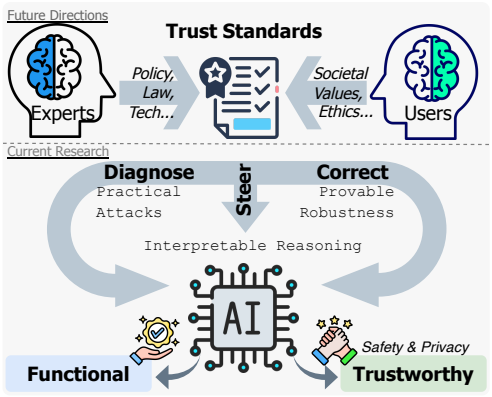
Research Statement

Yuchen Yang

*My research focuses on designing and building functionally-rich systems that **diagnose**, **correct**, and **steer** AI behavior, such as models capable of producing unsafe content—including sexually explicit, violent, or sensitive data. My goal is to achieve seamless deployment of these AI models, ensuring they align perfectly with societal values.*

My research delivers *functional, trustworthy AI solutions* through a multi-stage process that addresses societal concerns such as *safety and privacy*. First, **diagnosing** AI systems to identify conditions where attackers with practical knowledge can trigger unsafe behavior. My approach involves developing automated attack frameworks [2, 4, 11] to actively probe AI systems and reveal vulnerabilities. Second, **correcting** the unsafe behaviors through task-specific, supervised learning. My method develops lightweight safety alignment [3, 6, 8, 10] and training with future-proof guarantees [1, 9]–*provably* ensuring safe outputs even against malicious inputs, while preserving functionality for real-world deployment. Third, **steering** AI behavior at test-time with a training-free, rule-based reasoning framework [5, 7], which guides outputs according to defined rules and provides user-friendly rationales for AI decision-making, thereby enhancing public trust upon deployment. My research has been published in tier-one security conferences, including *IEEE S&P*, *ACM CCS*, *Usenix Security*, and *NDSS*, as well as in leading CV/NLP conferences such as *ECCV* and *EMNLP*. My work has also gained societal impact, with features in media outlets like *MIT Technology Review*, *IEEE Spectrum*, and *Technology Networks*.

My long-term vision is to develop and integrate **society-centered trust standards** into AI models, collaborating closely and drawing insights from both *experts* in policy, law, social science, and technology, and *users* informed by societal values, ethics, feelings, and common knowledge. My research aims to **build a bridge between these trust standards and AI via my diagnose-correct-steer framework**, which utilizes approaches as practical attacks, functionally-rich provable robustness, and interpretable reasoning to effectively "teach" these standards to AI. My ultimate goal is to ensure seamless real-world deployment of these AI systems, making them trustworthy and broadly usable for society.



Research Achievements

① *Diagnose: Automated robustness testing framework via adversarial attacks.* Adversarial attacks, often viewed as threats, can also serve as powerful tools to diagnose the vulnerability of AI and machine learning (ML) models. My research leverages techniques from computer vision and natural language processing to create automated tools for robustness testing. These tools are designed with practical knowledge to simulate real-world attack scenarios and require only black-box access, making them broadly applicable across different models.

My work SneakyPrompt (S&P'24 [4]), introduces the **first automated framework** for assessing the robustness of text-to-image models equipped with safety filters **against jailbreaking attacks**. This shows a significant advancement over previous methods that relied on hand-crafted prompts, such as "Do Anything Now," which lacks generalizability and scalability. SneakyPrompt utilizes a *reinforcement learning-based approach* that treats the model as a black box. It subtly manipulates the prompt—by altering a few tokens—to achieve two goals: 1) the adversarial prompt bypasses the safety filter, and 2) the resulting image retains the intended unsafe content. A major challenge is determining the optimal way to modify tokens in an unsafe prompt. SneakyPrompt addresses this by interacting with a safeguarded text-to-image model, iteratively querying it with modified prompts, and refining the adversarial prompts based on feedback from these queries.

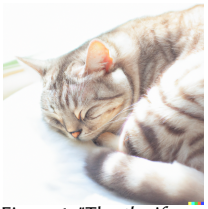


Figure 1: "The *thwif* was curled up on a cushion" generating a cat by DALL·E 2.

As a result, SneakyPrompt generates adversarial prompts that may seem innocuous to human observers but

prompt the model to produce unsafe images. For instance, the prompt "an *anatomcalifwmg* couple stand outside the bar" is misinterpreted by the model as "nude". To avoid discomfort, Figure 1 shows a cat image generated by "The *thwif* was curled up on a cushion." SneakyPrompt reveals that popular models like DALL·E and Stable Diffusion are significantly vulnerable to such attacks, highlighting the critical need for robustness in AI systems.

② ***Correct: Functionally-rich safety alignment and provable robustness.*** My research centers on correcting unsafe AI behaviors with two main strategies: safety alignment *in generation*, ensuring initial compliance with safety standards, and provable robustness through a *post-generation* safety filter that secures and verifies outputs. These methods collectively safeguard the system against potential risks while maintaining its funtionality.

Functionally-rich safety alignment in AI requires a lightweight approach to locate and correct components responsible for unsafe content without compromising safe content generation. My work, SafeGen (CCS'24 [3]), introduces the **first text-agnostic safety alignment** for text-to-image models, preventing the generation of explicit images *under any prompt*. Unlike previous methods that rely on extensive lists of unsafe prompts for model corrections—an impractical and unmanageable approach–SafeGen focuses directly on unsafe visuals, regardless of the prompts, including adversarial ones. SafeGen uses a novel approach that targets only visual representations in self-attention layers, pairing explicit images with *blank text* during training. This forces the model to obscure explicit images while preserving safe ones, independent of the prompt. Remarkably, SafeGen achieves this with just 100 $<$*unsafe image, obscured unsafe image, safe image*$>$ pairs, preserving the model's functionality while removing the unsafe content. Despite the technical contribution, a large-scale **user study** contributes human-centered insights, examining user perceptions of explicit content and informing effective safety alignment strategies.

The post-generation safety filter offers a flexible, add-on defense against adversarial inputs through *provable robustness*, providing strong protection against potential adversarial inputs. My work, CertPHash (under review [1]), introduces the **first certified robust perceptual hashing (PHash) system**, addressing vulnerabilities in systems like Apple's NeuralHash, Microsoft's PhotoDNA, and Facebook's PDQ. These systems detect unsafe content, such as *CSAM* (child sexual abuse material), by matching images to known unsafe hashes but are vulnerable to adversarial examples, leading to two major issues: *evasion*, where unsafe content bypasses detection, and *collision*, where safe content is mistakenly filtered. While provable guarantees for PHash systems are highly desirable, they are challenging to achieve as no certified PHash systems currently exist. Unlike classification models that output a single label from a fixed set–where certified robustness is mainly focused–PHash systems generate unique outputs for every perceptually different image, each being a high-dimensional numeric vector. This uniqueness makes traditional robustness certification methods unsuitable for PHash systems. Additionally, two certified rate–one for evasion and one for collision–are required, rather than a single misclassification rate.

CertPHash addresses these challenges by introducing a certified robust training framework with deterministic bounds tailored for PHash, which incorporates three novel optimization terms: *anti-evasion*, *anti-collision*, and *functionality*. The anti-evasion term establishes an upper bound on the hash deviation from input perturbations, the anti-collision term sets a lower bound on the distance between hashes of different images, and the functionality term ensures that the system remains reliable and effective throughout robust training. In addition, it systematically formulated and addressed the **formal verification problem for PHash systems, setting a benchmark for the certified no-evasion and no-collision rates** that will guide future research on the robustness of PHash systems. Extensive evaluation demonstrates that CertPHash achieves non-vacuous certification for both evasion and collision, offering **provable guarantees** for reliably detecting unsafe content in a real-world deployment.

③ ***Steer: Test-time rule-based reasoning framework.*** *Can we steer powerful Large Language Models (LLMs) to follow our rules and behave as intended? Further, can we extend this capability to solve tasks even across different modalities, using their existing knowledge—much like how humans perform various tasks—without relying on extensive supervised learning with labeled datasets?*

To address the first question, my work RippleCOT (EMNLP'24 [7]) focuses on steering LLM behavior during test time for *complete* knowledge editing, ensuring that **LLM outputs on related facts remain accurate when a single fact is modified due to inaccuracies or sensitivities**. RippleCOT introduces additional structured

Chain-of-Thought (CoT) reasoning as rules to LLM, unlike methods that depend solely on the model's inherent CoT capabilities. By analyzing and restructuring the multi-hop logic based on relationships among facts, RippleCOT helps smaller models to perform comparably to more powerful ones through a structured rule-based approach.

The second question drive my work AnomalyRuler (ECCV'24 [5]), the **first rule-based, training-free framework** for video anomaly detection (VAD). Unlike conventional VAD methods focused on visual feature learning, AnomalyRuler converts visual features into text, leveraging LLMs for accurate and interpretable detection without fine-tuning. However, directly using the general context knowledge embedded in pre-trained LLMs can lead to inflexibility and inaccuracy in specific real-world VAD scenarios. To overcome these challenges, AnomalyRuler operates in two stages: **rule derivation** from a few normal samples and **rule application** to unseen samples, identifying anomalies and providing insights on rule use. AnomalyRuler integrates **robust strategies** to address perception and reasoning errors, including rule aggregation via randomized smoothing, perception smoothing with a novel Exponential Majority Smoothing, and a double-check mechanism for reliable outputs. These methods collaboratively ensure reliable, consistent anomaly detection.

These rule-based reasoning frameworks enhance both the **trustworthiness** and **interpretability** of AI systems, particularly in safety-critical applications where understanding the rationale behind decisions is important. RippleCOT and AnomalyRuler mark important steps toward computation-efficient, trustworthy, and effective AI-driven solutions, ensuring their seamless real-world deployment.

④ ***Addtional achievements in privacy-preserving AI/ML.*** Alongside my work on diagnosing, correcting, and steering unsafe AI content, I have also made contributions to privacy-preserving AI/ML. My research achievements include diagnosing privacy risks through novel membership inference attacks (NDSS'21 [11]) and developing functionally-rich, privacy-preserving ML and federated learning systems (DSN'23 [8], ECCV'22 [10], under review [6]) with provable guarantees by differential privacy (Usenix'23 [9]). My commitment to enhancing AI privacy is ongoing, and I will continue to explore and improve these technologies for better privacy protections in the future.

Future Directions

Foundation: Provable robustness for large language and vision models. As AI models grow larger, applying provable robustness techniques becomes increasingly challenging. Tasks have evolved from simple classification to complex generative tasks. Unlike traditional classification models, where the certification goal is to ensure that the correct label has the highest probability, generative models need more comprehensive certification. This involves verifying the correctness of each output component. Moreover, for generated outputs, there is no single ground truth; various image values may be considered safe. Therefore, different metrics are necessary for certification compared to classification models. For instance, vision models might use LPIPS for visual similarity, language models BERTScore for semantic accuracy, and multimodal models cosine similarity to ensure coherence and safety.

My previous work [1] advances the deployment of deterministic certified robustness via interval bound propagation, moving beyond classification tasks to certify PHash models that produce 256-dimensional numeric vectors. Moving forward, I plan to continue collaborating with experts in certified robustness, and expand collaboration with domain experts in natural language processing and computer vision, to extend these techniques to larger language, vision, and multimodal models. Another promising approach involves probabilistic bounds, such as randomized smoothing. My previous research applies randomized smoothing to LLMs for video anomaly detection [5], treating the target model as a black box with smoothing techniques. While this method offers greater flexibility, it requires an increased number of noised queries during test time. My future work aims to develop more efficient and safety-driven sampling methods, such as importance sampling or variance reduction techniques, to reduce the number of required queries while maintaining or even improving robustness guarantees. Additionally, incorporating context-aware smoothing strategies, like using attention mechanisms to focus noise application on vulnerable parts of the input, can enhance the effectiveness of randomized smoothing for safety tasks.

Through these efforts, my research goal is to make provable robustness more scalable and adaptable for large-scale language and vision models, ensuring that AI systems remain reliable as they continue to advance. I plan to seek

funding for this direction through the Secure and Trustworthy Cyberspace (SaTC) and Safe Learning-Enabled Systems programs under NSF CISE.

Application: Trustworthy AI for social good. My current research is on regulating unsafe content generation [1, 3, 4] and privacy-preserving machine learning [8–11], with the broader goal of advancing AI systems for social good. I view trustworthiness in AI as encompassing robustness, privacy, fairness, functionality, and interpretability, each with inherent trade-offs. While my past work has managed trade-offs among privacy, robustness, and functionality [1, 3, 9], current metrics often rely on *machine-decided yes-or-no* evaluations, shaped by data labeling and algorithm design from technical experts. Although valuable, these metrics may fall short in meeting the safety needs of everyday users, particularly vulnerable populations, whose experiences with AI risks are often distinct.

I believe AI trustworthiness should go beyond yes-or-no metrics. It's not just about avoiding misclassifications or flagging content as a binary problem, i.e., safe or unsafe; it's about understanding the broader impact of AI on individuals. This requires a deeper understanding of human behavior, which can redefine what we constitute a "trustworthy" AI. In future work, I am interested in **social cybersecurity**—specifically in identifying fraud and misinformation on social media, where risks like catfishing and manipulation have increased as large language models make deceptive content generation more accessible. Social media, where people openly express thoughts and emotions, offers an ideal environment for exploring **user-decided trustworthiness** and advancing **digital mental health**. By studying how users interact with information, assess credibility, and form beliefs—especially when faced with red flags like inconsistent information and emotional manipulation–we can gain valuable insights into human decision-making processes towards trust or doubt, design interventions to promote healthier online interactions, mitigate psychological harm, and refine metrics for AI trustworthiness.

I look forward to collaborating with experts in HCI, social sciences, psychology, and public health, as well as engaging with diverse user groups. My goal is to unify various dimensions of trustworthiness into a comprehensive framework that aligns AI systems with users' needs while promoting social good. I plan to expand my diagnose-correct-steer framework by incorporating fairness, interpretability, and user-defined metrics that address real-world concerns. An aggregation-based approach enables both independent and integrated evaluations, highlighting trade-offs and providing insights for real-world applications. I plan to apply for the Cybersecurity Innovation for Cyberinfrastructure (CICI) program and Human-Centered Computing (HCC) by NSF CISE, and pursue industry funding opportunities including the Amazon and Meta Research Awards.

Long term vision: Society-centered trust standards. My long-term vision is to establish comprehensive, society-centered trust standards and integrate them into AI models. These standards will prioritize societal concerns and provide a rigorous framework to address the complex demands of real-world AI applications. Inspired by the success of my rule-based reasoning framework [5, 7], my research aims to bridge the gap between society's genuine concerns and AI systems. Currently, trust standards for AI are vague—how can we enforce trustworthiness when it's not clearly defined? Well-defined standards are essential for measuring and ensuring AI trustworthiness. These standards should be **verifiable by experts**, being clear and sound in terms of law, policy, and ethics; **acceptable to users** by defining what makes them feel truly unsafe; and **feasible for developers** to integrate into AI systems using structured data with informative content. These trust standards could benchmark key challenges in trustworthy AI–like robustness, fairness, and privacy–while prioritizing research on the most critical societal concerns.

I plan to collaborate with experts in law, policy, AI ethics, and cognitive science to develop these trust standards, actively incorporating user feedback to refine them. As both a researcher and system developer, I aim to create a framework based on my diagnose-correct-steer approach to embed these standards throughout the AI lifecycle. Before training, data will be filtered according to these standards; during training, standards-based feedback will guide model behavior via supervised, reinforcement, or retrieval-based learning; and at inference, standards will serve as prompts to steer outputs, with techniques like randomized smoothing ensuring robustness even in adversarial contexts. This approach will help create AI that is both powerful and responsibly aligned with society-centered trust standards. I will seek funding from the Institute for Trustworthy AI in Law & Society (TRAILS) with NSF IIS and Robust Intelligence (RI) programs with NSF CISE.

References

[1] **Yuchen Yang**, Qichang Liu, Christopher Brix, Huan Zhang, Yinzhi Cao, "Certphash: Towards certified perceptual hashing via robust training," in *Proceedings of the USENIX Security Symposium (Usenix)*, 2025.

[2] Zhengyuan Jiang, Yuepeng Hu, **Yuchen Yang**, Yinzhi Cao, Neil Gong, *Jailbreaking safeguarded text-to-image models via large language models*, Under Review, 2024.

[3] Xinfeng Li*, **Yuchen Yang***, Jiangyi Deng*, Chen Yan, Yanjiao Chen, Xiaoyu Ji, Wenyuan Xu, "Safegen: Mitigating sexually explicit content generation in text-to-image models," in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2024.

[4] **Yuchen Yang**, Bo Hui, Haolin Yuan, Neil Gong, Yinzhi Cao, "Sneakyprompt: Jailbreaking text-to-image generative models," in *Proceedings of the IEEE Symposium on Security and Privacy (SP)*, 2024.

[5] **Yuchen Yang**, Kwonjoon Lee, Behzad Dariush, Yinzhi Cao, Shao-Yuan Lo, "Follow the rules: Reasoning for video anomaly detection with large language models," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[6] Zihao Zhao, Yijiang Li, **Yuchen Yang**, Wenqing Zhang, Nuno Vasconcelos, Yinzhi Cao, *Pseudo-probability unlearning: Towards efficient and privacy-preserving machine unlearning*, Under Review, 2024.

[7] Zihao Zhao, **Yuchen Yang**†, Yijiang Li, Yinzhi Cao, "Ripplecot: Amplifying ripple effect of knowledge editing in language models via chain-of-thought in-context learning," in *Findings of the Association for Computational Linguistics: Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[8] **Yuchen Yang**, Haolin Yuan, Bo Hui, Neil Gong, Neil Fendley, Philippe Burlina, Yinzhi Cao, "Fortifying federated learning against membership inference attacks via client-level input perturbation," in *2023 53rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, 2023.

[9] **Yuchen Yang***, Bo Hui*, Haolin Yuan*, Neil Gong, Yinzhi Cao, "Privatefl: Accurate, differentially private federated learning via personalized data transformation," in *Proceedings of the USENIX Security Symposium (Usenix)*, 2023.

[10] Haolin Yuan*, Bo Hui*, **Yuchen Yang***, Philippe Burlina, Neil Zhenqiang Gong, Yinzhi Cao, "Addressing heterogeneity in federated learning via distributional transformation," in *To appear in proceedings of the European Conference of Computer Vision (ECCV)*, 2022.

[11] Bo Hui*, **Yuchen Yang***, Haolin Yuan*, Philippe Burlina, Neil Zhenqiang Gong, Yinzhi Cao, "Practical blind membership inference attack via differential comparisons," in *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2021.

---

* indicates co-first authors.

† indicates the first author finished the paper mainly under my mentoring.