors, Controls and Regulations. 2024. Available from: https://yourdataconnect.com/wp-content/uploads/2024/08/AI_Governance_Comprehensive_Sept_2024.pdf.

54. ^Qi X, Zeng Y, Xie T, Chen P-Y, Jia R, Mittal P, Henderson P. "Fine-tuning aligned language models compromises safety, even when users do not intend to!" In: The Twelfth International Conference on Learning Representations; 2024. Available from: https://openreview.net/forum?id=hTEGyKfodZ.

55. a, b, c, d, e, f, g, h, iMarkov T, Zhang C, Agarwal S, Eloundou Nekoul F, Lee T, Adler S, Jiang A, Weng L (2023). "A holistic approach to undesired content detection in the real world". Proceedings of the AAAI Conference on Artificial Intelligence. 37 (12): 15009–15018.

56. ^**JIGSAW** (2019). "Jigsaw Unintended Bias in Toxicity Classification Dataset". Jigsaw Unintended Bias in Toxicity Classification Dataset. Available from: https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification.

57. ^**JIGSAW** (2018). "Jigsaw Toxic Comment Classification Dataset". Jigsaw toxic comment classification dataset. Available from: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data.

58. ^**JIGSAW** (2020). "Jigsaw Multilingual Toxic Comment Classification Dataset". Article. Available from: https://www.kaggle.com/c/jigsaw-multilingual-toxic-comment-classification/data.

59. ^Zi Lin, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang, Jingbo Shang (2023). "ToxicChat: Unveiling Hidden Challenges of Toxicity Detection in Real-World User-AI Conversation". arXiv. arXiv:2310.17389 [cs.CL].

60. ^ Hartvigsen T, Gabriel S, Palangi H, Sap M, Ray D, Kamar E (2022). "Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection". arXiv preprint arXiv:2203.09509. arXiv:2203.09509.

61. ^ Zheng J, Liu X, Haque M, Qian X, Yang G, Yang W (2024). "HateModerate: Testing Hate Speech Detectors against Content Moderation Policies". In: Findings of the Association for Computational Linguistics: NAACL 2024. 2024: 2691–2710.

62. ^ Gehman S, Gururangan S, Sap M, Choi Y, Smith NA (2020). "Realtoxicityprompts: Evaluating neural toxic degeneration in language models". arXiv preprint arXiv:2009.11462.

63. ^ Lin S, Hilton J, Evans O (2021). "Truthfulqa: Measuring how models mimic human falsehoods". arXiv preprint arXiv:2109.07958.

64. ^ Mallen A, Asai A, Zhong V, Das R, Hajishirzi H, Khashabi D (2022). "When Not to Trust Language Models: Investigating Effectiveness and Limitations of Parametric and Non-Parametric Memories". arXiv preprint. 2022.

65. ^ Joshi M, Choi E, Weld DS, Zettlemoyer L (2017). "TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension". ArXiv. **abs/1705.03551**. S2CID 26501419.

66. ^ Lee K, Chang MW, Toutanova K. "Latent retrieval for weakly supervised open domain question answering." In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019 Jul; Florence, Italy. Association for Computational Linguistics. p. 6086-6096. doi:10.18653/v1/P19-1612. https://www.aclweb.org/anthology/P19-1612.

67. ^ Kwiatkowski T, Palomaki J, Redfield O, Collins M, Parikh A, Alberti C, Epstein D, Polosukhin I, Devlin J, Lee K, Toutanova K, Jones L, Kelcey M, Chang MW, Dai AM, Uszkoreit J, Le Q, Petrov S (2019). "Natural Questions: A Benchmark for Question Answering Research". Transactions of the Association for Computational Linguistics. 7: 453-466. doi:10.1162/tacl_a_00276.

68. ^ Thorne J, Vlachos A, Christodoulopoulos C, Mittal A. "FEVER: a Large-scale Dataset for Fact Extraction and VERification". In: NAACL-HLT; 2018.

69. ^ Wang W, Haddow B, Birch A, Peng W (2023). "Assessing the reliability of large language model knowledge". arXiv preprint arXiv:2310.09820. arXiv:2310.09820.

70. a, b, c Li J, Cheng X, Zhao WX, Nie JY, Wen JR. Halueval: A large-scale hallucination evaluation benchmark for large language models. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023:6449–6464.

71. ^Yang Z, Qi P, Zhang S, Bengio Y, Cohen WW, Salakhutdinov R, Manning CD. "HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering." In: Conference on Empirical Methods in Natural Language Processing (EMNLP); 2018.

72. a, bNiu C, Wu Y, Zhu J, Xu S, Shum K, Zhong R, Song J, Zhang T (2023). "Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models". arXiv preprint arXiv:2401.00396.

73. ^Hermann KM, Kocisky T, Grefenstette E, Espeholt L, Kay W, Suleyman M, Blunsom P (2015). "Teaching machines to read and comprehend". Advances in neural information processing systems. 28.

74. ^See A, Liu PJ, Manning CD. Get to the point: Summarization with pointer-generator networks. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada: Association for Computational Linguistics; 2017. p. 1073-1083. doi:10.18653/v1/P17-1099. https://www.aclweb.org/anthology/P17-1099.

75. ^Narayan S, Cohen SB, Lapata M (2018). "Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization". ArXiv. **abs/1808.08745**. Available from: https://arxiv.org/abs/1808.08745.

76. ^Dziri N, Kamalloo E, Milton S, Zaiane O, Yu M, Ponti E, Reddy S (2022). "FaithDial: A Faithful Benchmark for Information-Seeking Dialogue". arXiv preprint, arXiv:2204.10757. Available from: https://arxiv.org/abs/2204.10757.

77. ^Luo W, Shen T, Li W, Peng G, Xuan R, Wang H, Yang X (2024). "HalluDial: A Large-Scale Benchmark for Automatic Dialogue-Level Hallucination Evaluation". arXiv e-prints. pages arXiv--2406.

78. ^Lai G, Xie Q, Liu H, Yang Y, Hovy E. "RACE: Large-scale Reading Comprehension Dataset From Examinations." In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics; 2017. p. 785-794. doi:10.18653/v1/D17-1082. Available from: https://aclanthology.org/D17-1082.

79. ^Rajpurkar P, Jia R, Liang P. "Know What You Don{'}t Know: Unanswerable Questions for {SQ}u{AD}." In: Gurevych I, Miyao Y, editors. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Melbourne, Australia: Association for Computational Linguistics; 2018. p. 784-789. doi:10.18653/v1/P18-2124. Available from: https://aclanthology.org/P18-2124.

80. ^Rajpurkar P, Zhang J, Lopyrev K, Liang P. "SQuAD: 100,000+ Questions for Machine Comprehension of Text." In: Su J, Duh K, Carreras X, editors. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas: Association for Computational Linguistics; 2016. p. 2383-2392. doi:10.18653/v1/D16-1264. Available from: https://aclanthology.org/D16-1264.

81. ^Longpre S, Perisetla K, Chen A, Ramesh N, DuBois C, Singh S. Entity-based knowledge conflicts in question answering. In: Moens MF, Huang X, Specia L, Yih SW, editors. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics; 2021. p. 7052-7063. doi:10.18653/v1/2021.emnlp-main.565. Available from: https://aclanthology.org/2021.emnlp-main.565/.

82. a, b, c, dShen X, Chen ZJ, Backes M, Shen Y, Zhang Y (2023). "Do Anything Now: Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models". ArXiv. **abs/2308.03825**. S2CID 260704242.

83. ^Qiu H, Zhang S, Li A, He H, Lan Z (2023). "Latent Jailbreak: A Benchmark for Evaluating Text Safety and Output Robustness of Large Language Models". ArXiv. **abs/2307.08487**. S2CID 259937347.

84. ^Rube9n Daredo Jaramillo Romero (2023). "ChatGPT Jailbreak Prompts Dataset". https://huggingface.co/datasets/rubend18/ChatGPT-Jailbreak-Prompts.

85. ^Hao J (2023). "Jailbreak Classification Dataset". Hugging Face Datasets. Available from: https://huggingface.co/datasets/jackhhao/jailbreak-classification.

86. a, b, cZou A, Wang Z, Kolter JZ, Fredrikson M (2023). "Universal and transferable adversarial attacks on aligned language models". ArXiv. abs/2307.15043.

87. ^Liu F, Feng Y, Xu Z, Su L, Ma X, Yin D, Liu H (2024). "JAILJUDGE: A Comprehensive Jailbreak Judge Benchmark with Multi-Agent Enhanced Explanation Evaluation Framework". arXiv. arXiv:2410.12855 [cs.CL].

88. ^Qiu H, Zhang S, Li A, He H, Lan Z (2023). "Latent jailbreak: A benchmark for evaluating text safety and output robustness of large language models". arXiv preprint arXiv:2307.08487. arXiv:2307.08487.

89. ^Ren Q, Li H, Liu D, Xie Z, Lu X, Qiao Y, Sha L, Yan J, Ma L, Shao J (2024). "Derail Yourself: Multi-turn LLM Jailbreak Attack through Self-discovered Clues". arXiv. arXiv:2410.10700 [cs.CL].

90. a, bMazeika M, Phan L, Yin X, Zou A, Wang Z, Mu N, Sakhaee E, Li N, Basart S, Li B, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. arXiv preprint arXiv:2402.04249. 2024.

91. ^Yoo H, Yang Y, Lee H (2024). "CSRT: Evaluation and Analysis of LLMs using Code-Switching Red-Teaming Dataset". arXiv e-prints. pages arXiv--2406.

92. a, b, cBhatt M, Chennabasappa S, Nikolaidis C, Wan S, Evtimov I, Gabi D, Song D, Ahmad F, Aschermann C, Fontana L, et al. (2023). "Purple llama cyberseceval: A secure coding benchmark for language models". arXiv preprint arXiv:2312.04724.

93. ^Rudinger R, Naradowsky J, Leonard B, Van Durme B. "Gender bias in coreference resolution." In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2018 Jun; New Orleans, Louisiana. Association for Computational Linguistics.

94. a, bNadeem M, Bethke A, Reddy S. "StereoSet: Measuring stereotypical bias in pretrained language models." In: Zong C, Xia F, Li W, Navigli R, editors. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics; 2021. p. 5356–5371. doi:10.18653/v1/2021.acl-long.416. Available from: https://aclanthology.org/2021.acl-long.416.

95. ^Zhang T, Zeng Z, Xiao Y, Zhuang H, Chen C, Foulds J, Pan S (2024). "GenderAlign: An Alignment Dataset for Mitigating Gender Bias in Large Language Models". arXiv preprint arXiv:2406.13925.

96. ^Zeng Y, Yang Y, Zhou A, Tan JZ, Tu Y, Mai Y, Klyman K, Pan M, Jia R, Song D, et al. Air-bench 2024: A safety benchmark based on risk categories from regulations and policies. arXiv preprint arXiv:2407.17436. 2024.

97. ^Yin S, Pang X, Ding Y, Chen M, Bi Y, Xiong Y, Huang W, Xiang Z, Shao J, Chen S (2024). "SafeAgentBench: A Benchmark for Safe Task Planning of Embodied LLM Agents". arXiv preprint arXiv:2412.13178.

98. ^Yuan T, He Z, Dong L, Wang Y, Zhao R, Xia T, Xu L, Zhou B, Li F, Zhang Z, Wang R, Liu G (2024). "R-Judge: Benchmarking Safety Risk Awareness for LLM Agents". arXiv preprint arXiv:2401.10019.

99. ^Zhang H, Huang J, Mei K, Yao Y, Wang Z, Zhan C, Wang H, Zhang Y (2024). "Agent security bench (asb): Formalizing and benchmarking attacks and defenses in llm-based agents". arXiv preprint arXiv:2410.02644.

100. ^Andriushchenko M, Souly A, Dziemian M, Duenas D, Lin M, Wang J, Hendrycks D, Zou A, Kolter Z, Fredrikson M, et al. (2024). "Agentharm: A benchmark for measuring harmfulness of llm agents". arXiv preprint arXiv:2410.09024.

101. ^Ji J, Liu M, Dai J, Pan X, Zhang C, Bian C, Chen B, Sun R, Wang Y, Yang Y (2024). "Beavertails: Towards improved safety alignment of llm via a human-preference dataset". Advances in Neural Information Processing Systems. 36.

102. a, bJi J, Hong D, Zhang B, Chen B, Dai J, Zheng B, Qiu T, Li B, Yang Y (2024). "PKU-SafeRLHF: Towards Multi-Level Safety Alignment for LLMs with Human Preference". arXiv preprint arXiv:2406.15513.

103. ^Köpf A, Kilcher Y, von Rütte D, Anagnostidis S, Tam ZR, Stevens K, Barhoum A, Nguyen D, Stanley O, Nagyfi R, et al. Openassistant conversations-democratizing large language model alignment. Advances i

n Neural Information Processing Systems. 36; 2024.

104. ^*Askell A, Bai Y, Chen A, Drain D, Ganguli D, Henighan T, Jones A, Joseph N, Mann B, DasSarma N, Elha ge N, Hatfield-Dodds Z, Hernandez D, Kernion J, Ndousse K, Olsson C, Amodei D, Brown TB, Clark J, Mc Candlish S, Olah C, Kaplan J (2021). "A General Language Assistant as a Laboratory for Alignment". CoR R. **abs/2112.00861**. Available from: https://arxiv.org/abs/2112.00861.*

105. ^*Liu X, Lei X, Wang S, Huang Y, Feng Z, Wen B, Cheng J, Ke P, Xu Y, Tam WL, et al. Alignbench: Benchma rking chinese alignment of large language models. arXiv preprint arXiv:2311.18743. 2023.*

106. ^*Lee J, Kim M, Kim S, Kim J, Won S, Lee H, Choi E (2024). "KorNAT: LLM Alignment Benchmark for Kor ean Social Values and Common Knowledge". arXiv preprint arXiv:2402.13605.*

107. a, b*Wang B, Chen W, Pei H, Xie C, Kang M, Zhang C, Xu C, Xiong Z, Dutta R, Schaeffer R, et al. DecodingT rust: A Comprehensive Assessment of Trustworthiness in GPT Models. In: NeurIPS; 2023.*

108. ^*Huang Y, Sun L, Wang H, Wu S, Zhang Q, Li Y, Gao C, Huang Y, Lyu W, Zhang Y, Li X, Sun H, Liu Z, Liu Y, Wang Y, Zhang Z, Vidgen B, Kailkhura B, Xiong C, Xiao C, Li C, Xing EP, Huang F, Liu H, Ji H, Wang H, Zhang H, Yao H, Kellis M, Zitnik M, Jiang M, Bansal M, Zou J, Pei J, Liu J, Gao J, Han J, Zhao J, Tang J, Wa ng J, Vanschoren J, Mitchell J, Shu K, Xu K, Chang KW, He L, Huang L, Backes M, Gong NZ, Yu PS, Chen P Y, Gu Q, Xu R, Ying R, Ji S, Jana S, Chen T, Liu T, Zhou T, Wang WY, Li X, Zhang X, Wang X, Xie X, Chen X, Wang X, Liu Y, Ye Y, Cao Y, Chen Y, Zhao Y. "TrustLLM: Trustworthiness in Large Language Models." In: Forty-first International Conference on Machine Learning, 2024. Available from: https://openreview.ne t/forum?id=bWUUoLwwMp.*

109. ^*Li L, Dong B, Wang R, Hu X, Zuo W, Lin D, Qiao Y, Shao J (2024). "Salad-bench: A hierarchical and com prehensive safety benchmark for large language models". arXiv preprint arXiv:2402.05044.*

110. ^*Wang Y, Li H, Han X, Nakov P, Baldwin T (2023). "Do-Not-Answer: A Dataset for Evaluating Safeguar ds in LLMs". arXiv preprint arXiv:2308.13387. Available from: https://arxiv.org/abs/2308.13387.*

111. ^*Vidgen B, Scherrer N, Kirk HR, Qian R, Kannappan A, Hale SA, Röttger P (2023). "Simplesafetytests: a t est suite for identifying critical safety risks in large language models". arXiv preprint arXiv:2311.08370. arXiv:2311.08370.*

112. ^*Rose DJ (1969). "Engineering feasibility of controlled fusion: A review". Nuclear Fusion. 9 (3): 183.*

113. ^*Raeder J, Borrass K, Bunde R, Danner W, Klingelhofer R, Lengyel L, Leuterer F, Soll M. "Controlled nucl ear fusion: fundamentals of its utilization for energy supply". John Wiley and Sons, Inc., New York, NY; 1 986.*

114. a, b Bradford T. Solar revolution: the economic transformation of the global energy industry. MIT Press; 2008.

115. a, b Meneguzzo F, Ciriminna R, Albanese L, Pagliaro M (2015). "The great solar boom: a global perspecti ve into the far reaching impact of an unexpected energy revolution". Energy Science & Engineering. 3 (6): 499–509.

116. a, b, c, d, e McNeil SE (2005). "Nanotechnology for the biologist". Journal of leukocyte biology. 78 (3): 585 –594.

117. a, b, c, d, e Bhushan B. "Introduction to nanotechnology". Springer handbook of nanotechnology. 2017: 1 –19.

118. a, b, c, d, e Emerich DF, Thanos CG (2003). "Nanotechnology and medicine". Expert opinion on biological therapy. 3 (4): 655–663.

119. a, b, c Bousso R (2002). "The holographic principle". Reviews of Modern Physics. 74 (3): 825.

120. a, b, c Yaraş F, Kang H, Onural L (2010). "State of the art in holographic displays: a survey". Journal of dis play technology. 6 (10): 443–454.

121. a, b, c Benton SA, Bove Jr VM. Holographic imaging. John Wiley & Sons; 2008.

122. ^ Shahrubudin N, Lee TC, Ramlan RJPM (2019). "An overview on 3D printing technology: Technological, materials, and applications". Procedia Manufacturing. 35: 1286–1296.

123. ^ Gopinathan J, Noh I (2018). "Recent trends in bioinks for 3D printing". Biomaterials research. 22 (1): 1 1.

124. a, b Kaufman HR. "Technology of electron-bombardment ion thrusters." In: Advances in electronics and electron physics. Elsevier; 1975. p. 265–373.

125. a, b Dietz P, Gärtner W, Koch Q, Köhler PE, Teng Y, Schreiner PR, Holste K, Klar PJ (2019). "Molecular pro pellants for ion thrusters". Plasma Sources Science and Technology. 28 (8): 084001.

126. a, b Polk JE, Sekerak MJ, Ziemer JK, Schein J, Qi N, Anders A (2008). "A theoretical analysis of vacuum arc thruster and vacuum arc ion thruster performance". IEEE Transactions on Plasma Science. 36 (5): 2167– 2179.

127. ^ Musk E (2013). "Hyperloop Alpha". Available from: https://www.tesla.com/sites/default/files/blog_im ages/hyperloop-alpha.pdf.

128. ^ Premsagar S, Kenworthy J (2022). "A critical review of hyperloop (ultra-high speed rail) technology: U rban and transport planning, technical, environmental, economic, and human considerations". Frontier s in Sustainable Cities. 4: 842245.

129. ^Lang AJ, Connolly DP, de Boer G, Shahpar S, Hinchliffe B, Gilkeson CA (2024). "A review of Hyperloop aerodynamics". Computers & Fluids. Page 106202.

130. a, bGruska J, et al. Quantum computing. McGraw-Hill London; 1999. Volume 2005.

131. a, bRieffel E, Polak W (2000). "An introduction to quantum computing for non-physicists". ACM Computing Surveys (CSUR). 32 (3): 300–335.

132. a, b, cMahesh B. Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet]. 9(1):381–386, 2020.

133. ^Wettig A, Gupta A, Malik S, Chen D (2024). "Qurating: Selecting high-quality data for training language models". arXiv preprint arXiv:2402.09739.

134. ^Zhao P, Shan JW, Zhang YJ, Zhou ZH (2024). "Exploratory machine learning with unknown unknowns". Artificial Intelligence. 327: 104059.

135. ^Xia M, Malladi S, Gururangan S, Arora S, Chen D (2024). "Less: Selecting influential data for targeted instruction tuning". arXiv preprint arXiv:2402.04333.

136. a, bDixon MF, Halperin I, Bilokon P. Machine learning in finance. Springer; 2020. Volume 1170.

137. a, bErickson BJ, Korfiatis P, Akkus Z, Kline TL (2017). "Machine learning for medical imaging". Radiographics. 37 (2): 505–515.

138. a, bGuo LZ, Zhou Z, Li YF, Zhou ZH. "Identifying useful learnwares for heterogeneous label spaces." In: International Conference on Machine Learning. PMLR; 2023. p. 12122-12131.

139. a, bWu JH, Zhang SQ, Jiang Y, Zhou ZH (2023). "Theoretical exploration of flexible transmitter model". IEEE Transactions on Neural Networks and Learning Systems. 2023. Published by IEEE.

140. a, bQian YY, Zhang ZY, Zhao P, Zhou ZH (2024). "Learning with asynchronous labels". ACM Transactions on Knowledge Discovery from Data. ACM New York, NY.

141. ^Shinde PP, Shah S. "A review of machine learning and deep learning applications." In: 2018 Fourth international conference on computing communication control and automation (ICCUBEA). IEEE; 2018. p. 1-6.

142. ^Zhou ZH, Tan ZH (2024). "Learnware: Small models do big". Science China Information Sciences. 67 (1): 112102.

143. ^Guo LZ, Zhang ZY, Jiang Y, Li YF, Zhou ZH. "Safe deep semi-supervised learning for unseen-class unlabeled data." In: International conference on machine learning. PMLR; 2020. p. 3897-3906.

144. ^ Li Z, Zhou ZH, Gretton A (2021). "Towards an understanding of benign overfitting in neural networks". arXiv preprint arXiv:2106.03212.

145. ^ Yuan H, Cong G, Li G (2024). "Nuhuo: An Effective Estimation Model for Traffic Speed Histogram Imputation on a Road Network". Proceedings of the VLDB Endowment. **17** (7): 1605–1617.

146. ^ Ma J, Li S, Qin H, Hao A, Zhao Q (2022). "Self-adjustable hyper-graphs for video pose estimation based on spatial-temporal subspace construction". Science China. Information Sciences. **65** (3): 139101.

147. ^ Gong Y, Yi J, Chen DD, Zhang J, Zhou J, Zhou Z (2021). "Inferring the importance of product appearance with semi-supervised multi-modal enhancement: A step towards the screenless retailing". Proceedings of the 29th ACM International Conference on Multimedia. pp. 1120–1128.

148. ^ Dai WZ, Xu Q, Yu Y, Zhou ZH (2019). "Bridging machine learning and logical reasoning by abductive learning". Advances in Neural Information Processing Systems. **32**.

149. ^ Kendall A, Gal Y (2017). "What uncertainties do we need in bayesian deep learning for computer vision?" ArXiv. **abs/1703.04977**. S2CID 71134.

150. ^ Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2015). "Rethinking the Inception Architecture for Computer Vision". 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pages 2818–2826. S2CID 206593880.

151. ^ Wrobel BP (2001). "Multiple View Geometry in Computer Vision". Künstliche Intell.. **15**: 41. S2CID 261497446.

152. ^ Wang Y, Zou X, Yan L, Zhong S, Zhou J (2024). "SNIDA: Unlocking Few-Shot Object Detection with Non-linear Semantic Decoupling Augmentation". In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12544–12553.

153. ^ Dong L, Li Z, Xu K, Zhang Z, Yan L, Zhong S, Zou X (2022). "Category-aware transformer network for better human-object interaction detection". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pages 19538–19547.

154. ^ Tran AT, Hassner T, Masi I, Paz E, Nirkin Y, Medioni GG, et al. Extreme 3D Face Reconstruction: Seeing Through Occlusions. In: CVPR. 2018; **1**:2.

155. ^ Burger W, Burge M (2016). "Digital image processing". In: Texts in Computer Science. S2CID 519247.

156. ^ Sonka M, Hlavac V, Boyle R. Image processing, analysis and machine vision. Springer; 2013.

157. ^ Tang Y, Ding D, Rao Y, Zheng Y, Zhang D, Zhao L, Lu J, Zhou J (2019). "COIN: A Large-Scale Dataset for Comprehensive Instructional Video Analysis". 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pages 1207-1216. S2CID 71147568.

158. ^Xie S, Sun C, Huang J, Tu Z, Murphy KP (2017). "Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-offs in Video Classification." In: European Conference on Computer Vision. S2CID 51863579.

159. ^Xu N, Yang L, Fan Y, Yang J, Yue D, Liang Y, Price BL, Cohen SD, Huang TS. "YouTube-VOS: Sequence-to-Sequence Video Object Segmentation." In: European Conference on Computer Vision; 2018. Available from: https://api.semanticscholar.org/CorpusID:52154988.

160. ^Eronen AJ, Peltonen VT, Tuomi JT, Klapuri AP, Fagerlund S, Sorsa T, Lorho G, Huopaniemi J (2005). "Audio-based context recognition". IEEE Transactions on Audio, Speech, and Language Processing. **14** (1): 321–329.

161. ^Potamianos G, Neti C, Luettin J, Matthews I (2004). "Audio-visual automatic speech recognition: An overview". Issues in visual and audio-visual speech processing. **22**: 23. MIT Press Cambridge.

162. ^Afouras T, Chung JS, Senior A, Vinyals O, Zisserman A (2018). "Deep audio-visual speech recognition". IEEE Transactions on Pattern Analysis and Machine Intelligence. **44** (12): 8717–8727.

163. ^Ko T, Peddinti V, Povey D, Khudanpur S (2015). "Audio augmentation for speech recognition." In: Interspeech. **2015**: 3586.

164. a, b, cAbbasian M, Azimi I, Rahmani AM, Jain R (2023). "Conversational health agents: A personalized llm-powered agent framework". arXiv preprint arXiv:2310.02374. arXiv:2310.02374.

165. a, b, c, dZhang H, Hong J, Dong F, Drew S, Xue L, Zhou J (2023). "A privacy-preserving hybrid federated learning framework for financial crime detection". arXiv preprint arXiv:2302.03654.

166. a, b, c, dRahman MM, Watanobe Y (2023). "ChatGPT for education and research: Opportunities, threats, and strategies". Applied Sciences. **13** (9): 5783.

167. a, b, c, dMbakwe AB, Lourentzou I, Celi LA, Mechanic OJ, Dagan A (2023). "ChatGPT passing USMLE shines a spotlight on the flaws of medical education". PLOS Digital Health. **2** (2): e0000205. Public Library of Science San Francisco, CA USA.

168. a, b, cLi G, Hammoud H, Itani H, Khizbullin D, Ghanem B (2023). "Camel: Communicative agents for 'mind' exploration of large language model society". Advances in Neural Information Processing Systems. **36**: 51991–52008.

169. ^Wang L, Ma C, Feng X, Zhang Z, Yang H, Zhang J, Chen Z, Tang J, Chen X, Lin Y, et al. (2024). "A survey on large language model based autonomous agents". Frontiers of Computer Science. **18** (6): 186345.

170. ^Xu B, Yang A, Lin J, Wang Q, Zhou C, Zhang Y, Mao Z (2023). "Expertprompting: Instructing large language models to be distinguished experts". arXiv preprint arXiv:2305.14688.

171. ^Han S, Zhang Q, Yao Y, Jin W, Xu Z, He C (2024). "LLM multi-agent systems: Challenges and open problems". arXiv preprint arXiv:2402.03578.

172. ^Carmigniani J, Furht B, Anisetti M, Ceravolo P, Damiani E, Ivkovic M (2011). "Augmented reality technologies, systems and applications". Multimedia Tools and Applications. **51**: 341–377.

173. ^Chang G, Morreale P, Medicherla P. "Applications of augmented reality systems in education." In: Society for Information Technology & Teacher Education International Conference. Association for the Advancement of Computing in Education (AACE); 2010. p. 1380-1385.

174. ^Balantekin AB, Takigawa N (1998). "Quantum tunneling in nuclear fusion". Reviews of Modern Physics. **70** (1): 77.

175. ^Morse E. Nuclear Fusion. Springer; 2018.

176. ^Ichimaru S (1993). "Nuclear fusion in dense plasmas". Reviews of Modern Physics. **65** (2): 255.

177. ^Barbarino M (2020). "A brief history of nuclear fusion". Nature Physics. **16** (9): 890–893.

178. ^Miyamoto K. Plasma physics and controlled nuclear fusion. 38th ed. Springer Science & Business Media; 2005.

179. ^Knap V, Vestergaard LK, Stroe DI (2020). "A review of battery technology in CubeSats and small satellite solutions". Energies. **13** (16): 4097.

180. ^Chin KB, Brandon EJ, Bugga RV, Smart MC, Jones SC, Krause FC, West WC, Bolotin GG (2018). "Energy storage technologies for small satellite applications". Proceedings of the IEEE. **106** (3): 419--428.

181. ^Hill CA (2011). "Satellite battery technologyA tutorial and overview". IEEE Aerospace and Electronic Systems Magazine. **26**(6): 3843.

182. ^Verduci R, Romano V, Brunetti G, Yaghoobi Nia N, Di Carlo A, D'Angelo G, Ciminelli C (2022). "Solar energy in space applications: review and technology perspectives". Advanced Energy Materials. **12** (29): 2200125.

183. a, b, c, dD'Alessandro MA. Defending LLMs against Jailbreaking: Definition, examples and prevention [Internet]. 2024. Available from: https://www.giskard.ai/knowledge/defending-llms-against-jailbreaking-definition-examples-prevention.

184. a, b, c, d, e, fSwaelens N, Yaremchuk O (2024). "LLM Security: Going Beyond Firewalls". Available from: https://protectai.com/blog/llm-security-going-beyond-firewalls.

185. ^Palmer L. Emerging Technologies in Privacy: AI and Machine Learning [Internet]. 2024. Available from: https://trustarc.com/resource/emerging-technologies-privacy-ai-machine-learning/.

186. a, b, c Reinstein S. "What's New from DataGrail September 2024". 2024. Available from: https://www.datagrail.io/blog/product/whats-new-from-datagrail-september-2024/.

187. a, b Welsh S (2024). "Arize Release Notes: Embeddings Tracing, Experiments Details, and More". https://arize.com/blog/arize-release-notes-embeddings-tracing-experiments-details-and-more/.

188. a, b, c Dastra Team. Dastra. 2024. Available from: https://www.dastra.eu/en.

189. a, b, c Robb Taylor-Hiscock (2024). "Responsible use in the data lifecycle: Process, share, and use." Available from: https://www.onetrust.com/blog/responsible-use-in-the-data-lifecycle-process-share-and-use/.

190. ^ Bagdasaryan E, Veit A, Hua Y, Estrin D, Shmatikov V (2020). "How to backdoor federated learning". In: International Conference on Artificial Intelligence and Statistics. PMLR. pp. 2938–2948.

191. ^ Chen Y, Su L, Xu J (2017). "Distributed statistical machine learning in adversarial settings: Byzantine gradient descent". ACM on Measurement and Analysis of Computing Systems. **1** (2): 1–25.

192. ^ Li Y, Liu S, Chen K, Xie X, Zhang T, Liu Y (2023). "Multi-target backdoor attacks for code pre-trained models". arXiv preprint arXiv:2306.08350.

193. ^ Wan A, Wallace E, Shen S, Klein D. "Poisoning language models during instruction tuning." In: International Conference on Machine Learning. PMLR; 2023. p. 35413-35425.

194. ^ Shu M, Wang J, Zhu C, Geiping J, Xiao C, Goldstein T (2023). "On the exploitability of instruction tuning". Advances in Neural Information Processing Systems. **36**: 61836–61856.

195. ^ Yang H, Xiang K, Ge M, Li H, Lu R, Yu S (2024). "A comprehensive overview of backdoor attacks in large language models within communication networks". IEEE Network. IEEE.

196. ^ Greshake K, Abdelnabi S, Mishra S, Endres C, Holz T, Fritz M (2023). "Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection". Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security. 2023: 79–90.

197. a, b Liu Y, Deng G, Xu Z, Li Y, Zheng Y, Zhang Y, Zhao L, Zhang T, Liu Y (2023). "Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study". ArXiv. **abs/2305.13860**. doi:10.48550/arXiv.2305.13860.

198. a, b Chao P, Robey A, Dobriban E, Hassani H, Pappas GJ, Wong E (2023). "Jailbreaking black box large language models in twenty queries". arXiv preprint arXiv:2310.08419.

199. a, b, c, d, e Wu F, Zhang N, Jha S, McDaniel P, Xiao C (2024). "A new era in llm security: Exploring security concerns in real-world llm-based systems". arXiv preprint arXiv:2402.18649. arXiv:2402.18649.

200. a, b, c Bartolo M, Thrush T, Jia R, Riedel S, Stenetorp P, Kiela D (2021). "Improving question answering model robustness with synthetic adversarial data generation". arXiv preprint arXiv:2104.08678. Available

from: https://arxiv.org/abs/2104.08678.

201. a, b, c, d, e*Ganguli D, Lovitt L, Kernion J, Askell A, Bai Y, Kadavath S, Mann B, Perez E, Schiefer N, Ndouss e K, et al. (2022). "Red teaming language models to reduce harms: Methods, scaling behaviors, and less ons learned". arXiv preprint arXiv:2209.07858. arXiv:2209.07858.*

202. a, b, c, d, e*Perez E, Huang S, Song F, Cai T, Ring R, Aslanides J, Glaese A, McAleese N, Irving G (2022). "Re d teaming language models with language models". arXiv preprint arXiv:2202.03286.*

203. a, b, c*Ribeiro MT, Wu T, Guestrin C, Singh S. Beyond accuracy: Behavioral testing of NLP models with Ch eckList. In: Jurafsky D, Chai J, Schluter N, Tetreault J, editors. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics; 2020. p. 4902-4912. doi:10.18653/v1/2020.acl-main.442. Available from: https://aclanthology.org/2020.acl-main.442.*

204. a, b, c*Wallace E, Rodriguez P, Feng S, Yamada I, Boyd-Graber J (2019). "Trick me if you can: Human-in -the-loop generation of adversarial examples for question answering". Transactions of the Association for Computational Linguistics. 7: 387–401.*

205. a, b*Kumar D, AbuHashem YA, Durumeric Z (2024). "Watch Your Language: Investigating Content Mode ration with Large Language Models". Proceedings of the International AAAI Conference on Web and Soc ial Media. 18: 865–878.*

206. a, b*Wu J, Deng J, Pang S, Chen Y, Xu J, Li X, Xu W (2024). "Legilimens: Practical and Unified Content Mo deration for Large Language Model Services". arXiv preprint arXiv:2408.15488.*

207. a, b*Qiao W, Dogra T, Stretcu O, Lyu YH, Fang T, Kwon D, Lu CT, Luo E, Wang Y, Chia CC, et al. Scaling Up LLM Reviews for Google Ads Content Moderation. In: Proceedings of the 17th ACM International Confer ence on Web Search and Data Mining. 2024. p. 1174-1175.*

208. a, b*Elesedy H, Esperan7ca PM, Oprea SV, Ozay M (2024). "LoRA-Guard: Parameter-Efficient Guardrail Adaptation for Content Moderation of Large Language Models". arXiv preprint arXiv:2407.02987. Avail able from: https://arxiv.org/abs/2407.02987.*

209. ^*Cao B, Cao Y, Lin L, Chen J (2023). "Defending against alignment-breaking attacks via robustly aligne d llm". arXiv preprint arXiv:2309.14348. arXiv:2309.14348.*

210. a, b*Varshney N, Dolin P, Seth A, Baral C (2023). "The Art of Defending: A Systematic Evaluation and Ana lysis of LLM Defense Strategies on Safety and Over-Defensiveness". ArXiv. **abs/2401.00287**. S2CID 2666 93612.*

211. ^Zollo TP, Morrill T, Deng Z, Snell JC, Pitassi T, Zemel R (2023). "Prompt Risk Control: A Rigorous Framework for Responsible Deployment of Large Language Models". arXiv preprint arXiv:2311.13628. arXiv:2311.13628.

212. ^Wang Y, Zhong W, Li L, Mi F, Zeng X, Huang W, Shang L, Jiang X, Liu Q (2023). "Aligning large language models with human: A survey". arXiv preprint arXiv:2307.12966. arXiv:2307.12966.

213. ^Perez E, Ringer S, Lukosiute K, Nguyen K, Chen E, Heiner S, Pettit C, Olsson C, Kundu S, Kadavath S, Jones A, Chen A, Mann B, Israel B, Seethor B, McKinnon C, Olah C, Yan D, Amodei D, Amodei D, Drain D, Li D, Tran-Johnson E, Khundadze G, Kernion J, Landis J, Kerr J, Mueller J, Hyun J, Landau J, Ndousse K, Goldberg L, Lovitt L, Lucas M, Sellitto M, Zhang M, Kingsland N, Elhage N, Joseph N, Mercado N, DasSarma N, Rausch O, Larson R, McCandlish S, Johnston S, Kravec S, El Showk S, Lanham T, Telleen-Lawton T, Brown T, Henighan T, Hume T, Bai Y, Hatfield-Dodds Z, Clark J, Bowman SR, Askell A, Grosse R, Hernandez D, Ganguli D, Hubinger E, Schiefer N, Kaplan J. Discovering language model behaviors with model-written evaluations. In: Rogers A, Boyd-Graber J, Okazaki N, editors. Findings of the Association for Computational Linguistics: ACL 2023. Toronto, Canada: Association for Computational Linguistics; 2023. p. 13387-13434. doi:10.18653/v1/2023.findings-acl.847. Available from: https://aclanthology.org/2023.findings-acl.847.

214. ^Tian JJ, Emerson D, Pandya D, Seyyed-Kalantari L, Khattak F. Efficient evaluation of bias in large language models through prompt tuning. In: Socially Responsible Language Modelling Research; 2023.

215. ^Shen T, Jin R, Huang Y, Liu C, Dong W, Guo Z, Wu X, Liu Y, Xiong D (2023). "Large language model alignment: A survey". arXiv preprint arXiv:2309.15025. **2023**.

216. ^Wang H, Hong J, Zhou J, Wang Z (2023). "How robust is your fairness? evaluating and sustaining fairness under unseen distribution shifts". Transactions on machine learning research. **2023**. NIH Public Access.

217. a, bLi Q, Hong J, Xie C, Tan J, Xin R, Hou J, Yin X, Wang Z, Hendrycks D, Wang Z, et al. LLM-PBE: Assessing Data Privacy in Large Language Models. arXiv preprint arXiv:2408.12787. 2024.

218. ^Li Z, Hong J, Li B, Wang Z (2024). "Shake to leak: Fine-tuning diffusion models can amplify the generative privacy risk". In: 2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML). IEEE. pp. 18–32.

219. a, bHong J, Wang JT, Zhang C, Li Z, Li B, Wang Z (2023). "Dp-opt: Make large language model your privacy-preserving prompt engineer". arXiv preprint arXiv:2312.03724.

220. a, b *Li H, Guo D, Fan W, Xu M, Huang J, Song Y (2023). "Multi-step Jailbreaking Privacy Attacks on Chat GPT". ArXiv. **abs/2304.05197**. S2CID 258060250.*

221. a, b, c *Birkstedt T, Minkkinen M, Tandon A, Me4ntyme4ki M (2023). "AI governance: themes, knowledge gaps and future agendas". Internet Research. **33** (7): 133–167.*

222. a, b *Zou A, Phan L, Chen S, Campbell J, Guo P, Ren R, Pan A, Yin X, Mazeika M, Dombrowski AK, et al. (2023). "Representation engineering: A top-down approach to ai transparency". arXiv preprint arXiv:2310.01405. Available from: https://arxiv.org/abs/2310.01405.*

223. a, b *Me4ntyme4ki M, Minkkinen M, Birkstedt T, Viljanen M (2022). "Defining organizational AI governance". AI and Ethics. 2 (4): 603–609.*

224. ^ *Wang Y, Ma H, Yang J, Wang K (2017). "Industry 4.0: a way from mass customization to mass personalization production". Advances in Manufacturing. **5**: 311–320. S2CID 168907632.*

225. ^ *Liu Y, Deng G, Li Y, Wang K, Wang Z, Wang X, Zhang T, Liu Y, Wang H, Zheng Y, et al. (2023). "Prompt Injection attack against LLM-integrated Applications". arXiv preprint arXiv:2306.05499.*

226. ^ *Pedro R, Castro D, Carreira P, Santos N (2023). "From prompt injections to sql injection attacks: How protected is your llm-integrated web application?" arXiv preprint arXiv:2308.01990.*

227. a, b *Choquet G, Aizier A, Bernollin G (2024). "Exploiting Privacy Vulnerabilities in Open Source LLMs Using Maliciously Crafted Prompts".*

228. ^ *Piet J, Alrashed M, Sitawarin C, Chen S, Wei Z, Sun E, Alomair B, Wagner D. "Jatmo: Prompt injection defense by task-specific finetuning." In: European Symposium on Research in Computer Security. Springer; 2024. p. 105–124.*

229. a, b, c *Perez F, Ribeiro I (2022). "Ignore previous prompt: Attack techniques for language models". arXiv preprint arXiv:2211.09527.*

230. a, b, c *Suo X. "Signed-Prompt: A new approach to prevent prompt injection attacks against LLM-integrated applications". arXiv preprint arXiv:2401.07612. 2024.*

231. a, b, c *Toyer S, Watkins O, Mendes EA, Svegliato J, Bailey L, Wang T, Ong I, Elmaaroufi K, Abbeel P, Darrell T, et al. Tensor trust: Interpretable prompt injection attacks from an online game. arXiv preprint arXiv:2311.01011. 2023.*

232. a, b, c *Liu X, Zhu Y, Lan Y, Yang C, Qiao Y (2023). "Query-relevant images jailbreak large multi-modal models". arXiv preprint arXiv:2311.17600. arXiv:2311.17600.*

233. a, b *Wei A, Haghtalab N, Steinhardt J (2024). "Jailbroken: How does llm safety training fail?" Advances in Neural Information Processing Systems. **36**.*

234. a, b, c Diaa A, Aremu T, Lukas N (2024). "Optimizing Adaptive Attacks against Content Watermarks for L anguage Models". arXiv preprint arXiv:2410.02440.

235. a, b Song J, Huang Y, Zhou Z, Ma L (2024). "Multilingual Blending: LLM Safety Alignment Evaluation wi th Language Mixture". arXiv preprint arXiv:2407.07342.

236. a, b, c Yan S, Wang S, Duan Y, Hong H, Lee K, Kim D, Hong Y (2024). "An LLM-Assisted Easy-to-Trigger Backdoor Attack on Code Completion Models: Injecting Disguised Vulnerabilities against Strong Detecti on". arXiv preprint arXiv:2406.06822. arXiv:2406.06822.

237. ^ Teng M, Jia X, Duan R, Li X, Huang Y, Chu Z, Liu Y, Ren W (2024). "Heuristic-Induced Multimodal Risk Distribution Jailbreak Attack for Multimodal Large Language Models". arXiv preprint arXiv:2412.05934.

238. a, b PromptArmor Team (2024). "Slack AI data exfiltration from private channels via indirect prompt inj ection". https://promptarmor.substack.com/p/slack-ai-data-exfiltration-from-private.

239. a, b Longpre S, Kapoor S, Klyman K, Ramaswami A, Bommasani R, Blili-Hamelin B, Huang Y, Skowron A, Yong ZX, Kotha S, et al. A safe harbor for ai evaluation and red teaming. arXiv preprint arXiv:2403.04 893. 2024.

240. a, b Yu J, Lin X, Yu Z, Xing X (2024). "$\{$LLM-Fuzzer$\}$: Scaling assessment of large language model jailbreaks". In: 33rd USENIX Security Symposium (USENIX Security 24). pp. 4657–4674.

241. ^ Derczynski L, Galinkin E, Martin J, Majumdar S, Inie N (2024). "garak: A Framework for Security Probi ng Large Language Models". arXiv preprint arXiv:2406.11036.

242. ^ Virtue AI Team (2024). "Accelerating Trust in AI: The Rivos and Virtue AI Approach to AI Safety and Se curity." Available from: https://www.virtueai.com/2024/10/17/accelerating-trust-in-ai-the-rivos-an d-virtue-ai-approach-to-ai-safety-and-security/.

243. a, b, c, d DynamoAI (2024). "Dynamo AI x Lenovo Pioneering Safeguarded Generative AI Solutions with DynamoGuard". Available from: https://dynamo.ai/dynamo-x-lenovo.

244. ^ Zukovs HG (2024). "Caveat Podcast: The Startup Leading AI Security in the UK with Mindgard's CEO". Available from: https://mindgard.ai/resources/caveat-podcast-the-startup-leading-ai-security-in-th e-uk-mindgard.

245. ^ Ma H, Zhang C, Fu H, Zhao P, Wu B (2023). "Adapting large language models for content moderation: Pitfalls in data engineering and supervised fine-tuning". arXiv preprint arXiv:2310.03400.

246. ^ Jha P, Jain R, Mandal K, Chadha A, Saha S, Bhattacharyya P (2024). "MemeGuard: An LLM and VLM-based Framework for Advancing Content Moderation via Meme Intervention". arXiv preprint arXiv:240 6.05344. Available from: https://arxiv.org/abs/2406.05344.

247. ^Peng Y, Yan S, Lu Z (2019). "Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets." In: Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019), pp. 58–65.

248. ^Xu J, Ju D, Li M, Boureau Y, Weston J, Dinan E (2020). "Recipes for safety in open-domain chatbots". arXiv preprint arXiv:2010.07079. arXiv:2010.07079.

249. ^Chen J, Wang X, Yao Z, Bai Y, Hou L, Li J (2024). "Finding Safety Neurons in Large Language Models". arXiv preprint arXiv:2406.14144.

250. ^Kumar A, Agarwal C, Srinivas S, Li AJ, Feizi S, Lakkaraju H (2023). "Certifying llm safety against adversarial prompting". arXiv preprint arXiv:2309.02705.

251. a, b, cLees A, Tran VQ, Tay Y, Sorensen J, Gupta J, Metzler D, Vasserman L. A new generation of perspective api: Efficient multilingual character-level transformers. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022. p. 3197–3207.

252. a, b, cHanu L, Unitary team (2020). Detoxify. Github. https://github.com/unitaryai/detoxify.

253. a, b, cInan H, Upasani K, Chi J, Rungta R, Iyer K, Mao Y, Tontchev M, Hu Q, Fuller B, Testuggine D, et al. (2023). "Llama guard: Llm-based input-output safeguard for human-ai conversations". arXiv preprint arXiv:2312.06674.

254. ^Niknazar M, Haley PV, Ramanan L, Truong ST, Shrinivasan Y, Bhowmick AK, Dey P, Jagmohan A, Maheshwari H, Ponoth S, et al. (2024). "Building a Domain-specific Guardrail Model in Production". arXiv preprint arXiv:2408.01452.

255. a, bCalypsoAI (2024). "CalypsoAI Moderator Thwarts Attempts to Bypass System Safeguards". Available from: https://calypsoai.com/jailbreak-prevention/.

256. ^Lakera Team (2024). "Lakera CEO Joins Leaders from Meta, Cohere and MIT for AI Safety Session at AI House Davos". https://www.lakera.ai/news/ai-safety-session-at-ai-house-davos.

257. ^BreezeML (2024). "How BreezeML Can Help Enterprises Follow the NIST AI Risk Management Framework." Available from: https://breezeml.ai/2024/02/14/how-breezeml-can-help-enterprises-follow-the-nist-ai-risk-management-framework/.

258. a, bMohiuddin S (2024). "Guardrails AI's Commitment to Responsible Vulnerability Disclosure". GuardRails AI. Available from: https://www.guardrailsai.com/blog/commitment-to-responsible-vulnerability.

259. a, bRebedea T, Dinu R, Sreedhar M, Parisien C, Cohen J (2023). "Nemo guardrails: A toolkit for controllable and safe llm applications with programmable rails". arXiv preprint arXiv:2310.10501. arXiv:2310.10501.

260. ^Duan H, Dziedzic A, Yaghini M, Papernot N, Boenisch F (2023). "On the privacy risk of in-context learning". In: The 61st Annual Meeting Of The Association For Computational Linguistics.

261. ^Miranda M, Ruzzetti ES, Santilli A, Zanzotto FM, Bratières S, Rodolà E (2024). "Preserving privacy in large language models: A survey on current threats and solutions". arXiv preprint arXiv:2408.05212.

262. ^Carlini N, Tramer F, Wallace E, Jagielski M, Herbert-Voss A, Lee K, Roberts A, Brown T, Song D, Erlingsson U, et al. Extracting training data from large language models. In: 30th USENIX Security Symposium (USENIX Security 21); 2021. p. 2633–2650.

263. a, bZhu L, Liu Z, Han S (2019). "Deep leakage from gradients". Advances in neural information processing systems. **32**.

264. a, bHu H, Salcic Z, Sun L, Dobbie G, Yu PS, Zhang X (2022). "Membership inference attacks on machine learning: A survey". ACM Computing Surveys (CSUR). **54** (11s): 1–37.

265. a, bZhang H, Hong J, Deng Y, Mahdavi M, Zhou J (2024). "Understanding deep gradient leakage via inversion influence functions". Advances in neural information processing systems. **36**.

266. ^Sebastian G (2023). "Privacy and data protection in ChatGPT and other AI Chatbots: strategies for securing user information". International Journal of Security and Privacy in Pervasive Computing (IJSPPC). **15** (1): 1–14.

267. ^Gardhouse K (2024). "How Private AI Can Help Financial Institutions Comply with OSFI Guidelines". Available from: https://private-ai.com/en/2024/09/24/osfi-compliance/.

268. a, b, cKavishwar S. Relyance AI: Driving Innovation in Privacy, Security, and AI Governance [Internet]. 2024. Available from: https://www.relyance.ai/resources/relyance-ai-driving-innovation-in-privacy-security-and-ai-governance.

269. a, bZendata Team (2024). "Mastering The AI Supply Chain: From Data to Governance". Available from: https://www.zendata.dev/post/mastering-the-ai-supply-chain-from-data-to-governance.

270. ^Robey A, Wong E, Hassani H, Pappas GJ (2023). "SmoothLLM: Defending Large Language Models Against Jailbreaking Attacks". ArXiv. **abs/2310.03684**. doi:10.48550/arXiv.2310.03684.

271. a, b, cHuang C, Zhao W, Zheng R, Lv H, Dou S, Li S, Wang X, Zhou E, Ye J, Yang Y, et al. SafeAligner: Safety Alignment against Jailbreak Attacks via Response Disparity Guidance. arXiv preprint arXiv:2406.18118. 2024.

272. a, bPhute M, Helbling A, Hull M, Peng S, Szyller S, Cornelius C, Chau DH (2023). "Llm self defense: By self examination, llms know they are being tricked". arXiv preprint arXiv:2308.07308. **2023**.

273. a, b *Wu D, Wang S, Liu Y, Liu N (2024). "Llms can defend themselves against jailbreaking in a practical manner: A vision paper". arXiv preprint arXiv:2402.15727.*

274. a, b *Xiong C, Qi X, Chen P-Y, Ho T-Y (2024). "Defensive prompt patch: A robust and interpretable defens e of llms against jailbreak attacks". arXiv preprint arXiv:2405.20099.*

275. a, b *Cao Z, Yang Y, Zhao H (2024). "Nothing in excess: Mitigating the exaggerated safety for llms via safe ty-conscious activation steering". arXiv preprint arXiv:2408.11491.*

276. a, b *Fu Y, Xiao W, Chen J, Li J, Papalexakis E, Chien A, Dong Y (2024). "Cross-task defense: Instruction-tu ning llms for content safety". arXiv preprint arXiv:2405.15202.*

277. a, b *Wang X, Wu D, Ji Z, Li Z, Ma P, Wang S, Li Y, Liu Y, Liu N, Rahmel J (2024). "SelfDefend: LLMs Can De fend Themselves against Jailbreaking in a Practical Manner". arXiv preprint arXiv:2406.05498. Availabl e from: https://arxiv.org/abs/2406.05498.*

278. a, b *Zeng Y, Wu Y, Zhang X, Wang H, Wu Q (2024). "AutoDefense: Multi-Agent LLM Defense against Jailb reak Attacks". ArXiv. **abs/2403.04783**. S2CID 268297202.*

279. a, b *Papernot N, McDaniel P, Wu X, Jha S, Swami A. Distillation as a defense to adversarial perturbations against deep neural networks. In: 2016 IEEE symposium on security and privacy (SP). IEEE; 2016. p. 582 –597.*

280. a, b *Gowal S, Dvijotham K, Stanforth R, Bunel R, Qin C, Uesato J, Arandjelovic R, Mann T, Kohli P (2018). "On the effectiveness of interval bound propagation for training verifiably robust models". arXiv preprin t arXiv:1810.12715.*

281. a, b *Cohen J, Rosenfeld E, Kolter Z (2019). "Certified adversarial robustness via randomized smoothing". In: international conference on machine learning. PMLR. pp. 1310–1320.*

282. a, b *Shi W, Shea R, Chen S, Zhang C, Jia R, Yu Z (2022). "Just fine-tune twice: Selective differential privac y for large language models". arXiv preprint arXiv:2204.07667. arXiv:2204.07667.*

283. ^ *Shi W, Cui A, Li E, Jia R, Yu Z (2021). "Selective differential privacy for language modeling". arXiv prepr int arXiv:2108.12944. Available from: https://arxiv.org/abs/2108.12944.*

284. a, b *Ram{\'e} A, Vieillard N, Hussenot L, Dadashi R, Cideron G, Bachem O, Ferret J. "Warm: On the benefit s of weight averaged reward models (2024)". arXiv preprint arXiv:2401.12187. Available from: https://a rxiv.org/abs/2401.12187.*

285. a, b *Kavishwar S. "Transcend Gives Enterprises New Ways to Drive Privacy-Compliant Growth". 2024. Av ailable from: https://transcend.io/blog/preference-management-compliant-growth.*

286. a, b, cYu S, Hong J, Zeng Y, Wang F, Jia R, Zhou J (2023). "Who Leaked the Model? Tracking IP Infringers in Accountable Federated Learning".

287. a, b, cYu S, Hong J, Zhang H, Wang H, Wang Z, Zhou J (2023). "Safe and Robust Watermark Injection wit h a Single OoD Image". arXiv preprint arXiv:2309.01786. arXiv:2309.01786.

288. a, b, c, d, eAlabdulakreem A, Arnold CM, Lee Y, Feenstra PM, Katz B, Barbu A (2024). "SecureLLM: Using Compositionality to Build Provably Secure Language Models for Private, Sensitive, and Secret Data". ar Xiv preprint arXiv:2405.09805. arXiv:2405.09805.

289. ^Kirchenbauer J, Geiping J, Wen Y, Katz J, Miers I, Goldstein T. "A watermark for large language model s." In: International Conference on Machine Learning. PMLR; 2023. p. 17061-17084.

290. a, bEhsan U, Liao QV, Muller M, Riedl MO, Weisz JD. "Expanding explainability: Towards social transpar ency in AI systems." In: Proceedings of the 2021 CHI conference on human factors in computing systems. 2021. p. 1-19.

291. a, bChao P, Dobriban E, Hassani H (2024). "Watermarking Language Models with Error Correcting Code s". arXiv preprint arXiv:2406.10281. Available from: https://arxiv.org/abs/2406.10281.

292. ^Saidot (2024). "Quick guide: How to identify, assess, and manage AI risks throughout the AI lifecycle". Available from: https://www.saidot.ai/insights/quick-guide-how-to-identify-assess-and-manage-ai -risks-throughout-the-ai-lifecycle.

293. ^Singh N. "10/10 Vision: Seeing the Future of AI & its Governance Clearly". 2024. Available from: http s://www.credo.ai/blog/10-10-vision-seeing-the-future-of-ai-its-governance-clearly.

294. a, bRaji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, Smith-Loud J, Theron D, Barnes P (2 020). "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic a uditing". Proceedings of the 2020 conference on fairness, accountability, and transparency. 2020: 33–4 4.

295. a, b, cDathathri S, See A, Ghaisas S, Huang PS, McAdam R, Welbl J, Bachani V, Kaskasoli A, Stanforth R, Matejovicova T, et al. (2024). "Scalable watermarking for identifying large language model outputs". N ature. **634** (8035): 818–823.

296. a, b, cGoogle DeepMind (2024). "Watermarking AI-generated text and video with SynthID". https://dee pmind.google/discover/blog/watermarking-ai-generated-text-and-video-with-synthid/.

297. ^OpenAI (2023). "OpenAI Red Teaming Network". Available from: https://openai.com/index/red-teami ng-network/.

298. ^a, ^b*Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, Chen Q, Peng W, Feng X, Qin B, et al. (2023). "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions". arXiv preprint arXiv:2311.05232. arXiv:2311.05232.*

299. ^*Cheng Y, Zhang C, Zhang Z, Meng X, Hong S, Li W, Wang Z, Wang Z, Yin F, Zhao J, et al. Exploring large language model based intelligent agents: Definitions, methods, and prospects. arXiv preprint arXiv:2401.03428. 2024.*

300. ^*Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, Batra S, Bhargava P, Bhosale S, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288. 2023.*

301. ^*Yang A, Yang B, Hui B, Zheng B, Yu B, Zhou C, Li C, Li C, Liu D, Huang F, et al. Qwen2 technical report. arXiv preprint arXiv:2407.10671. 2024.*

302. ^*Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, de las Casas D, Bressand F, Lengyel G, Lample G, Saulnier L, et al. Mistral 7B. arXiv preprint arXiv:2310.06825. 2023.*

303. ^*Hong J, Wang Z, Zhou J (2022). "Dynamic privacy budget allocation improves data efficiency of differentially private gradient descent". Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. pp. 11–35.*

304. ^*Triastcyn A, Faltings B (2020). "Bayesian differential privacy for machine learning". In: International Conference on Machine Learning. PMLR. pp. 9583–9592.*

305. ^*Zhao J, Chen Y, Zhang W (2019). "Differential privacy preservation in deep learning: Challenges, opportunities and solutions". IEEE Access. **7**: 48901–48911.*

306. ^*Zhang S, He X (2023). "Dprovdb: Differentially private query processing with multi-analyst provenance". Proceedings of the ACM on Management of Data. **1** (4): 1–27.*

307. ^*Zhu Z, Hong J, Drew S, Zhou J (2022). "Resilient and communication efficient learning for heterogeneous federated systems". Proceedings of machine learning research. **162**: 27504.*

308. ^*Han S, Buyukates B, Hu Z, Jin H, Jin W, Sun L, Wang X, Wu W, Xie C, Yao Y, et al. "FedSecurity: A Benchmark for Attacks and Defenses in Federated Learning and Federated LLMs." In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2024. p. 5070–5081.*

309. ^*Zhang C, Xie Y, Bai H, Yu B, Li W, Gao Y (2021). "A survey on federated learning". Knowledge-Based Systems. **216**: 106775.*

310. ^*Ouyang X, Xie Z, Fu H, Cheng S, Pan L, Ling N, Xing G, Zhou J, Huang J (2023). "Harmony: Heterogeneous multi-modal federated learning through disentangled model training." Proceedings of the 21st Ann*

ual International Conference on Mobile Systems, Applications and Services. pp. 530–543.

311. ^*Lee JW, Kang H, Lee Y, Choi W, Eom J, Deryabin M, Lee E, Lee J, Yoo D, Kim YS, et al. Privacy-preservin g machine learning with fully homomorphic encryption for deep neural network. IEEE Access. **10**:30039 –30054, 2022.*

312. ^*Pulido-Gaytan LB, Tchernykh A, Cortés-Mendoza JM, Babenko M, Radchenko G. "A survey on privacy -preserving machine learning with fully homomorphic encryption." In: Latin American High Performa nce Computing Conference. Springer; 2020. p. 115–129.*

313. a, b*Jin W, Yao Y, Han S, Joe-Wong C, Ravi S, Avestimehr S, He C (2023). "FedML-HE: An efficient homo morphic-encryption-based privacy-preserving federated learning system". arXiv preprint arXiv:2303.1 0837. Available from: https://arxiv.org/abs/2303.10837.*

314. ^*Yu D, Kairouz P, Oh S, Xu Z (2024). "Privacy-Preserving Instructions for Aligning Large Language Mo dels". arXiv preprint arXiv:2402.13659.*

315. ^*Zhang L, Li B, Thekumparampil KK, Oh S, He N (2024). "DPZero: Private Fine-Tuning of Language M odels without Backpropagation". In: Forty-first International Conference on Machine Learning.*

316. ^*Tang X, Shin R, Inan HA, Manoel A, Mireshghallah F, Lin Z, Gopi S, Kulkarni J, Sim R (2023). "Privacy- preserving in-context learning with differentially private few-shot generation". arXiv preprint arXiv:23 09.11765. 2023.*

317. ^*Duan H, Dziedzic A, Papernot N, Boenisch F (2024). "Flocks of stochastic parrots: Differentially private prompt learning for large language models". Advances in Neural Information Processing Systems. **36**.*

318. ^*Zhou C, Liu P, Xu P, Iyer S, Sun J, Mao Y, Ma X, Efrat A, Yu P, Yu L, et al. Lima: Less is more for alignme nt. Advances in Neural Information Processing Systems. **36**, 2024.*

319. a, b*Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, e t al. (2022). "Training language models to follow instructions with human feedback". Advances in neur al information processing systems. **35**: 27730–27744.*

320. ^*Bai T, Luo J, Zhao J, Wen B, Wang Q (2021). "Recent advances in adversarial training for adversarial ro bustness". arXiv preprint arXiv:2102.01356. Available from: https://arxiv.org/abs/2102.01356.*

321. ^*Chu Z, Chen J, Chen Q, Yu W, He T, Wang H, Peng W, Liu M, Qin B, Liu T (2023). "Navigate through eni gmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future". Annual Meeti ng of the Association for Computational Linguistics. S2CID 263153015.*

322. ^*Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, Zhang L. Deep learning with differe ntial privacy. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications sec*

urity. 2016. p. 308–318.

323. ^*Ji Z, Lipton ZC, Elkan C (2014). "Differential privacy and machine learning: a survey and review". arXiv preprint arXiv:1412.7584. arXiv:1412.7584.*

324. ^*Sun X, Zhang P, Liu JK, Yu J, Xie W (2018). "Private machine learning classification based on fully hom omorphic encryption". IEEE Transactions on Emerging Topics in Computing. 8 (2): 352–364.*

325. ^*Damgrd I, Escudero D, Frederiksen T, Keller M, Scholl P, Volgushev N. "New primitives for actively-sec ure MPC over rings with applications to private machine learning." In: 2019 IEEE Symposium on Securit y and Privacy (SP). IEEE; 2019. p. 1102-1120.*

326. ^*Knott B, Venkataraman S, Hannun A, Sengupta S, Ibrahim M, van der Maaten L (2021). "Crypten: Secu re multi-party computation meets machine learning". Advances in Neural Information Processing Syste ms. 34: 4961–4973.*

327. ^*OWASP (2023). "OWASP Top 10 List for Large Language Models". Available from: https://www.llmtop1 0.com/assets/downloads/OWASP-Top-10-for-LLM-Applications-v1_1.pdf.*

328. a, b, c ***OWASP** (2023). "Supply Chain Vulnerabilities in OWASP Top 10 List for Large Language Models". A vailable from: https://llmtop10.com/llm05/.*

329. a, b *OWASP (2023). "Model Theft in OWASP Top 10 List for Large Language Models". Available from: htt ps://llmtop10.com/llm10/.*

330. a, b *OWASP (2023). "Model Denial of Service in OWASP Top 10 List for Large Language Models". Availabl e from: https://llmtop10.com/llm04/.*

331. a, b, c ***OWASP** (2023). "Insecure Plugin Design in OWASP Top 10 List for Large Language Models". Availa ble from: https://llmtop10.com/llm07/.*

332. ^***OWASP** (2023). "Training Data Poisoning in OWASP Top 10 List for Large Language Models". Availabl e from: https://llmtop10.com/llm03/.*

333. ^*OWASP (2023). "Prompt Injection in OWASP Top 10 List for Large Language Models". Available from: https://llmtop10.com/llm01/.*

334. ^*Rehberger J (2023). "LLM Apps: Don't Get Stuck in an Infinite Loop!" https://embracethered.com/blog/ posts/2023/llm-cost-and-dos-threat/.*

335. ^*Page C. "OpenAI blames DDoS attack for ongoing ChatGPT outage". https://techcrunch.com/2023/11/0 9/openai-blames-ddos-attack-for-ongoing-chatgpt-outage/, 2023.*

336. a, b *Rehberger J. Plugin vulnerabilities: Visit a website and have your source code stolen [Internet]. 2023. Available from: https://embracethered.com/blog/posts/2023/chatgpt-cross-plugin-request-forgery-a*

nd-prompt-injection./.

337. ^Rehberger J. "ChatGPT: Lack of Isolation between Code Interpreter sessions of GPTs". 2024. Available from: https://embracethered.com/blog/posts/2024/lack-of-isolation-gpts-code-interpreter/.

338. ^Bertino E, Ghinita G, Kamra A, et al. (2011). "Access control for databases: Concepts and systems". Foundations and Trends® in Databases. 3 (1–2): 1–148.

339. ^Ouaddah A, Mousannif H, Abou Elkalam A, Ait Ouahman A (2017). "Access control in the Internet of Things: Big challenges and new opportunities". Computer Networks. 112: 237–262.

340. ^Sandhu RS, Samarati P (1994). "Access control: principle and practice". IEEE Communications Magazine. 32 (9): 40–48.

341. ^Bennett CJ. Regulating privacy: Data protection and public policy in Europe and the United States. Ithaca, NY: Cornell University Press; 1992.

342. ^Voigt P, Von dem Bussche A (2017). "The eu general data protection regulation (gdpr)". A Practical Guide, 1st Ed., Cham: Springer International Publishing. 10 (3152676): 10–5555.

343. a, bKandolo W (2024). "Ensuring AI Data Access Control in RDBMS: A Comprehensive Review." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8400–8407.

344. ^Miguel BS, Naseer A, Inakoshi H (2021). "Putting accountability of AI systems into practice". In: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence. pp. 5276–5278.

345. a, b, cNovelli C, Taddeo M, Floridi L (2024). "Accountability in artificial intelligence: what it is and how it works". Ai & Society. 39 (4): 1871–1882.

346. a, bRaja AK, Zhou J (2023). "AI accountability: approaches, affecting factors, and challenges". Computer. 56 (4): 61–70.

347. ^European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). 2024. Available from: https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

348. ^European Union (2016). "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulatio

n) (Text with EEA relevance)". Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02016R0679-20160504.

349. ^The White House (2022). "Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People." Available from: https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

350. ^National Institute of Standards and Technology (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. Available from: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf.

351. a, b, c, dGilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. "Explaining explanations: An overview of interpretability of machine learning." In: 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA). IEEE; 2018. p. 80–89.

352. ^Reyes M, Meier R, Pereira S, Silva CA, Dahlweid F-M, von Tengg-Kobligk H, Summers RM, Wiest R (2020). "On the interpretability of artificial intelligence in radiology: challenges and opportunities". Radiology: artificial intelligence. 2 (3): e190043.

353. ^Xu F, Uszkoreit H, Du Y, Fan W, Zhao D, Zhu J. "Explainable AI: A brief survey on history, research areas, approaches and challenges." In: Natural language processing and Chinese computing: 8th cCF international conference, NLPCC 2019, dunhuang, China, October 9--14, 2019, proceedings, part II 8. Springer; 2019. p. 563--574.

354. ^Wren H (2024). "What is AI transparency? A comprehensive guide". https://www.zendesk.com/blog/ai-transparency/.

355. ^Larsson S, Heintz F (2020). "Transparency in artificial intelligence". Internet policy review. 9 (2).

356. ^Tjoa E, Guan C (2020). "A survey on explainable artificial intelligence (xai): Toward medical xai". IEEE transactions on neural networks and learning systems. 32 (11): 4793–4813.

357. ^Covert I, Lundberg S, Lee S-I (2021). "Explaining by removing: A unified framework for model explanation". Journal of Machine Learning Research. 22 (209): 1–90.

358. a, bLundberg SM, Lee S-I (2017). "A Unified Approach to Interpreting Model Predictions". In: Neural Information Processing Systems. S2CID 21889700.

359. ^Rozemberczki B, Watson L, Bayer P, Yang H, Kiss O, Nilsson S, Sarkar R (2022). "The shapley value in machine learning". arXiv preprint arXiv:2202.05594. arXiv:2202.05594.

360. ^\u0160trumbelj E, Kononenko I (2014). "Explaining prediction models and individual predictions with feature contributions". Knowledge and Information Systems. 41: 647–665.

361. ^Datta A, Sen S, Zick Y (2016). "Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems". In: 2016 IEEE symposium on security and privacy (SP). IEEE. pp. 598–617.

362. ^Wachter S, Mittelstadt B, Russell C (2017). "Counterfactual explanations without opening the black box: Automated decisions and the GDPR". Harv. JL & Tech.. 31: 841.

363. ^Karimi AH, Barthe G, Balle B, Valera I. "Model-agnostic counterfactual explanations for consequential decisions." In: International conference on artificial intelligence and statistics. PMLR; 2020. p. 895–905.

364. ^Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: International conference on machine learning. PMLR; 2018. p. 2668–2677.

365. ^Adebayo J, Gilmer J, Muelly M, Goodfellow IJ, Hardt M, Kim B (2018). "Sanity Checks for Saliency Maps". In: Neural Information Processing Systems. S2CID 52938797.

366. ^Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, Le QV, Zhou D, et al. (2022). "Chain-of-thought prompting elicits reasoning in large language models". Advances in neural information processing systems. 35: 24824–24837.

367. ^Zhang Z, Zhang A, Li M, Smola A (2022). "Automatic chain of thought prompting in large language models". arXiv preprint arXiv:2210.03493.

368. ^Diao S, Wang P, Lin Y, Pan R, Liu X, Zhang T (2023). "Active prompting with chain-of-thought for large language models". arXiv preprint arXiv:2302.12246. arXiv:2302.12246.

369. ^Fu Y, Ou L, Chen M, Wan Y, Peng H, Khot T (2023). "Chain-of-Thought Hub: A Continuous Effort to Measure Large Language Models' Reasoning Performance". arXiv preprint arXiv:2305.17306.

370. ^Long J (2023). "Large language model guided tree-of-thought". arXiv preprint arXiv:2305.08291. arXiv:2305.08291.

371. ^Yao S, Yu D, Zhao J, Shafran I, Griffiths T, Cao Y, Narasimhan K (2024). "Tree of thoughts: Deliberate problem solving with large language models". Advances in Neural Information Processing Systems. 36.

372. ^Cao S, Zhang J, Shi J, Lv X, Yao Z, Tian Q, Li J, Hou L (2023). "Probabilistic Tree-of-thought Reasoning for Answering Knowledge-intensive Complex Questions". arXiv preprint arXiv:2311.13982. arXiv:2311.13982.

373. ^Besta M, Blach N, Kubicek A, Gerstenberger R, Podstawski M, Gianinazzi L, Gajda J, Lehmann T, Niewiadomski H, Nyczyk P, et al. Graph of thoughts: Solving elaborate problems with large language models. Proceedings of the AAAI Conference on Artificial Intelligence. 2024; 38(16):17682–17690.

374. ^Lei B, Liao C, Ding C, et al. (2023). "Boosting logical reasoning in large language models through a new framework: The graph of thought". arXiv preprint arXiv:2308.08614. Available from: https://arxiv.org/abs/2308.08614.

375. ^Jiang Z, Xu FF, Gao L, Sun Z, Liu Q, Dwivedi-Yu J, Yang Y, Callan J, Neubig G (2023). "Active retrieval augmented generation". arXiv preprint arXiv:2305.06983. 2023.

376. a, bLewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W, Rocktäschel T, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems. 33:9459–9474, 2020.

377. ^Erdogan LE, Lee N, Jha S, Kim S, Tabrizi R, Moon S, Hooper C, Anumanchipalli G, Keutzer K, Gholami A (2024). "Tinyagent: Function calling at the edge". arXiv preprint arXiv:2409.00608.

378. ^Lin Q, Wen M, Peng Q, Nie G, Liao J, Wang J, Mo X, Zhou J, Cheng C, Zhao Y, et al. Hammer: Robust Function-Calling for On-Device Language Models via Function Masking. arXiv preprint arXiv:2410.04587. 2024.

379. ^Ran Y, Xu Z, Yao Y, Hu Z, Han S, Jin H, Shah AD, Zhang J, Stripelis D, Zhang T, et al. Alopex: A computational framework for enabling on-device function calls with llms. arXiv preprint arXiv:2411.05209. 2024.

380. ^Nakano R, Hilton J, Balaji S, Wu J, Ouyang L, Kim C, Hesse C, Jain S, Kosaraju V, Saunders W, et al. (2021). "Webgpt: Browser-assisted question-answering with human feedback". arXiv preprint arXiv:2112.09332. arXiv:2112.09332.

381. ^Ribeiro MT, Singh S, Guestrin C (2016). "Why should I trust you? Explaining the predictions of any classifier". In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144.

382. ^Ribeiro MT, Singh S, Guestrin C (2016). "Model-agnostic interpretability of machine learning". arXiv preprint arXiv:1606.05386.

383. ^Bills S, Cammarata N, Mossing D, Tillman H, Gao L, Goh G, Sutskever I, Leike J, Wu J, Saunders W (2023). "Language models can explain neurons in language models". URL https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html (Date accessed: 14.05. 2023), 2.

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.