Research Statement

Yun-Yun Tsai

My research focuses on making AI systems robust and safe beyond training. As foundation models are increasingly deployed in real-world applications, relying on offline retraining to address emerging failures is often too slow and with high computational cost. Beyond training, AI systems encounter unforeseen inputs and adversarial perturbations that undermine their reliability during inference time. As models are further adapted, unauthorized reuse and modifications make it difficult to trace their provenance and ownership. Moreover, as generative models produce increasingly realistic content and perform complex tasks through multiple interactions, identifying synthetic content and evaluating whether generated outputs behave as intended become more challenging. These challenges motivate my PhD research on improving the robustness, accountability, and security of AI systems throughout their deployment lifecycle.

My research philosophy is grounded in a simple principle: "The little things are infinitely the most important." In AI systems, subtle signals in model predictions, generated outputs, and execution behavior can reveal failures and vulnerabilities that are otherwise difficult to detect. Guided by this principle, my research addresses three complementary directions. First, I develop **test-time input adaptation** methods for detecting and mitigating deployment-time failures (e.g., adversarial examples, out-of-distribution data). My approach is to use prediction uncertainty, learned self-supervised representations, and generative priors to identify distribution shifts. The main advantage is that these methods can restore reliable predictions under new conditions without requiring labeled target data or retraining the entire mode. Second, I develop **behavioral fingerprinting** method to generate rare input–output patterns and use them to verify whether a black-box model is derived from a protected model. Our proposed fingerprinting framework is non-invasive, and can succefully verify the model provenance even after fine-tuning, quantization, or prompt change. Third, to **improve the trustworthiness of agent systems**, I leverage visual and temporal cues to detect AI-generated videos, and use executable interaction tests and agent feedback to evaluate the functional correctness of generatedn WebUI applications and guide model improvement. These efforts contribute to my broader research vision of developing AI systems that can recognize emerging failures, adapt to unforeseen conditions, and maintain safety and reliability throughout their deployment.

My research has made practical contributions to the robustness, security, and reliability of AI systems through three research thrusts. First, I developed efficient methods to improve model robustness under unseen distribution shifts and adversarial perturbations, with representative work published at NeurIPS 2023 and CVPR 2023–2024 [1, 2, 3]. This research has received support from *GE/DARPA, Google, and Accenture*, reflecting interest in robust AI deployment from government and industry. Second, I developed RAFP [4], a non-invasive, black-box framework for verifying model provenance and ownership despite downstream modifications, building on my earlier research on model stealing attacks [5] (NDSS 2020). Third, I developed methods to detect AI-generated videos and evaluate the functional correctness of generated web applications, including REVEAL [6] (ICCV 2025 KnowledgeMR workshop) and WebUIProof [7], which help identify synthetic content and failures that conventional evaluation methods may overlook. My broader research projects, such as model reprogramming [8, 9] (ICML 2020–2021), captcha reasoning (AACL 2026) [10], CodeQL vulnerability detection [11], etc [12, 13] have also informed subsequent research in robust adaptation and visual understanding.

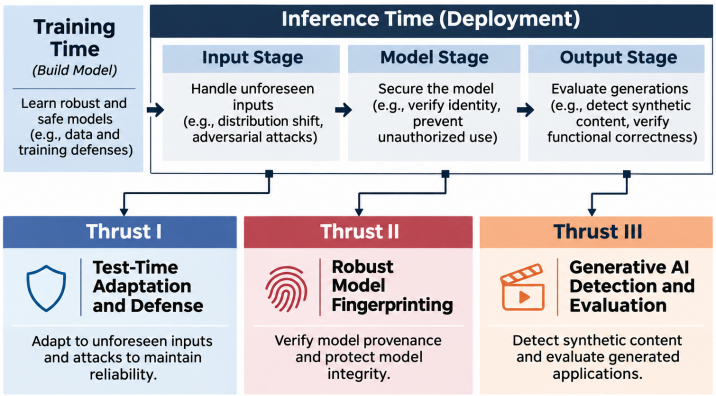


Figure 1: Safeguarding AI across the model cycle.

Thrust I: Test-Time Adaptation and Defense Against Unforeseen Shifts

Modern AI systems are often deployed in environments that differ substantially from their training conditions. Distribution shifts, noise, corruptions, and adversarial perturbations can cause severe prediction failures. Existing defense methods largely rely on **training-time robustness**, requiring models to anticipate and learn from possible failure modes before deployment (e.g., adversarial training). However, many real-world shifts and attacks are unforeseen, and retraining may be costly or infeasible once a model is deployed. My work addresses this gap by developing **test-time intervention methods** that use signals exposed during inference to detect unreliable behavior and adapt model predictions without requiring labeled test data or full model retraining.

• Lightweight Test-Time Adaptation from Self-Supervised Signals.

My first work asks whether a deployed model can adapt to unfamiliar inputs without updating its parameters. I developed **Convolutional Visual Prompt (CVP)** [1], a self-supervised and label-free method that learns a lightweight transformation over test inputs rather than modifying the underlying vision model. The key insight is that self-supervised consistency provides an observable signal of whether an input is compatible with the model's learned representation. CVP uses this signal to optimize a compact convolutional prompt at inference time, allowing the model to recover from out-of-distribution shifts while keeping the backbone fixed. This work demonstrated that lightweight input-space intervention can outperform conventional test-time adaptation baselines

while substantially reducing adaptation cost.

• **Generative Priors for Robust Test-Time Adaptation.**

While self-supervised signals can indicate that a model is uncertain, they may provide limited guidance about how a corrupted input should be corrected. To address this limitation, I developed **Generalized Diffusion Adaptation (GDA)** [2], which uses diffusion models as structured generative priors for test-time adaptation. The central insight is that diffusion models encode complementary information about both semantic content and appearance, which can be extracted to guide adaptation under previously unseen shifts. GDA combines prediction-based objectives with diffusion-derived structural and style signals to refine model behavior at inference time, achieving strong robustness across diverse out-of-distribution conditions. In parallel, my work on **compositional adversarial robustness** [3] showed that models robust to individual perturbations can still fail under combinations of semantic attacks, highlighting the need for test-time defenses that reason about more complex and previously unseen failure modes.

Thrust II: Behavioral Fingerprinting for Model Provenance

While robust AI research traditionally asks how to protect models from malicious or unexpected inputs, an equally important deployment-time challenge is how to **protect the model itself**. Once a foundation model is released through APIs, shared with downstream users, or adapted by third parties, it can be copied, fine-tuned, quantized, or redistributed without authorization. Existing protection mechanisms often rely on access to model parameters, embedded watermarks, or modifications during training, making them difficult to apply when only black-box access is available or when the model has been substantially transformed. My work, **RAFP : Identifying LLM Lineages via Rare-Region Fingerprints** [4] addresses this challenge through behavioral verification, using distinctive input–output patterns to establish model provenance and ownership without accessing model parameters. I developed RAFP, a black-box fingerprinting framework that identifies prompt–response pairs in low-probability, rare-token regions, where model behaviors are highly distinctive yet resilient to downstream modifications. Starting from rare-token responses, RAFP optimizes prompts to elicit behaviors that distinguish the target model from independently trained models. Beyond empirical validation, I provide a theoretical analysis of fingerprint persistence, showing that sufficiently distinctive rare-region behaviors remain separable from unrelated models when modifications induce bounded changes in response scores. This explains RAFP's robustness under supervised fine-tuning, LoRA adaptation, quantization, and system-prompt changes. Together, these results enable reliable model provenance verification with only a small number of black-box queries, providing a practical deployment-time safeguard against unauthorized model copying, adaptation, and redistribution.

Thrust III: Generative AI Detection and Agentic Evaluation

As generative AI systems become increasingly capable of producing realistic content and performing complex tasks, ensuring their reliability requires more than inspecting their final outputs. AI-generated videos may appear visually convincing while containing subtle inconsistencies in motion, geometry, or temporal coherence. Similarly, AI-generated applications may render correctly but fail to respond to user interactions or satisfy functional requirements. These challenges expose a common limitation of existing evaluation methods: surface-level features, visual similarity, and static checks often fail to capture whether generated content is physically consistent or generated systems behave correctly. My research addresses this gap by developing **evidence-driven detection and evaluation methods** that combine multiple sources of evidence to identify failures that are difficult to detect through output inspection alone.

• **Detecting AI-generated content through complementary evidence.** Existing synthetic-content detectors often rely on visual artifacts or patterns learned from specific generation models, which may not generalize to new generators or increasingly realistic content. In my REVEAL work [6], I investigate how complementary signals, including optical flow, depth, edges, and temporal consistency, can provide additional evidence of synthetic manipulation. Rather than treating these signals independently, my approach uses vision-language models to reason over multiple forms of evidence and identify inconsistencies that may be overlooked by individual detectors. A key research challenge is determining which evidence is most informative for a given video, since different generation failures may manifest through different visual or temporal cues. By dynamically selecting relevant evidence and refining the detection process, I aim to develop more interpretable and generalizable detection methods that can adapt to evolving generative models.

• **Evaluating generative and agentic systems through execution.** As language models increasingly generate interactive applications and execute multi-step workflows, evaluating their outputs requires examining what systems actually do rather than how they appear. For example, a generated web application may compile successfully and closely match a reference screenshot while containing broken interactions, incorrect state transitions, or incomplete functionality. To address this limitation, I create **WebUIProof** [7], a WebUI benchmark with execution-based evaluation frameworks that translate functional requirements into actionable tests and use UI agents to interact with generated applications, inspect resulting states, and identify behavioral failures. This approach provides more informative feedback than static code analysis or visual comparison alone. Beyond measuring task success, I am interested in identifying the underlying causes of failures, distinguishing implementation errors from incorrect reasoning or incomplete task execution, and using these findings to guide targeted improvements. Such feedback can also support model optimization by providing training signals grounded in observed execution outcomes rather than subjective assessments of output quality.

Although synthetic-content detection and agentic evaluation address different applications, both require reasoning over evidence

that is not immediately apparent from the final output. My long-term goal is to develop evaluation methods that can identify relevant evidence, verify whether expected properties hold, and explain why failures occur. By connecting detection, execution-based verification, and failure diagnosis, this research aims to establish more reliable ways to assess generative AI systems and provide actionable feedback for improving their safety and robustness.

Future Directions

My long-term research vision is to develop AI systems that can recognize unseen threats, develop effective defenses, and maintain safety as their capabilities and environments evolve. My past research on test-time input adaptation, model fingerprinting, and generative AI evaluation provides a foundation for this goal. Building upon these contributions, I plan to pursue three connected directions from short to long terms: (1.) understanding and defending against emerging attacks on agentic systems, (2.) developing self-evolving AI safeguards, and (3.) extending these capabilities to embodied AI in real-world environments. These three directions span from identifying new safety vulnerabilities to developing systems that can proactively improve their own defense mechanisms and ultimately operate safely in complex physical environments.



Figure 2: Future research direction from emerging agentic AI threats to self-evolving safeguards and safe embodied AI.

• **Near-term Direction: Emerging Attacks and Defenses for Agentic AI Systems.** As AI systems evolve from standalone models into autonomous agents capable of reasoning, using tools, maintaining memory, and interacting with external environments, they introduce new attack surfaces that cannot be fully addressed by conventional model-level defenses. An adversary may manipulate intermediate observations, exploit interactions among multiple tools, or decompose harmful objectives into seemingly benign subtasks to bypass existing safeguards. These attacks are particularly challenging because unsafe behavior may emerge from a sequence of individually harmless actions rather than a single malicious instruction.

My near-term research will investigate new attack mechanisms and defense strategies for agentic AI systems, focusing on vulnerabilities that arise from multi-step reasoning, tool use, memory, and interactions with external environments. Building on my ongoing work on covert misuse detection, I will study how harmful objectives can be concealed across multiple subtasks and how their underlying intent can be recovered through structured reasoning over task dependencies and interaction trajectories. Beyond detecting known attacks, I aim to systematically identify previously unexplored vulnerabilities and develop defenses that reason about the relationships among observations, actions, and intermediate states. The goal is to establish a deeper understanding of how agentic systems can be compromised and how emerging attacks can be detected and mitigated before they lead to unsafe outcomes.

• **Mid-term Direction: Self-Evolving Safeguard and Automated Defense Discovery** While identifying emerging attacks is essential, developing a separate defense for every new vulnerability is difficult to sustain as AI capabilities and attack strategies evolve. Current safeguards often depend on manually designed rules, fixed safety classifiers, or additional training on known attacks, limiting their ability to respond to unforeseen threats. My mid-term research will investigate whether AI systems can automatically discover new defense strategies, evaluate their effectiveness, and improve their own safeguards with limited human intervention.

A central direction is to leverage world models to anticipate potential failures before they occur. By learning or simulating how an agent's actions influence future states, a world model could help predict unsafe trajectories, explore possible adversarial behaviors, and evaluate alternative defensive interventions. Building on this capability, I plan to develop methods that allow AI systems to generate candidate defense algorithms, test them against simulated and real attack scenarios, and iteratively refine their safeguards using execution feedback. This requires addressing fundamental challenges in automated defense discovery, including how to search over possible defense mechanisms, distinguish genuine improvements from overfitting to simulated attacks, and verify that newly proposed defenses do not introduce additional vulnerabilities.

Beyond language-based agents, I aim to establish a general framework for self-evolving safeguards across multimodal AI systems, where risks may arise from interactions among text, images, video, audio, and external observations. Rather than designing independent defenses for each modality, I will investigate shared representations of system behavior, safety constraints, and potential failure trajectories that enable defensive knowledge to transfer across different models and environments. Ultimately, this direction aims to move AI safety from reactive, manually designed defenses toward proactive and continuously improving safeguards that can anticipate emerging risks and develop new protection mechanisms.

• **Long-term Direction: Safe and Adaptive Embodied AI for Real-World Deployment** My long-term goal is to extend my research on AI safety and adaptive defenses from digital systems to embodied agents and robots operating in the physical world. Unlike software agents, embodied systems must make safety-critical decisions under incomplete observations, limited computational resources, and real-time constraints, where failures can have irreversible consequences. I envision developing safety mechanisms that combine predictive world models, adaptive safeguards, and continuous behavioral verification to help embodied agents anticipate

risks and respond to unforeseen situations. A key challenge is transferring safety mechanisms developed for foundation models and software agents to physical systems with different sensing, action, and computational constraints.

Ultimately, I aim to build toward AI systems that can recognize emerging threats, adapt their defenses, and maintain reliable behavior as they interact with increasingly complex environments.

References

[1] **Yun-Yun Tsai**, Chengzhi Mao, and Junfeng Yang. "Convolutional visual prompt for robust visual perception". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 27897–27921.

[2] **Yun-Yun Tsai**, Fu-Chen Chen, Albert YC Chen, Junfeng Yang, Che-Chun Su, Min Sun, and Cheng-Hao Kuo. "Gda: Generalized diffusion for robust test-time adaptation". In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2024, pp. 23242–23251.

[3] Lei Hsiung**, Yun-Yun Tsai**, Pin-Yu Chen, and Tsung-Yi Ho. "Towards compositional adversarial robustness: Generalizing adversarial training to composite semantic perturbations". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2023, pp. 24658–24667.

[4] **Yun-Yun Tsai**, Jia Hao Liang, Chuan Guo, Junfeng Yang, and Laurens van der Maaten. *RAFP: Identifying LLM Lineages via Rare-Region Fingerprints*. 2026. arXiv: 2505.12682 [cs.LG]. URL: https://arxiv.org/abs/2505.12682.

[5] Honggang Yu, Kaichen Yang, Teng Zhang**, Yun-Yun Tsai**, Tsung-Yi Ho, and Yier Jin. "CloudLeak: Large-Scale Deep Learning Models Stealing Through Adversarial Examples". In: *Proceedings 2020 Network and Distributed System Security Symposium* (2020). URL: https://api.semanticscholar.org/CorpusID:211268411.

[6] **Yun-Yun Tsai**, Qingyuan Liu, Ruijian Zha, Victoria Li, Pengyuan Shi, Chengzhi Mao, and Junfeng Yang. *REVEAL: Robust Evolution of Vision-Language Models for Explainable AI-Video Detection*. 2026. arXiv: 2502.14994 [cs.CV]. URL: https://arxiv.org/abs/2502.14994.

[7] **Yun-Yun Tsai**, Yuning Mao, Shiqi Wang, Junfeng Yang, and Sinong Wang. *WebUIProof: Benchmarking WebUI Code Generators with UI-Agent Execution Harness*. 2026. arXiv: 2610.02617 [cs.SE]. URL: https://arxiv.org/abs/2610.02617.

[8] **Yun-Yun Tsai**, Pin-Yu Chen, and Tsung-Yi Ho. "Transfer Learning without Knowing: Reprogramming Black-box Machine Learning Models with Scarce Data and Limited Resources". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by H. D. III and A. Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 13–18 Jul 2020, pp. 9614–9624. URL: https://proceedings.mlr.press/v119/tsai20a.html.

[9] Chao-Han Huck Yang**, Yun-Yun Tsai**, and Pin-Yu Chen. "Voice2series: Reprogramming acoustic models for time series classification". In: *International conference on machine learning*. PMLR. 2021, pp. 11808–11819.

[10] Python Song, Luke Tenyi Chang**, Yun-Yun Tsai**, Penghui Li, and Junfeng Yang. *Reasoning under Vision: Understanding Visual-Spatial Cognition in Vision-Language Models for CAPTCHA*. 2025. arXiv: 2510.06067 [cs.CV]. URL: https://arxiv.org/abs/2510.06067.

[11] Chunyi Wang, Yunfei Ke, Junfeng Yang**, Yun-Yun Tsai**, and Penghui Li. *ARQ: Agentic CodeQL Query Refinement for C/C++ Vulnerability Detection*. 2026. arXiv: 2608.20637 [cs.CR]. URL: https://arxiv.org/abs/2608.20637.

[12] **Yun-Yun Tsai**, Ju-Chin Chao, Albert Wen, Zhaoyuan Yang, Chengzhi Mao, Tapan Shah, and Junfeng Yang. "Test-time detection and repair of adversarial samples via masked autoencoder". In: *arXiv preprint arXiv:2303.12848* (2023).

[13] Qingyuan Liu, Pengyuan Shi**, Yun-Yun Tsai**, Chengzhi Mao, and Junfeng Yang. *Turns Out I'm Not Real: Towards Robust Detection of AI-Generated Videos*. 2024. arXiv: 2406.09601 [cs.CV]. URL: https://arxiv.org/abs/2406.09601.