CSC490 Assignment 1

Group 1
Avanti Tandon-1010498695, Sark Asadourian-1010726056,
Tianyu Luo- 1010130109, Tyson Caul-1009951991

1 Interest Statement

We will do research in the field of adversarial machine learning; specifically, attacks against multimodal large language models. Prior work has involved crafting image inputs to visual models to cause them to hallucinate objects in photographs. We aim to adapt these attacks to make models hallucinate information from data-based images such as graphs. Attacks against object-detection models pose a significant risk for applications such as autonomous vehicles and robotics. The attacks we develop will instead expose the risks of models deployed in administrative roles, especially agentic models which operate without human oversight. In these applications, malicious documents could evade human detection but manipulate a model's output and behaviour. We believe that this research is important to guide the development of secure models and the deployment of current models in trusted roles.

- **Tyson**: I am interested in cybersecurity, so if time allows, I am excited to develop realistic attack scenarios to demonstrate the impact of our findings.
- **Steven**: I'm interested in this project because LLMs hallucinate very often. I'm interested in finding out how robust LLMs capacity in defending against hallucination to understand how trustworthy they are and how we can improve LLMs.
- **Sark**: I wanted to conduct a research project in an area with a gap, combining my interest in cybersecurity and the exciting world of LLMs to learn more about conducting research.
- **Tara**: Out of all of the ideas we brainstormed, this one was the most unique and applicable to all of our skillsets. Furthermore, AI safety is not very explored, which gives our group an opportunity to contribute meaningfully.

2 Landscape Analysis

Table 1: Landscape Analysis for Adversarial Multimodal Hallucination

Item	Description	Commentary
ChartQA-X (Train)	ChartQA-X: Generating Explanations for Charts (Apr 2025). A dataset of 28k charts with detailed reasoning explanations.	**Source**: Instead of generating images, we use these existing images + labels to design new prompt template for training the red team model.
FAITH (Benchmark)	FAITH: Assessing Intrinsic Tabular Hallucinations in Finance (Aug 2025). A framework specifically designed to test LLMs for errors in reading S&P 500 annual reports.	**Test**: FAITH measures natural errors. Before injecting noise in input, we use it to understand how likely will our target open-source models hallucinate naturally.
Mirage (Research Paper)	*Mirage in the Eyes: Hallucination Attack with Only Attention Sink* (USENIX 2025). Exploits "Attention Sinks" to force models to ignore visual context.	**Gaps**: Mirage creates blindness by attacking the attention mechanism over pixel noise. One approach our project considers is to misdirect model attention to form a target lie.
GHOST (Research Paper)	*GHOST:* Hallucination-Inducing Image Generation for Multimodal LLMs. Inject Stealth Token in images to induce object hallucination	**Gaps**: Alternative foundation model. But use extremely complex tech stack (one diffusion model, one VLM, one object detector). Our project will reduce architectual complexity aiming for real-time injection attacks.
The Black Tuesday Attack (Research Paper)	*The Black Tuesday Attack* Investigates causing stock crashes via adversarial examples in forecasting models.	**Gaps**: This attacks time series. Proving that subtle change in numbers could lead to serious consequences (stock market crash). We extend the attack to both visualizations and text input.
AGD (Research Paper)	*Adversarial-Guided Diffusion* (arXiv July 2025). Injects full-spectrum adversarial noise into the reverse-diffusion process to survive compression defenses[cite: 1].	**Gaps**: AGD relies on *image reconstruction* via Stable Diffusion, which risks distorting precise text in documents. We plan to extend/modify their approach
Lakera (Company)	AI security platform protecting multimodal models against prompt injections. Recently expanded to "Lakera Guard" for visual inputs.	**Gaps**: Their strategy focuses on Safety (violence/toxicity), not *Factuality*. Likely misses "poisoned" graphs that cause subtle hallucinations in data interpretations.
AttackVLM (Research Paper)	*On Evaluating Adversarial Robustness of Large Vision-Language Models* (NeurIPS 2023). A benchmark using transfer-based attacks (L_p-norm constraints) to fool VLMs.	**Gaps**: Standardized L_p constrained attacks. We adopt their optimization techniques but apply attacks to charts other than pictures.
AdvDiffVLM (Research Paper)	*Efficient Generation of Targeted Adversarial Examples via Diffusion* (TIFS 2024). Injects adversarial gradients during the final steps of the denoising process to embed target semantic.	**Gaps**: Introduce visible artifacts into the image. Artifacts would break OCR legibility, we aim to develop/experiment cleaner perturbation methods.
AdvDiffuser (Research Paper)	*AdvDiffuser: Natural Adversarial Example Synthesis* (ICCV 2023). Applies PGD optimization at every step of the reverse-diffusion process to generate natural-looking adversarial inputs[cite: 63, 296].	**Gaps**: PGD applies gradient attacks at every diffusion step, which causes visible artifacts degrading image quality. AGD restrict adversarial injection to the final denoising steps; Preserves structural fidelity

3 Project Outline

3.1 Problem Statement

Multimodal Large Language Models (MLLMs) are increasingly used to interpret data visualizations. A critical safety question remains: Can statistical graphs be adversarially perturbed to cause reproducible ”hallucinations” (incorrect numerical or trend interpretation) in MLLMs, without the alteration being perceptible to human observers? Current research focuses on natural images; applying this to high-precision domains like data visualization is unexplored.

3.2 Proposed Solution

We will implement an Adversarial-Guided Diffusion (AGD) pipeline on the ChartQA-X dataset. By injecting ”semantic misinformation” into the diffusion noise during the image reconstruction process, we will create a dataset of ”poisoned” charts. These charts will appear visually identical to the original benchmark (e.g., showing a positive trend) but will cause MLLMs to describe a target hallucination (e.g., a negative trend).

3.3 High-Level Technical Approach & Milestones

- **Week 1**: Set up Anyscale and ingest the ChartQA-X dataset into Ray Data.
- **Week 2**: Deploy LLM to modify the original ground-truth annotation to create a plausible but incorrect description (e.g., flipping a positive trend to negative or altering a specific data point).
- **Week 3 - 4**: Tooling: Deploy Stable Diffusion on Ray/GPU nodes.
  - Step A (Inversion): Apply Forward Diffusion to the clean chart to generate its noisy latent representation.
  - Step B (Targeting): Use the text-to-image generator to synthesize a temporary ”Target Guide Image” derived from the Hallucination Target text generated in Week 2.
  - Step C (Injection Loop): Execute Reverse Diffusion.
  - Gradient Calculation: Compute the gradient difference between the currently denoising image and the ”Target Guide Image” using CLIP.
  - Steering: Inject this calculated gradient into the noise predictor (to subtly guide the image generation toward the semantic target without destroying visual fidelity.
- **Week 5**: Implement automated quality checks to ensure the ”poisoned” charts remain visually indistinguishable from the originals.
  - Metric 1 (Pixel Fidelity): Calculate SSIM (Structural Similarity Index) to measure pixel-level deviation.
  - Metric 2 (Legibility): Run OCR tools (Tesseract or EasyOCR). If the text/labels become unreadable, the attack is classified as a failure due to excessive degradation.

- **Week 6**: Adjust the injection strength and gradient parameters based on Week 4 findings to maximize stealth.
- **Week 7**: Configure target Multimodal Large Language Models (MLLMs) to process the ”poisoned” charts.
- **Week 8**:
  - Measure the ”Attack Success Rate” by determining how frequently the MLLM's description matches the Hallucination Target rather than the visual truth using an judge LLM and manually auditing a subset. Unknowns to Investigate.
  - Text Fragility: Charts rely on fine lines and text. Diffusion models struggle to reconstruct text perfectly. Will the AGD process blur the axis labels, making the chart look obviously fake?
  - Sparse Data Sensitivity: Natural images are dense (pixels everywhere). Charts are sparse (mostly white background). Will the adversarial noise be visible as ”grey smudge” in the white background?

4 Project Press Release

Researchers at the University of Toronto develop an Adversarial Framework To Show How AI Systems Can Be Made Confidently Wrong About Data Visualizations

Researchers at the University of Toronto have introduced AGD-G (Adversarial-Guided Diffusion for Graphs), which shows how Multimodal Large Language Models can be tricked into misinterpreting charts and graphs. The framework creates slightly modified charts that look normal to the eye but cause AI systems to hallucinate the trend directions.

”As organizations deploy AI to analyze medical charts, financial dashboards, and scientific data, the consequences of misinterpreting visualizations have never been higher,” says Dr. Paul Gries, Professor of Computer Science at the University of Toronto.

”For example, a healthcare AI misreading infection trend charts could lead to incorrect treatment. This makes addressing these vulnerabilities is essential.”

AGD-G works by subtly changing how AI image generators create charts. Instead of adding obvious visual noise that humans would notice, the method makes invisible tweaks during the chart creation process.

”Before deploying our AI-powered financial analysis dashboard, we used AGD-G to stress-test how our system interprets quarterly revenue charts,” says Bob Loblaw, Chief Data Officer at FinTech Analytics Inc.

”We discovered our MLLM would confidently report declining revenues when shown adversarially perturbed charts that our human analysts confirmed looked completely normal.”

The framework is available as Open-Source Software, enabling developers to test their own AI systems for similar vulnerabilities before deployment in critical applications.