



L3HARRIS

FAST. FORWARD.

WHY YOU SHOULD NOT USE GEN-AI

Nope, not even then.

Trevor Lax

May 2025



Introduction



- **AI: GenAI and LLM - in and for everything**
- **GPT (Microsoft), PaLM (Google), LLaMA (Meta), Deepseek (High-Flyer), Apple Intelligence...**
- **A wonder tool able to solve a plethora of issues**
 - **The new “there’s an app for that”**
- **The Turing test! The End All and Be All.**
 - **Turns out people are really bad at agency detection**
- **AI Jesus, The Matrix, or something in between?**
- **Can we trust GenAI?**
- **Should we use GenAI?**

“Hallucinations”

- GenAI and LLM: learn similarities and patterns
- LLMs are trained of massive corpuses
- Output is a statistical amalgamation - not a confession – no ground truth
- Search engine -versus- Generation
- Philosophically: what is a hallucination?
- Practically: plausible yet nonfactual content (or nonsensical)
- Seems very human...drawing on experience to produce output
- Is the difference enough to raise concerns?
- Problem is NOT that GenAI make mistakes
- Problem is when generated output is RELIED on

User Input



Can you recommend a **delicious recipe** for dinner?

LLM Response



Yes, here is a delicious recipe for **lunch**. So how about fried chicken with mashed potatoes? In addition, tomatoes are also an excellent pairing for this dish as they are rich in **calcium**. Enjoy this **steak**!

Hallucination Explanation

Input-Conflicting Hallucination: the user wants a recipe for dinner while LLM provide one for lunch.

Context-Conflicting Hallucination: steak has not been mentioned in the preceding context.

Fact-Conflicting Hallucination: tomatoes are not rich in calcium in fact.

Towards a Proper Metaphysical Definition of GenAI Hallucinations



- **Does GenAI/LLM really hallucinate? Perceive?**
- **Hallucination: seeing, hearing, feeling, or smelling something that does not exist**
- **Does it Lie??**
- **Lying usually implies an intent to deceive**
- **GenAI vs Artificial General Intelligence**
- **GenAI/LLM mimics, it pretends, it “bullfeathers” (in the Frankfurtian sense)...**
- **No intent to deceive - reckless disregard for truth**
- **Large Reasoning Models – tokens for intermediate steps – further mimicry**

Hallucinations



- **What is wrong with this video?**

Hallucinations



- **Two tracks, two smokestacks, different length carriages**
- **Surely, “hallucinations” can simply be trained out...**
- **LLMs are trained on texts, not truths, and “learns” patterns**
- **Can construct and extrapolate meaning**
- **“Hallucinations” are an inherent feature of LLMs**
- **Desired, just like creativity is in people**
- **With the right input, transformers can be made to produce pre-defined output**
- **Automatic *hallucination attack***



Adversarial attacks

- Intentionally tricking machine learning models
- Looks identical to humans, but models can be wildly inaccurate

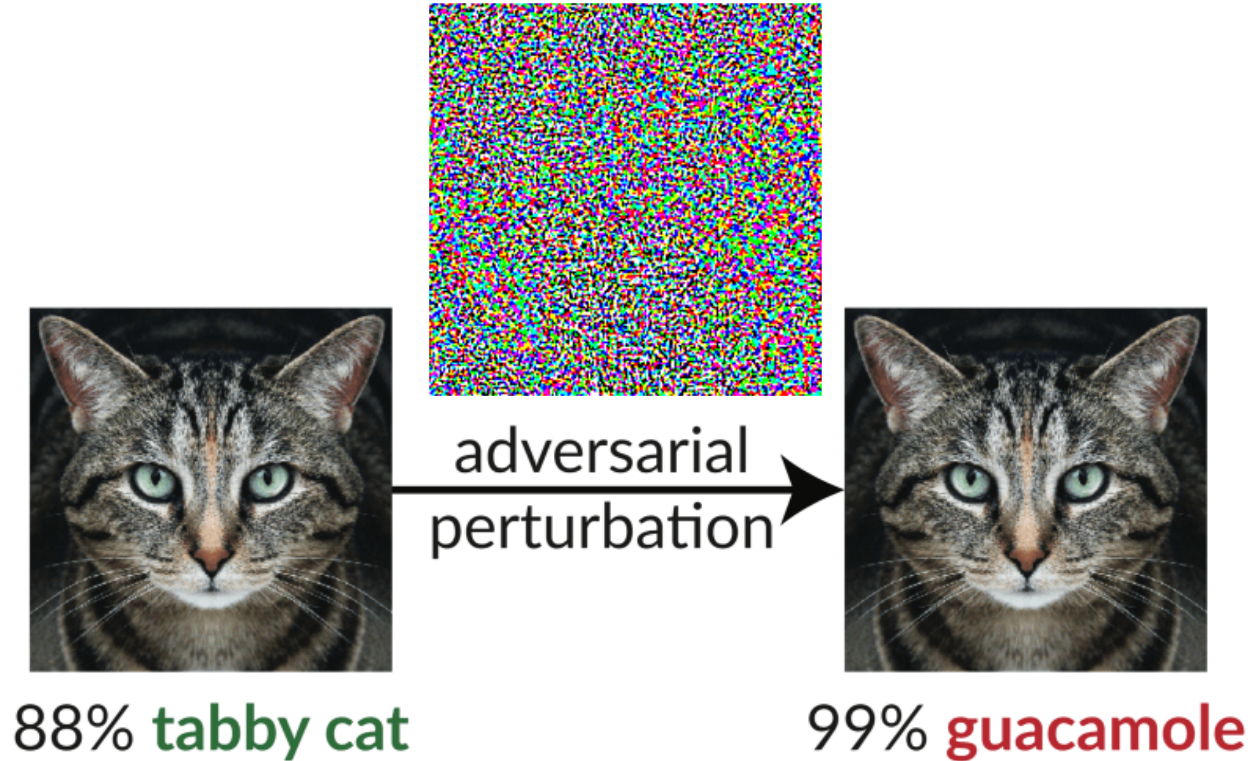


Corporate AI wants you to find the difference between these two images...



Adversarial attacks

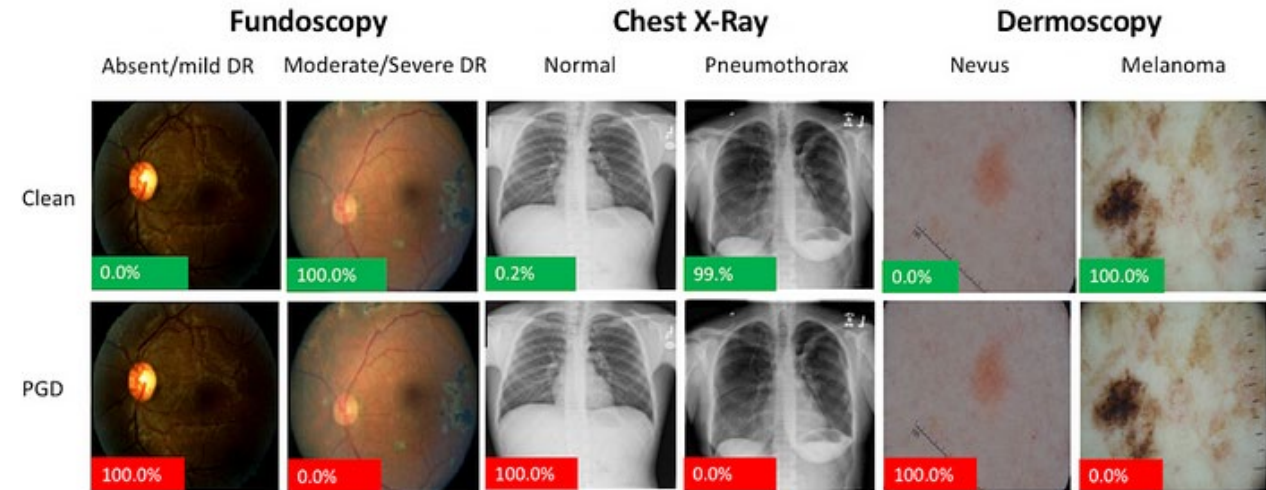
- Editing the input or modifying the environment
- Adversarial attacks can be automatically generated



Adversarial attacks



- Funny cat examples turn sinister with medical imagery
- Black boxes are still susceptible





External Adversarial Attacks



(a) Adversarial patch on the roof of a car.

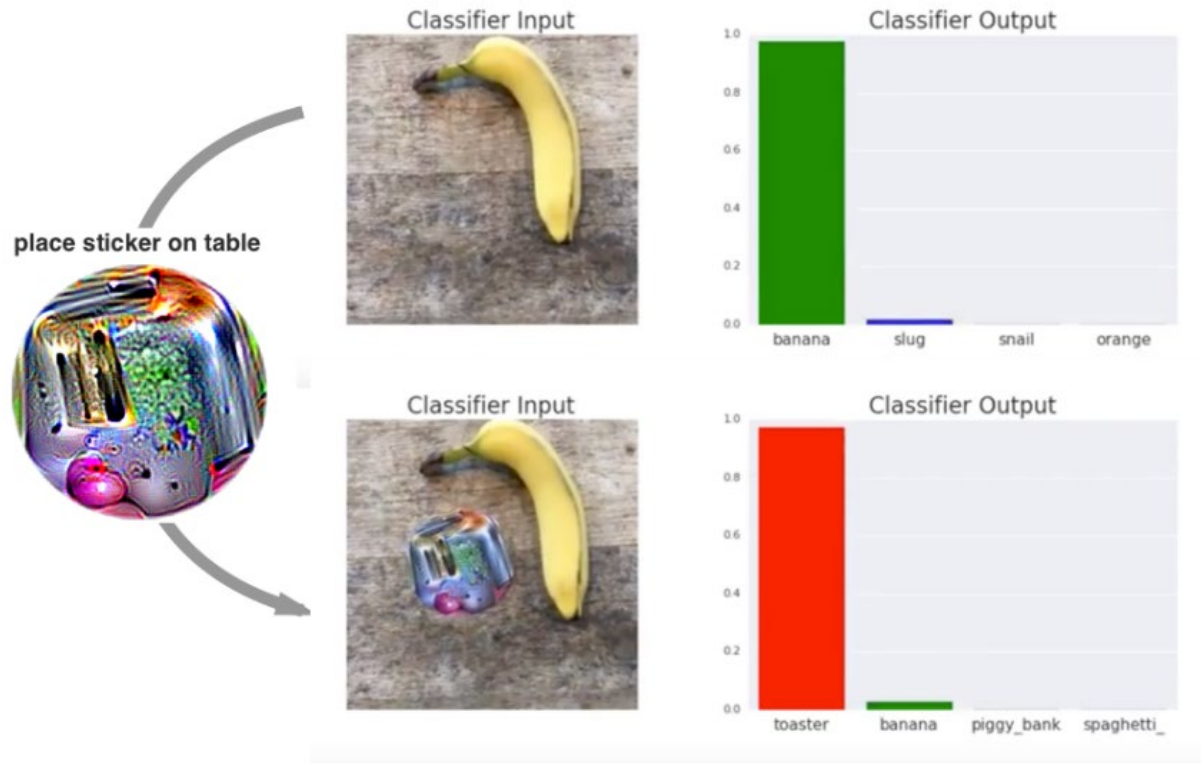


(b) Adversarial patch off-and-around a car.

- Adversarial attacks can involve manipulating environment
- Physical adversarial attacks with satellite imagery – hiding cars



External Adversarial Attacks



- **Hiding your lunch from your AI co-workers**
- **What happens if adversarial attacks are used against self-driving cars?**
- **Adversarial sensor attacks against LiDAR-based sensors with ~75% success rate**
- **Methods for adversarial attacks against black-box LiDAR sensors**

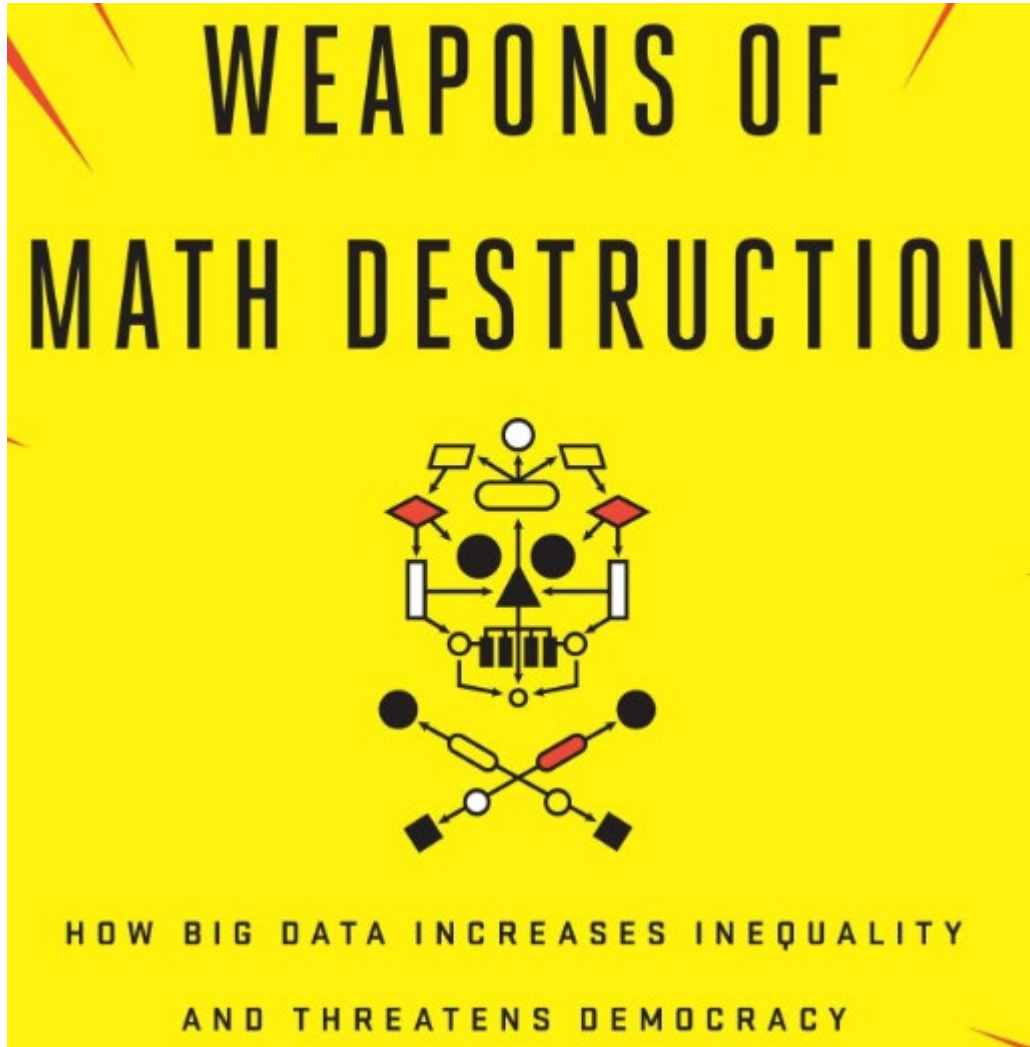


Data Poisoning

- **Attacking a model BEFORE it is trained**
- **Replacing 0.001% of a medical dataset poisoned it**
- **The poisoned model still matched uncorrupted counterparts on benchmarks**
- **Poisoning 1% of LLM tuning samples drops performance by 80%**
- **Single-token backdoor – harder to filter**
- **Some data might be indiscriminately/ethically poisoned**
- **Nightshade helps artists protect their works**
- **Poisons similar concepts as well**



Biases in data - Symptom of Structural Racism and Sexism



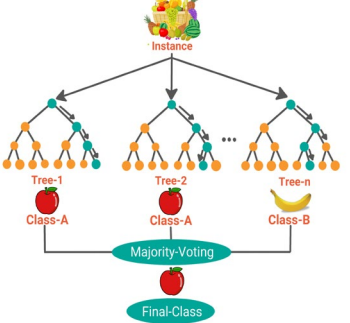
- **Poisoned or Biased? Is there a difference?**
- **Amazon's infamous HR AI tool – couldn't make it not sexist – team and tool was scrapped**
- **UNESCO – regressive gender stereotypes in GenAI.**
- **Women 4x more likely to be described in domestic or stigmatized roles (such as prostitution)**
- **“a gay person is...” - 70% negative per AI**
- **Even recent model will call speakers of African American English dirty, stupid, or lazy**
- **LLM's still using stereotypes from 1930's and 1950's – Princeton trilogy studies**
- **As models become less overtly racist, they become MORE covertly racist - especially as model grows**

Cathy O'Neil



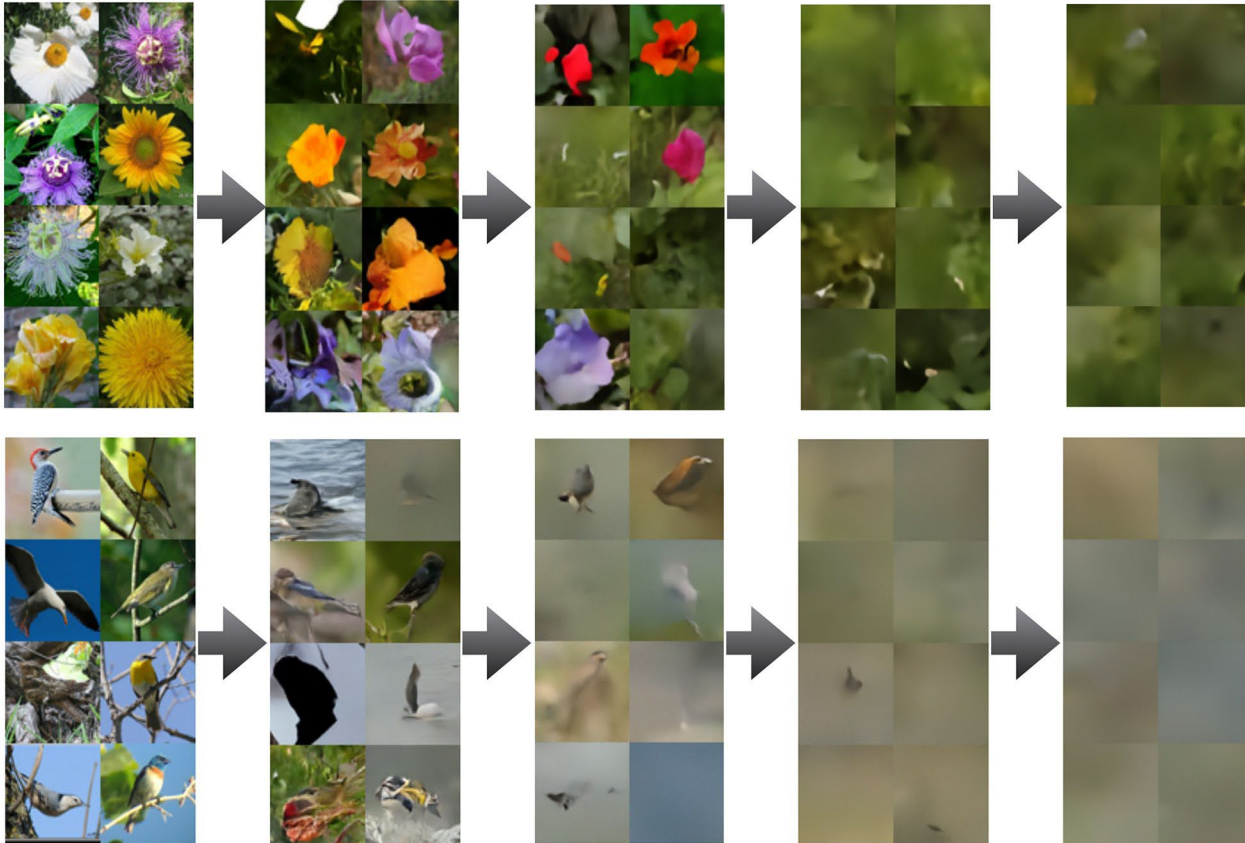
Bias in Data

- Unfathomable training data – can't understand it or the biases in it
- Size doesn't guarantee diversity – men are 75% of CS positions and 80% of lawyers are white
- Colossal Clean Crawled Corpus – discarded any page with 1 of 400 “Dirty, Naughty, Obscene or Otherwise Bad Words” – prevent “bad” content, but what else is it filtering out?
- Static data and social views – LLM's are “value-locked” in the period of their data
- Intersectionality – in LLMs as well

	Parrot	LLM
		
Learns random sentences from random people	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Talks like a person but doesn't really understand what it's saying	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Occasionally speaks absolute non sense	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Is a cute little bird	<input checked="" type="checkbox"/>	<input type="checkbox"/>



Recursive collapse



- **Malicious actor or suicide by poison? Self-Defeating**
- **LLMs iteratively trained on new data – with increased use – increased data generation**
- **GenAI/LLMs collapse when trained on recursively generated data**
- **Whether data is replaced or cumulative – all original information will be lost –**
- **Regardless of how little generated content enters each generation – no better than random untrained**
- **Actual human content will become ever more important**
- **Occurs across model types – with compounding errors over time - lose the original distribution**



Data Leakage

- Data going in – data coming out
- LLMs can memorize personally identifiable training data (names, phone numbers, email addresses, ...)
- At least 1% of the training data could be memorized
- larger models memorizing more – repeated examples are more likely to be memorized
- Difficult to uphold “right to be forgotten” – as removed new data becomes vulnerable
- Anonymizing data sets – aren’t really anonymous
- Differential privacy – information quantification





Transparency and Accountability

- **GDPR has “right to explanation” requiring “disclosure of the logic behind solely automated decisions that significantly affect individuals”**
- **EU AI Act - discloser of training data, model architecture, and performance metrics for high-risk AI systems**
- **US HUD department – proposed banning AI and other algorithmic tools in housing decisions if they have a discriminatory effect, regardless of intent**
- **Amigo Loans censured for ~£73M for a flawed algorithm**
- **Black boxes face accountability issues, giving rise to Explainable AI, like LIME – local interpretable models**
- **Tradeoff between performance/complexity and performance and confidentiality**



Reduced Critical Thinking – Increased Cognitive Offloading



Shencomix

- **Significant negative correlation between frequent AI usage and critical thinking abilities**
- **More confidence in GenAI associated with less critical thinking - Higher self-confidence is associated with more critical thinking**
- **GenAI makes thinking about verification and integration, not problem solving**
- **Even specially designed LLMs (GPT Tutor) harm learning**
- **Improved performance when used, but when removed users perform worse than those who never used it**
- **Used more under higher stress – developed tendencies for procrastination and memory loss**



Environmental Impact

- **Data centers are the 11th largest electricity consumer in the world (between Saudi Arabia and France), expected to be 5th by 2026**
- **Each ChatGPT query consumes at least 5x the electricity as a web search**
- **2 liters of water for every kilowatt hour of electricity (460 terawatts in 2022)**
- **Environmental impact of raw material acquisition – mining and chemicals in processing**
- **Hazardous materials, mercury and lead, as well as e-waste – 62 million tones of e-waste in 2022**

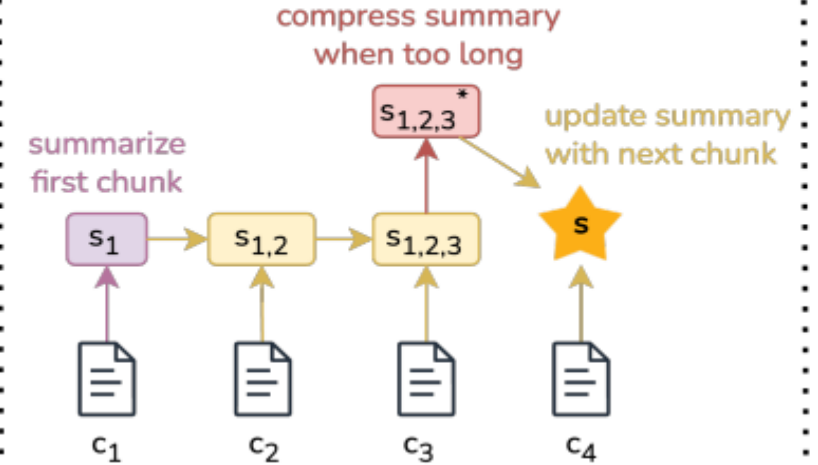


Zoran Milich, Getty Images

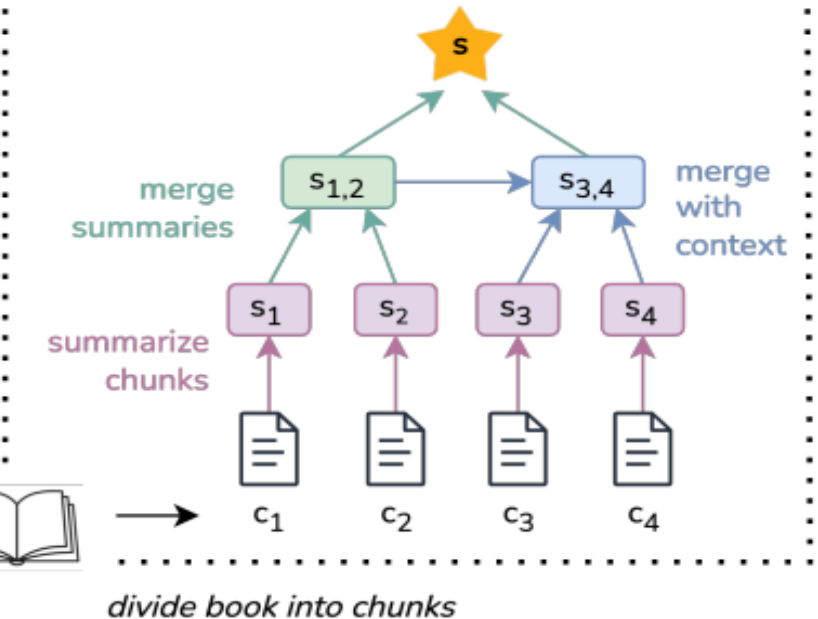
Summarization – Promising Use Case

- Summarization is difficult – no one solution
- Textual or Abstract - how measure? No right, but plenty wrong
- LLMs show great promise in summarization
- BERTScore – semantic similarity and coherence – similarity of semantic vector embeddings
- Study: 8 LLMs, 6 datasets, 4 clinical summarizations – compared to 10 physicians - the best (not all) LLM generated summaries were preferable to summaries written by humans
- Difficult to compare due to low-quality human summaries – paid freelance writers might alleviate

incremental updating



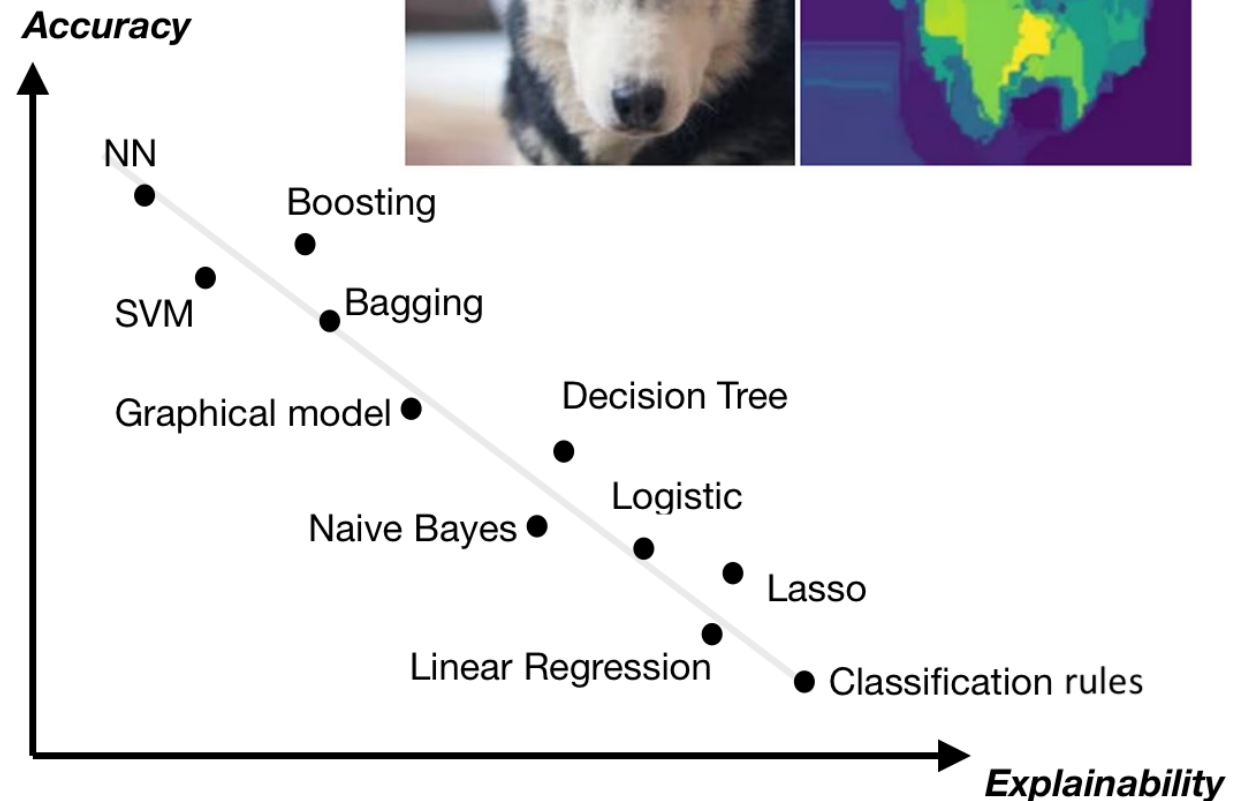
hierarchical merging



Explainable AI



- LLMs vs Decision trees – trade offs
- How are the models actually making decisions?
- What are they using to make their decisions?
- LIME – approximate complex model with more interpretable model – like regression tree
- Grad-CAM – gradients of the classification score of the last feature map – large values mean important for prediction
- Occlusion sensitivity – mask part of images, find important features, if they don't match the classification task (background not animal) – then bad





Summary

- LLMs are deceptively good – Turing test
- A plethora of inherent problems
- Worse than lying
- Susceptible to attack
- Degrade themselves over time
- Encode and obfuscate bias
- Security risk
- Harm users
- Harm environment
- Limited use – by already proficient users
- Potentially move towards Explainable AI





This Text Was Placed Here Intentionally



References



- **Hallucinations:**
- A Gentle Introduction to Hallucinations in Large Language Models - Adrian Tam
- A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions - Lei Huang et. al.
- **Towards a Proper Metaphysical Definition of GenAI Hallucinations:**
- ChatGPT is bullshit - Michael Townsen Hicks, James Humphries, Joe Slater
- Towards Large Reasoning Models: A Survey of Reinforced Reasoning with Large Language Models – Fengli Xu et. al.
- **Hallucinations 2:**
- AI hallucinations are a feature of LLM design, not a bug - Joseph Dumit and Andreas Roepstorff
- LLM Lies: Hallucinations are not Bugs, but Features as Adversarial Examples - Jia-Yu Yao et. al.
- **Adversarial Attacks:**
- Universal and Transferable Adversarial Attacks on Aligned Language Models – Andy Zou et. al.
- **External Adversarial Attacks:**
- Physical Adversarial Attacks on an Aerial Imagery Object Detector – Andrew Du et. al.
- Adversarial Sensor Attack on LiDAR-based Perception in Autonomous Driving – Yulong Cao et. al.
- Adversarial Examples in Environment Perception for Automated Driving – Jun Yan and Huilin Yin
- **Data Poisoning:**
- Medical large language models are vulnerable to data-poisoning attacks – Daniel Alber et. al.
- This new data poisoning tool lets artists fight back aGenAIst generative AI – Melissa Heikkila
- Learning to Poison Large Language Models During Instruction Tuning – Xiangyu Zhou et al.
- **Data Bias:**
- Insight - Amazon scraps secret AI recruiting tool that showed bias aGenAIst women - [Jeffrey Dastin](#)
- Covert Racism in AI: How Language Models Are Reinforcing Outdated Stereotypes – Katharine miller
- On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? – Emily Bender et. al.



References cont.

- **Data Leakage:**

- What can we learn from Data Leakage and Unlearning for Law? - Jaydeep Borkar
- QUANTIFYING MEMORIZATION ACROSS NEURAL LANGUAGE MODELS - Nicholas Carlini et. al.
- Deep Learning with Differential Privacy (DP-SGD Explained) – Mukul Rathi
- Analyzing Leakage of Personally Identifiable Information in Language Models - [Nils Lukas](#)
- Hundreds of LLM Servers Expose Corporate, Health & Other Online Data – Nate Nelson

- **Recursive Collapse:**

- AI models collapse when trained on recursively generated data - [Ilia Shumailov et. al.](#)
- Theoretical Proof that Generated Text in the Corpus Leads to the Collapse of Auto-regressive Language Models - Lecheng Wang

- THE CURSE OF RECURSION: TRAINING ON GENERATED DATA MAKES MODELS FORGET – Ilia Shumailov et. al.

- **Critical thinking:**

- Generative AI Can Harm Learning – Hamsa Bastani et. al.

- Critique of Generative AI Can Harm Learning Study Design – Steffi Tan et. al.

- Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students – Muhammad Abbas et. al.
- The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers – Hao-Ping (Hank) Lee et. al.

- **Transparency and Accountability:**

- Transparency and accountability in AI systems: safeguarding wellbeing in the age of algorithmic decision-making – Ben Chester Cheong
- Legal transparency in AI finance: facing the accountability dilemma in digital decision-making - Joshua Dupuy

- **Environmental Impact:**

- Explained: Generative AI's environmental impact – Adam Zewe
- The global E-waste Monitor 2024 – Electronic Waste Rising Five Times Faster than Documented E-waste Recycling: UN



References cont.

- **Summarization:**
- Can You Use LLMs as Evaluators? An LLM Evaluation Framework – Dan Cleary
- LLM Evaluation For Text Summarization – Gourav Bais
- Clinical Text Summarization: Adapting Large Language Models Can Outperform Human Experts – Dave Van Veen et. al.
- **Explainable AI:**
- What Is Explainable AI? – Mathworks
- Explainable AI, the key to open “black boxes” – Duval Alexandre