Data Council | March 27, 2024

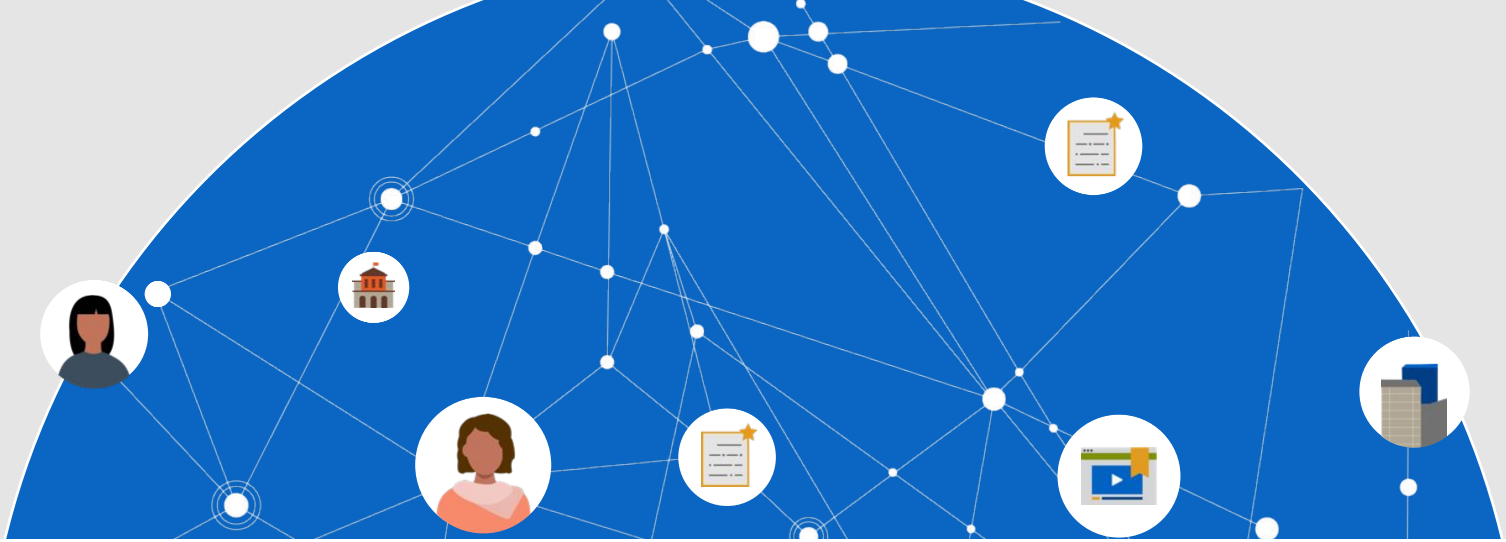
Building Responsible & Trustworthy Generative AI Products @ LinkedIn

Daniel Olmedilla

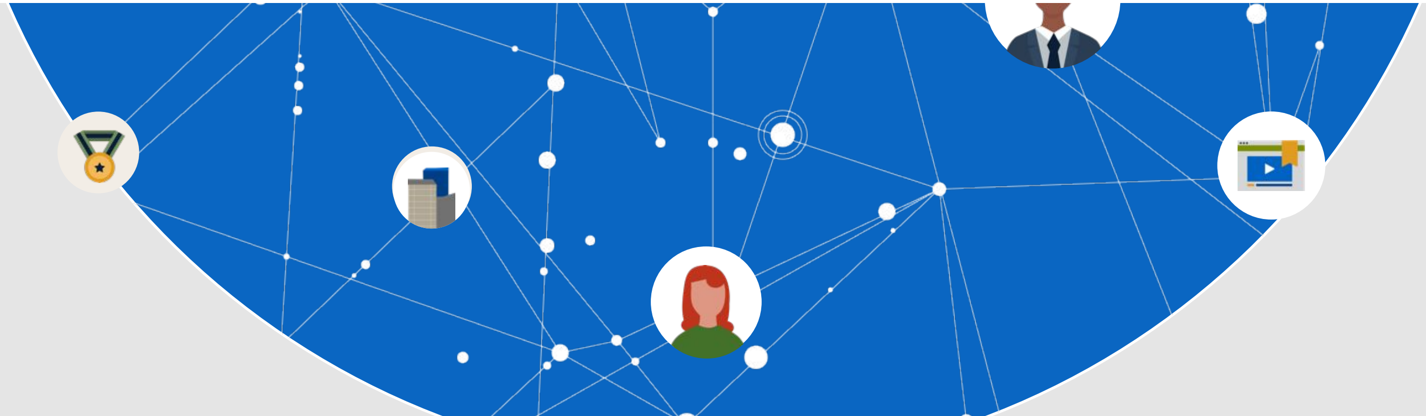
Sr. Director, Trust, Privacy and Responsible AI

Agenda

- What Does Trust and Responsible AI Mean?
- Generative AI Revolution at LinkedIn
- Our Strategy
- Other Uses of Generative AI in Trust



Create economic opportunity for every member of the global workforce



AI-Powered Products Across 3 Major Areas

Knowledge Sharing

1B Members
1.5M feed updates viewed / minute

Talent and Learning

4.3M+ active talent professionals

65M+ weekly job seekers

6 people are hired every minute

Products & Services

$6B+ Marketing and Sales Solutions annual revenue
2x # advertisers in past 5 years

AI-Powered Products: Search, Recommendations, Pacing and Bidding Optimization, ...
Generative AI-Powered Products: Collaborative Articles, Coach, write with AI, …

Trust, Privacy and Responsible AI

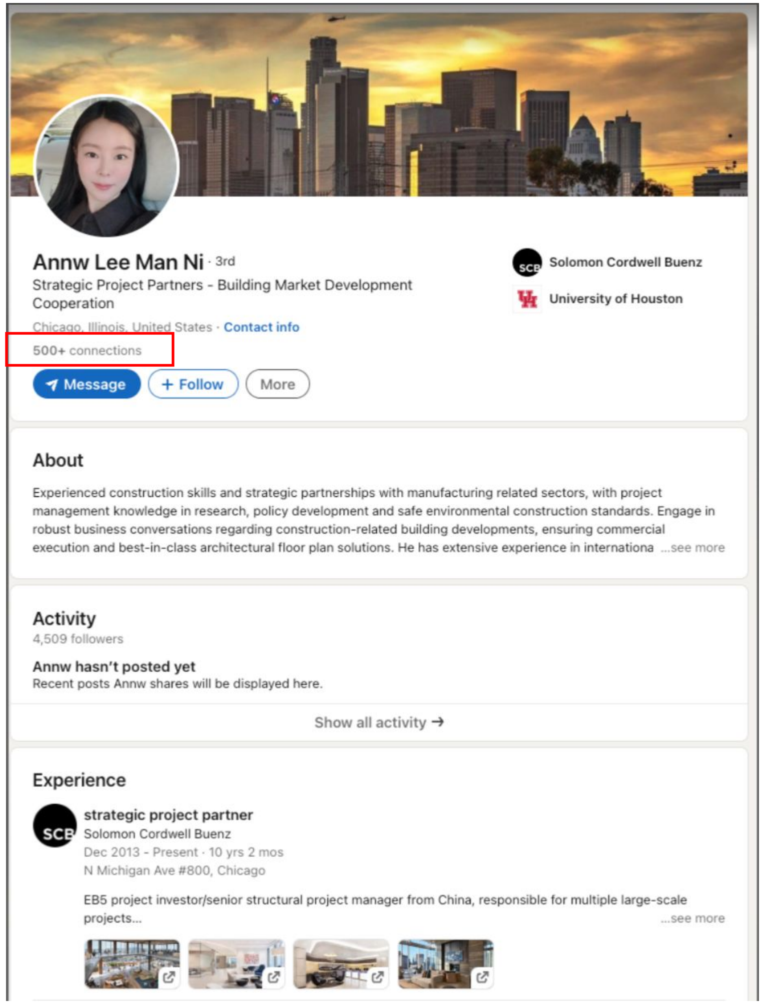
What Does Trust and Responsible AI Mean?

Examples of Abuse / Low Quality / Inauthenticity

Hate Speech, Illegal, ...

Unoriginal, Misinformation

Fake Account, Impersonation, ...



Responsible AI

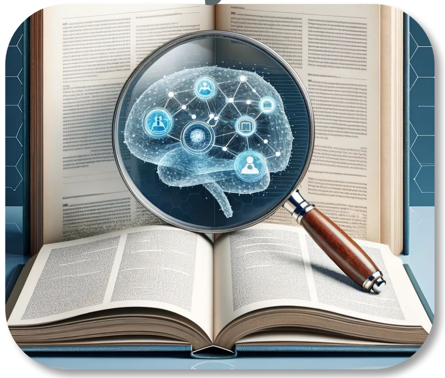
Fairness

Ensuring equal treatment in our AI models

Measurement and detection of algorithmic bias in AI models

Mitigation of model bias with algorithmic debiasing methods

Transparenc
y



Maintaining clarity in AI operations through systematic documentation and decision understanding

AI Governance framework for systematic model documentation and accountability

Explanations of AI model's decision-making process and feature importance

Privacy

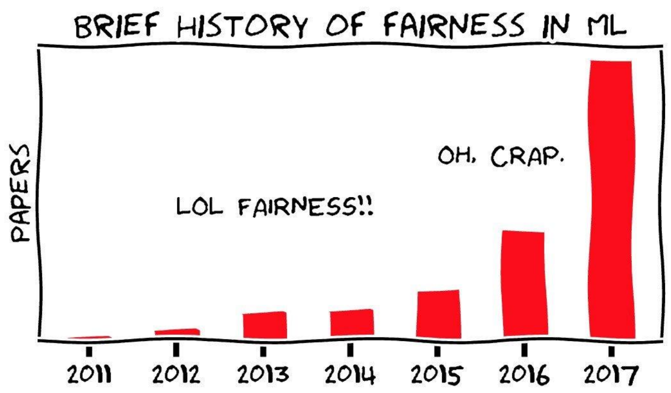
Protecting member data and identity

Ensuring compliance with privacy regulations

Increasing privacy and anonymity through privacy enhancing technologies

Responsible AI: Fairness

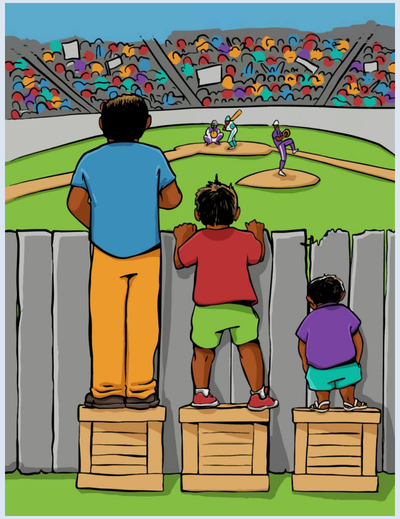
Fairness has **multiple definitions**, which are **mutually incompatible**



Translation tutorial:
21 fairness definitions and their politics

Arvind Narayanan
(Computer scientist, Princeton University)

Equal Treatment



Equitabl
e
Outcome
s

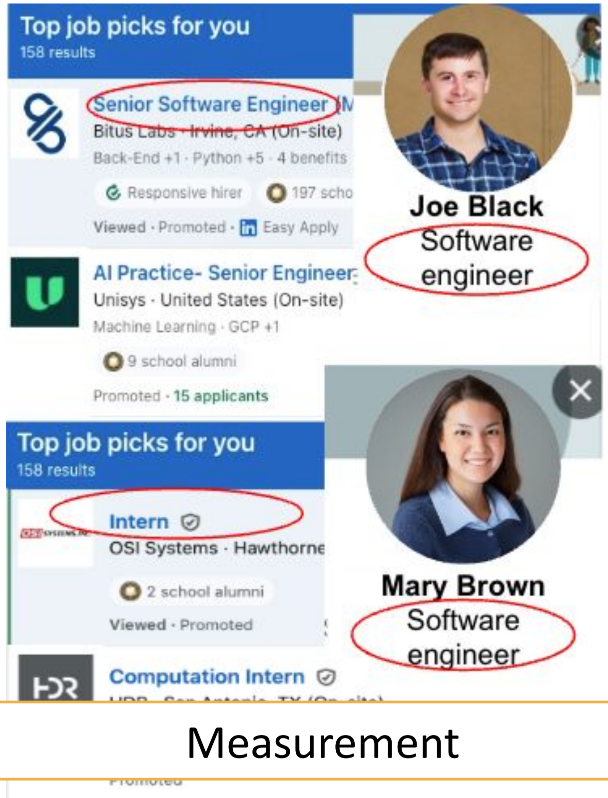
AI Fairness = Equal AI Treatment + Product Equity

ACM FAccT'23: Disentangling and Operationalizing AI Fairness at LinkedIn

Fairness Varies by Use Case

Candidate Side	Viewer Side	Content Subject

When I became the new Chief Learning Officer of the Navy last month, new LinkedIn algorithms kicked in. Now, my suggested new network partners are much less diverse, particularly with respect to gender. This morning, for example, 11 out of 12 suggested top new network members are men. I understand why that is...



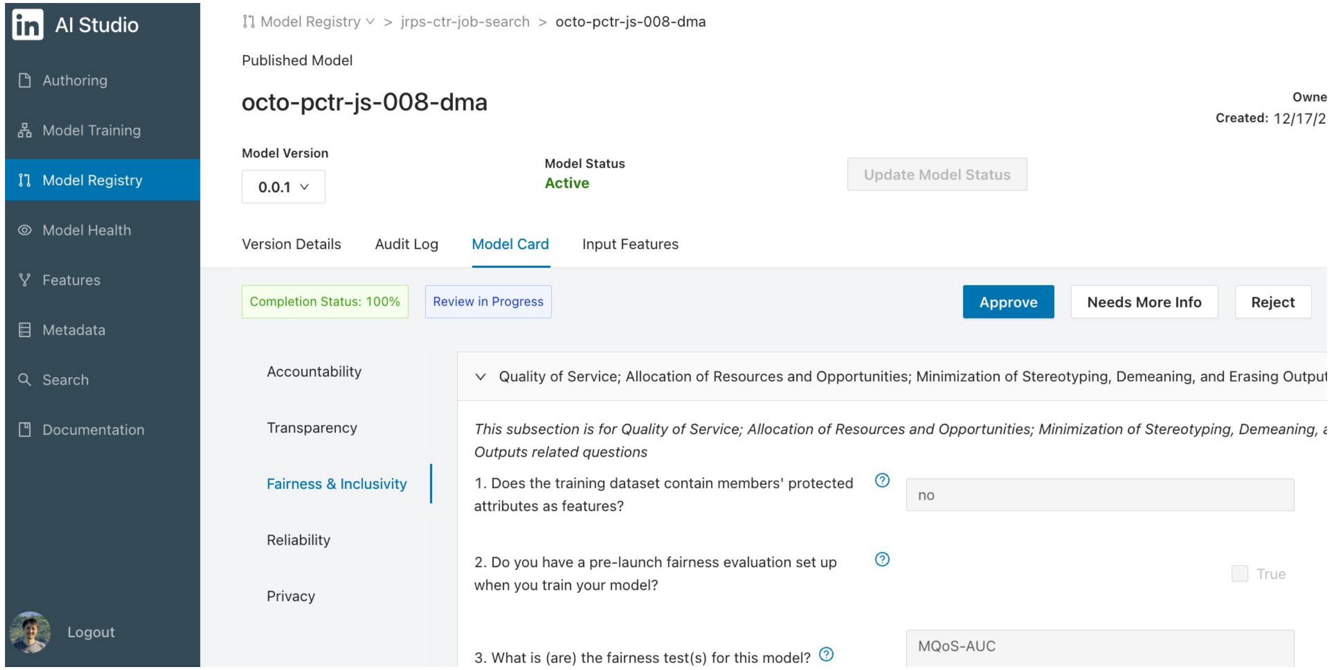
Measurement

Mitigation

Responsible AI: Transparency and Governance

Model cards: collect information that are essential for adherence to regulations and our RAI principles, with high integration into our systems

- Accountability – POCs, model purpose, data source, data processing, output format, adverse effects, etc.
- Transparency - System intelligibility for decision making
- Fairness and inclusivity measurements
- Reliability – performance, failure modes, health monitoring
- Privacy – PII used, data pipeline security



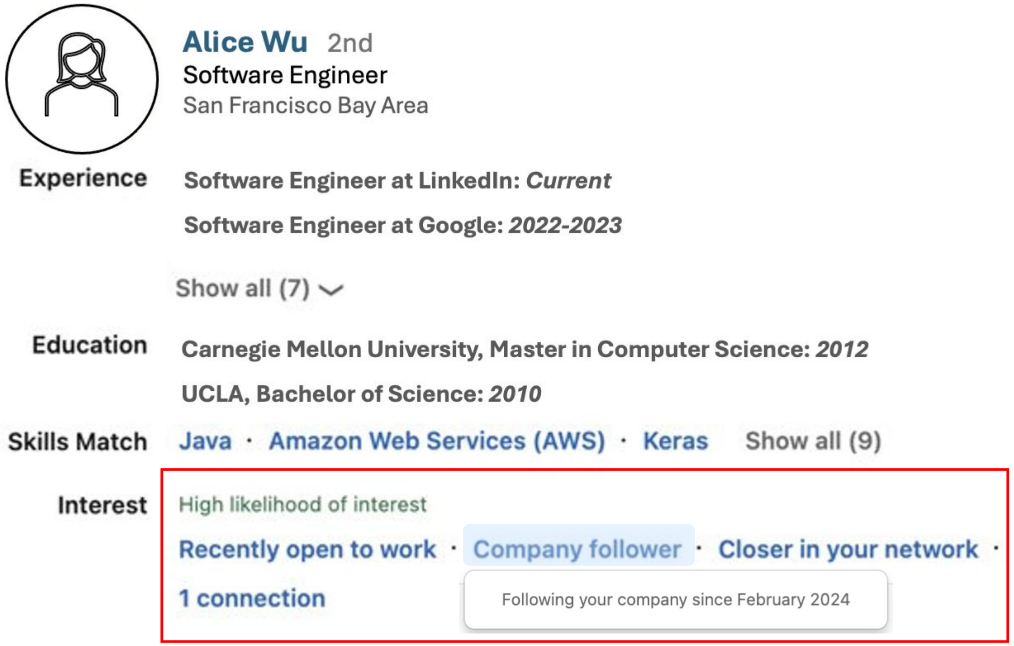
Responsible AI: Transparency (Example)

Explainable AI: Integrated gradients for finding important features

Explaining Recruiter Search: “Why am I recommended this candidate”

Method: Highlighting important features using Feature Attribution

Impact: Improve recruiter search metrics and enhance recruiter trust



Above: surfacing “likelihood of interest” insights to recruiters.

Responsible AI: Privacy

- **Compliance and data protection by design**

  Regulatory and platform changes (e.g. Apple ATT, Google Privacy Sandbox, GDPR and DMA) define what data can be collected, processed, including sensitive third-party data

- **Sensitive Data**

  Some AI products may need to leverage sources of sensitive data

- Using **Privacy Enhancing Technologies (PETs),** our AI products are built with privacy by design, complying with regulations and protecting sensitive data

Privacy Risks and Tech Capabilities

Re-identification of members using contextual information

- Non-PII identity (role, company) and engagement (views, clicks) data can sometimes be de-anonymized using context information.
  - “a CEO working at LinkedIn viewed your post” fully identifies Ryan Roslansky’s engagement.
- We build novel privacy metrics that quantify the re-identification risk, and we apply PETs like **Differential Privacy** to mitigate products with high risk.

Data is confidential and/or must be kept separated

- Traditional protections like encryption at rest and access controls don't minimize the collection or safe processing of sensitive data.
- We build PETs like **Federated Learning** and **Secure Multiparty Computation** to train models in a privacy-preserving way on:
  - Data distributed across edge devices/data silos.
  - Data that remains encrypted or de-identified throughout compute.

Generative AI
Revolution at LinkedIn

The Revolution of Generative AI

- Revolutionary technology advances opportunity, but also brings new and increased risks

LinkedIn **faces threats** by bad actors using GenAI to carry out harm through inauthentic accounts and harmful content

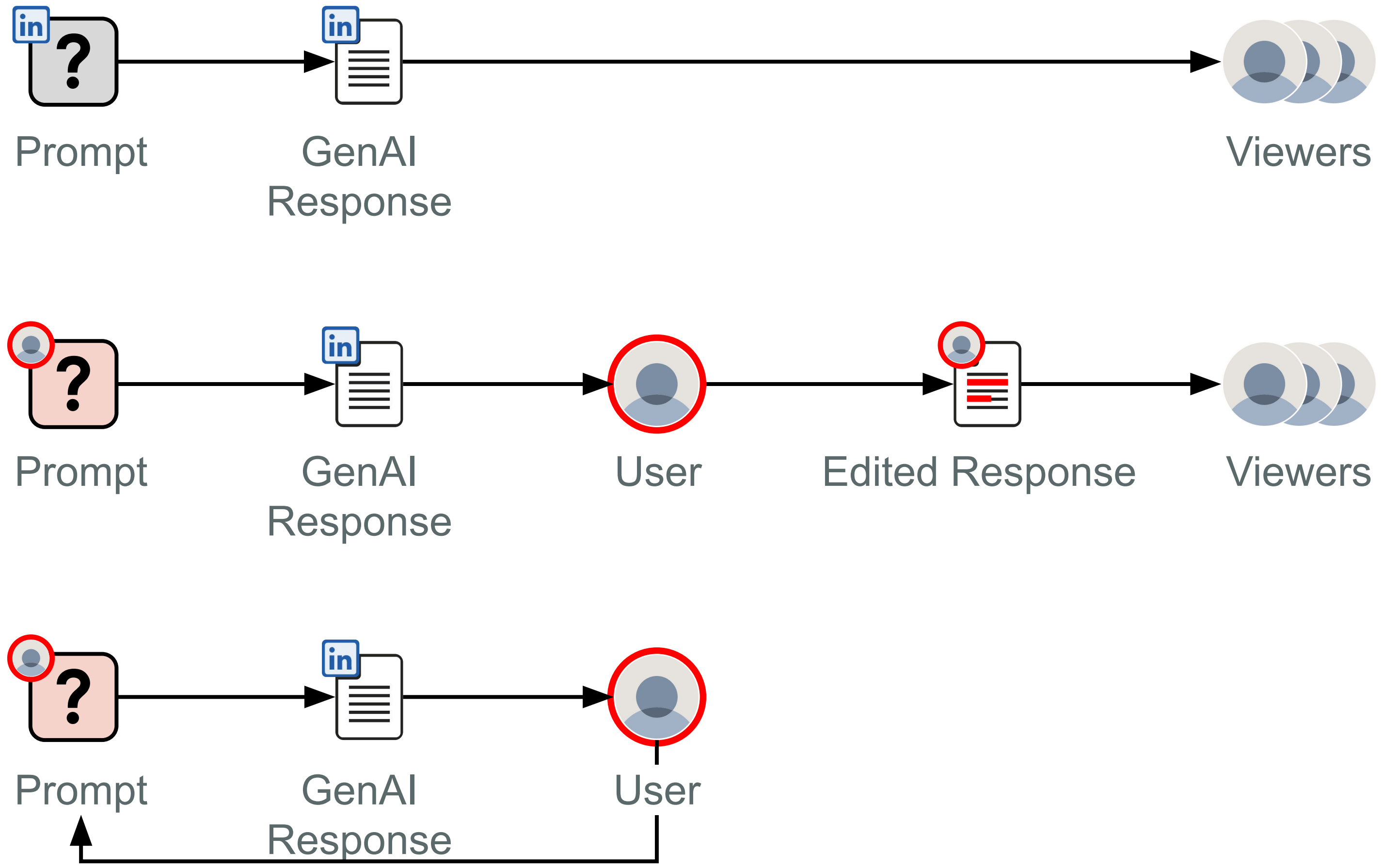
LinkedIn **uses** GenAI to improve its ability to measure, prevent, and mitigate abuse

LinkedIn **builds** GenAI products that need to be trustworthy and safeguarded from misuse

Overview of LinkedIn GenAI Products

- GenAI products breakdown into 3 types that differ in interaction pattern.

1. Direct Content Creation (e.g., Collaborative Articles)
2. Assisted Content Creation (e.g., GenAI Job Posting)
3. Copilot (e.g., AI Powered Coach)



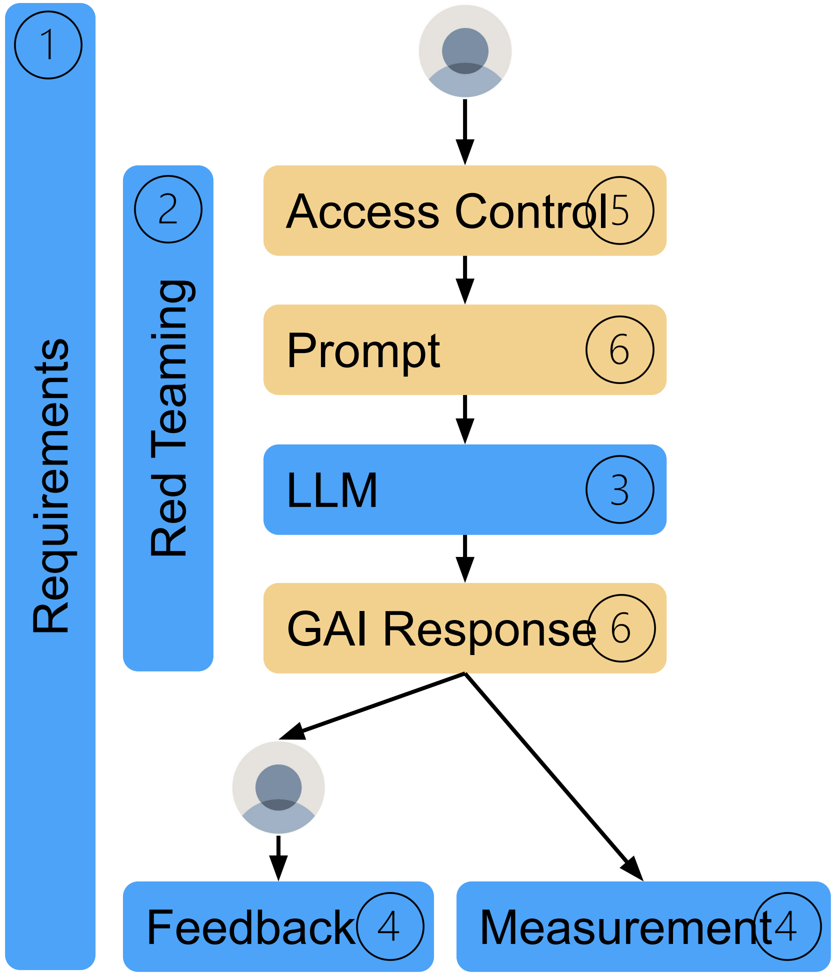
What is New from Trust Perspective

- GenAI products introduce challenges not usually seen in User Generated Content (UGC) moderation

	Challenge	Consequence
Product Experience	Interactive	• Low latency moderation that is fully automated
	User control over the input	• New attack vectors (jailbreaks and prompt leakage attacks) • Product misuse can enable generation of high-quality harmful content at scale
	Private interactions between user and GenAI	• Cannot rely on other users of the platform to report abuse
Development Process	High volume & rapid product launches & iterations	• Automated and routine risk assessments are needed to scale
Perception	LinkedIn is the content author	• Higher content standards, equating to improved detection for traditional risks and mitigations for new risks

Our Strategy

Trust and RAI for GenAI Products



Process

1. Launch requirements that ANY product needs to meet, ranging from legal disclosures, to user input character limits, to moderation filters

2. Manual Red Teaming sources new adversarial inputs while Automated Red Teaming regularly tests for regressions

3. Foundational LLMs used are aligned to be safe, ethical, and fair

4. Centrally tracked feedback-based metrics and bias measurement metrics are monitored as guardrails

Runtime Defenses

5. User Access Control limits access and misuse by bad actors

6. Moderation filters harmful inputs/outputs using LinkedIn classifiers, Azure AI Content Safety, and/or selected open source models

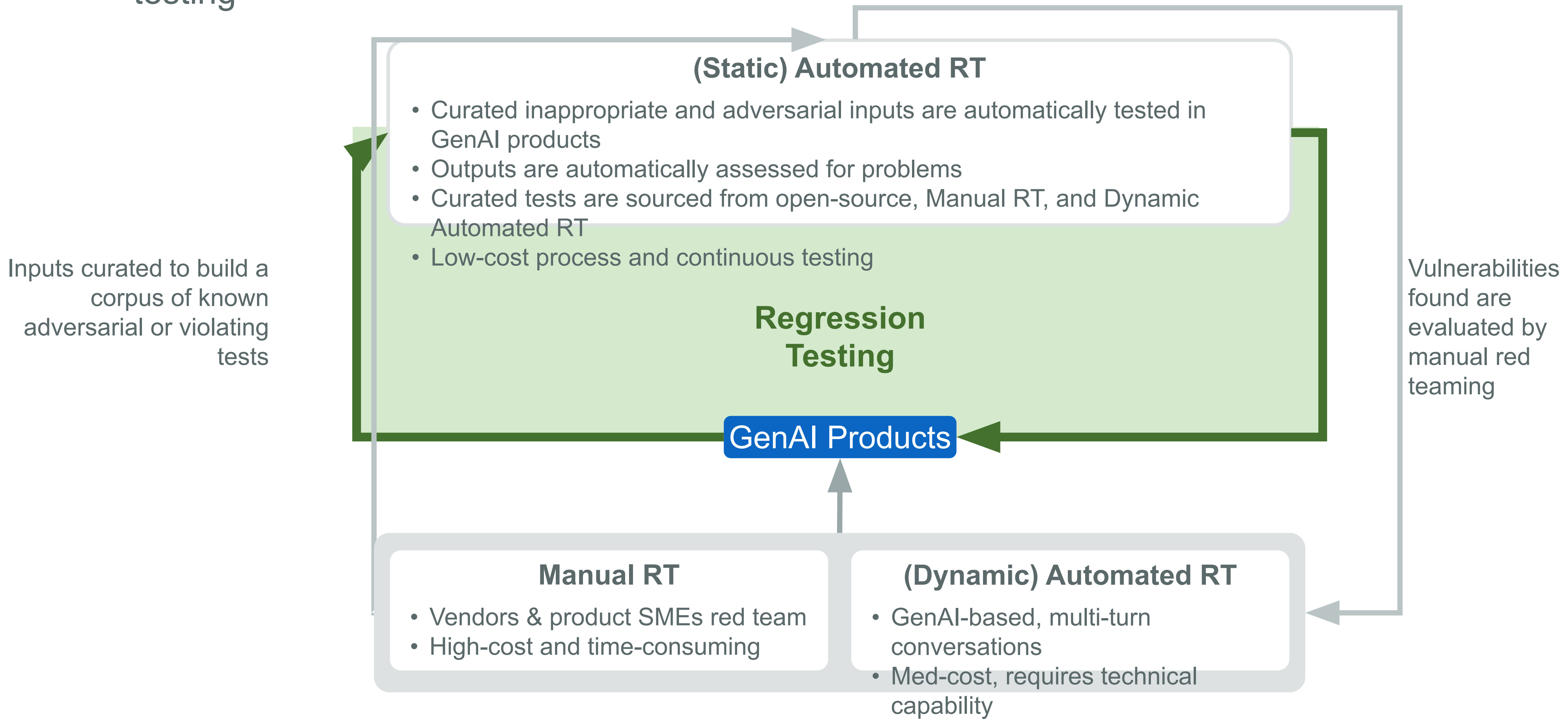
Requirements for GenAI-Powered Products

- For every GenAI product, we have **Process**, **UX**, **Measurement**, **Backend Safeguards**, and **Testing** requirements that increase based on ramp stage to account for increased risk
- Example (non-exhaustive) of Copilot Requirements

Ramp Stage	PP (Private Preview) Manually curated external users (<= 1000 total users)	MVP Any ramp to randomized users (>1000 total users)	GA 100% ramp to the target population
Risk	Lower Risk	Higher Risk	
Process	Trust / Legal / Security Review	MVP Review	GA Review
UX	User Feedback	Access Control (Trustworthy Members)	Usage Limits
Measurement	Guardrail Metrics (Tracking)	Bias / Stereotype Measurement	Scalable / Centralized Tracking
Backend Safeguards	Constrained Product Scope	Moderation Defenses Calibration	Full Moderation Defenses
Testing	Manual Quality Evaluation	Manual Red Teaming	Automated Red Teaming

Red Teaming (RT)

- Aims to identify the weaknesses and gaps within GenAI products through adversarial testing

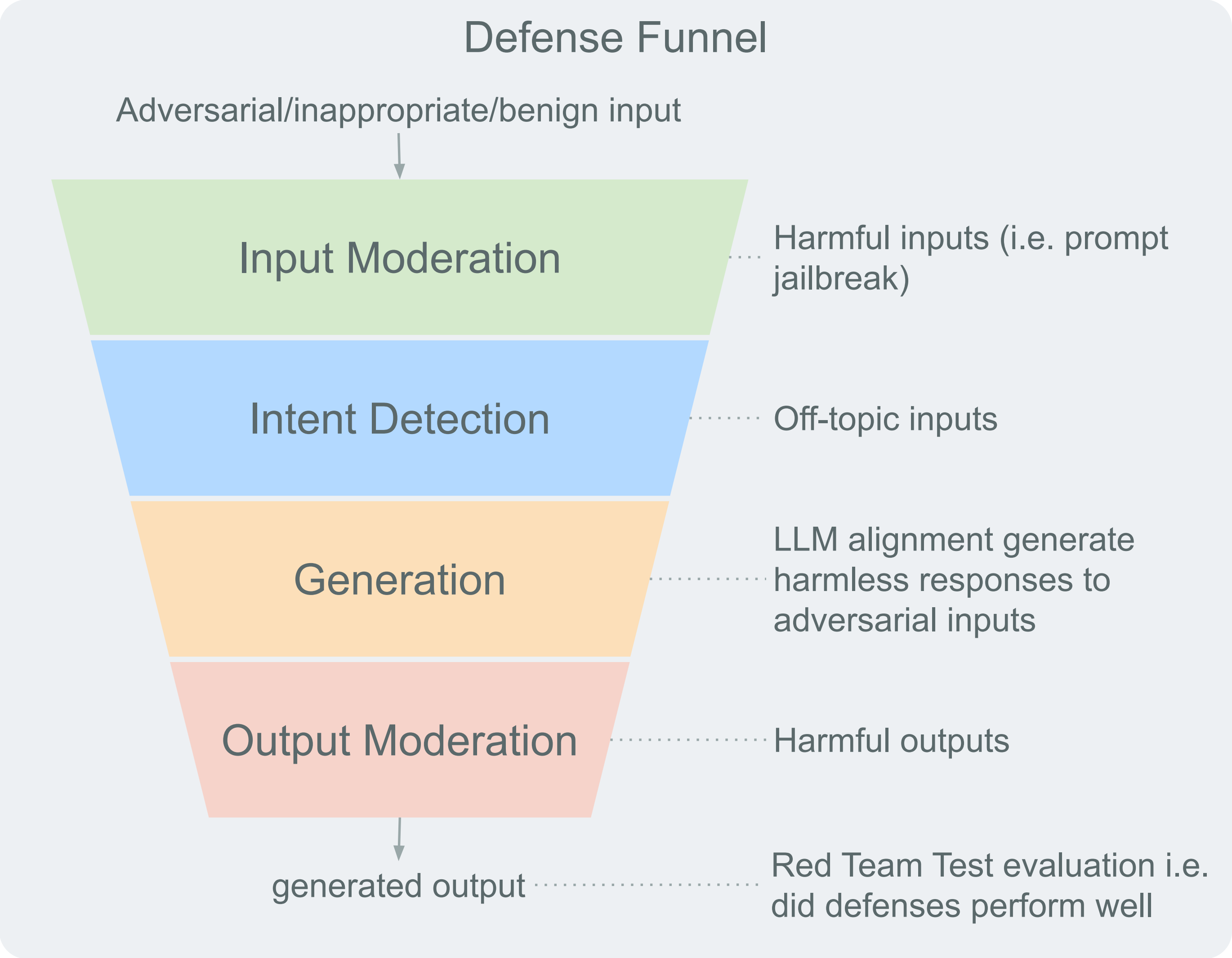


Red Teaming (RT): Automated

- Provides scalable, continuous, & configurable risk evaluation

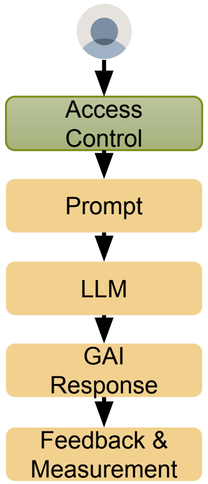
- Run cadence is decoupled from product deployments as risk can be introduced without deployments

- Weekly metrics
  - Goal: does the system correctly stop inappropriate and adversarial inputs
  - Recall: do we stop all of them
    Represents product robustness to misuse
  - Precision: do we stop good ones by mistake
    Represents defense funnel overenforcing



Access Control

- Bad actors can leverage GAI products to more efficiently and effectively conduct harm (e.g., through inauthentic accounts, harmful content, scam jobs)
- Access to advanced features is gated on our confidence of them being good or posing low risk (i.e. authenticity level, past behavior, etc.)



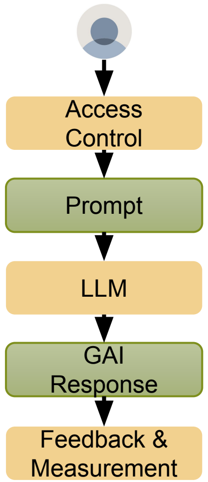
- Different products might provide access to different levels depending on the level of risk. The following represents an example

Tier	Account Risk Level	Action
1	High confidence good	Full access
2	Likely good	Full access
3	Unknown	Limited access
4	Low confidence inauthentic	Deny

Some actions might increase our confidence level
i.e. Account Verification via identity, workplace, or educational institution

- Note: equitable access is a critical part of this logic

Input & Generated Content Moderation



- Removes problematic content being returned to the user
- Applies to both the input and response from the GAI product with a few key differences

Addressable Risks

Select risks are better addressed at a given stage i.e. input vs generated content moderation

Jailbreak is typically detected during input moderation as that is where the signature is present.

Semantics

Input and generated content often require different moderation capabilities

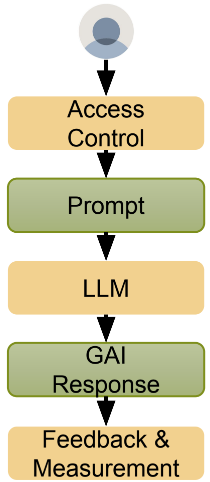
Illegal Input: *I want to get high. Where can I get some crack in the Bay Area?*
Illegal Generated Content: Crack is reported to be fairly accessible in Berkeley and has the key benefit of providing a "rush"

Moderation Mode

UI may be streaming (shows generated text incrementally), introducing additional complexity

Chunking Logic: *every 1000 tokens, additive*
User Experience: *once harmful content is detected, vanish the content and replace with a canned message*

Input & Generated Content Moderation - Examples



- All inputs and generated content (text and image) to/from a GAI product are moderated
- Problematic contents are blocked, and self-harm inputs are blocked with a “help hotline” message

Risks	Input	Generated Response
Jailbreak	Yes	No
Prompt Leakage	No	Yes
Self-Harm	Yes	Yes
Hate/Harassment	Select Products	Yes
Violence	Select Products	Yes
Illegal Regulated Content	No	Yes
Sensitive Topics	No	Select Products, Select Topics
Discriminatory Jobs	No	Job Products

Where possible, moderation focuses on the generated content as inputs are more varied and more prone to false positives

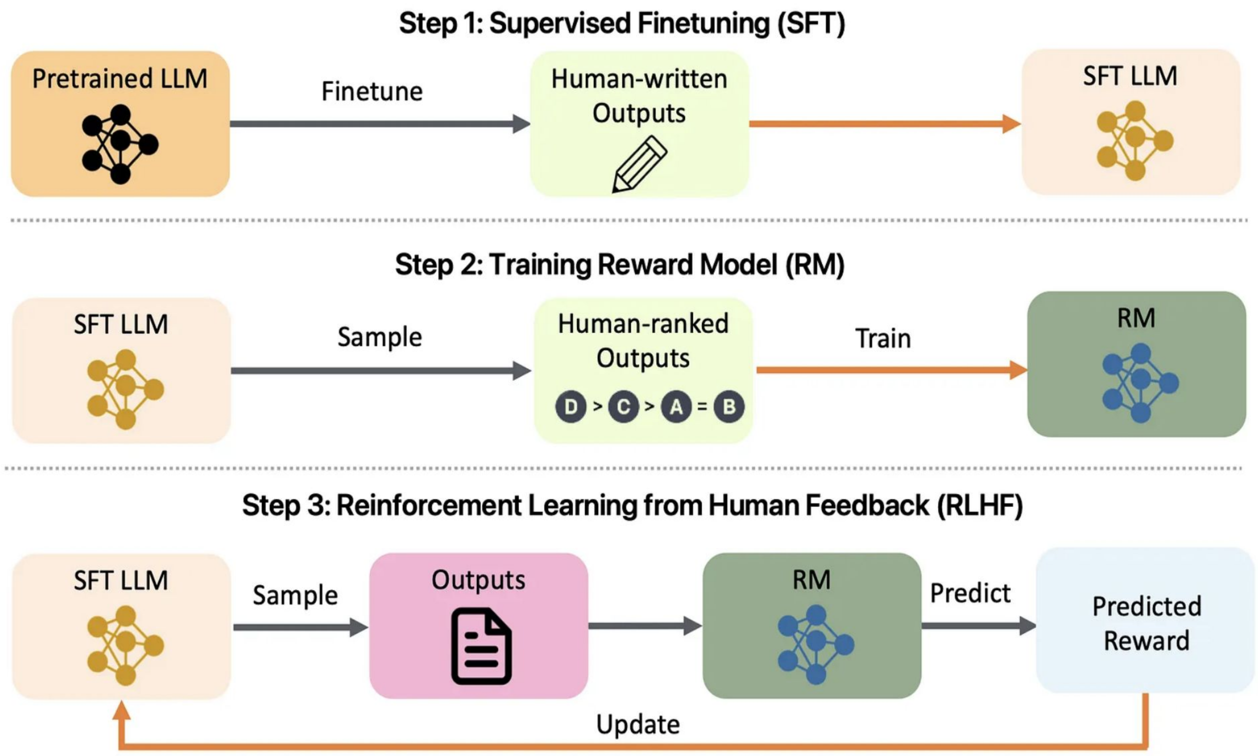
Products grounding the response from curated content (e.g., LinkedIn Learning) can operate on sensitive topics

In-House Classifiers

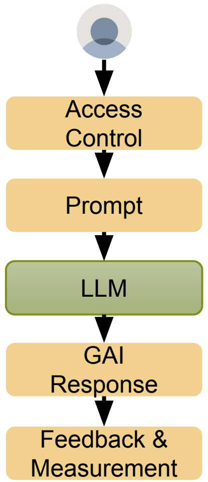
Azure AI Content Safety + we also test and use selected open source models

Model Alignment

- We ensure that any foundational model we use is compliant with LinkedIn's Responsible AI Principles
- Alignment refers to the process of ensuring that LLMs behave according to human values and preferences
- Alignment focuses on fundamental risks (e.g., sexual content) and not on product specific risks (e.g., political content)
- We have a test bed to compare various methods: RLHF, DPO, KPO, etc.



Source: https://arxiv.org/abs/2308.05374

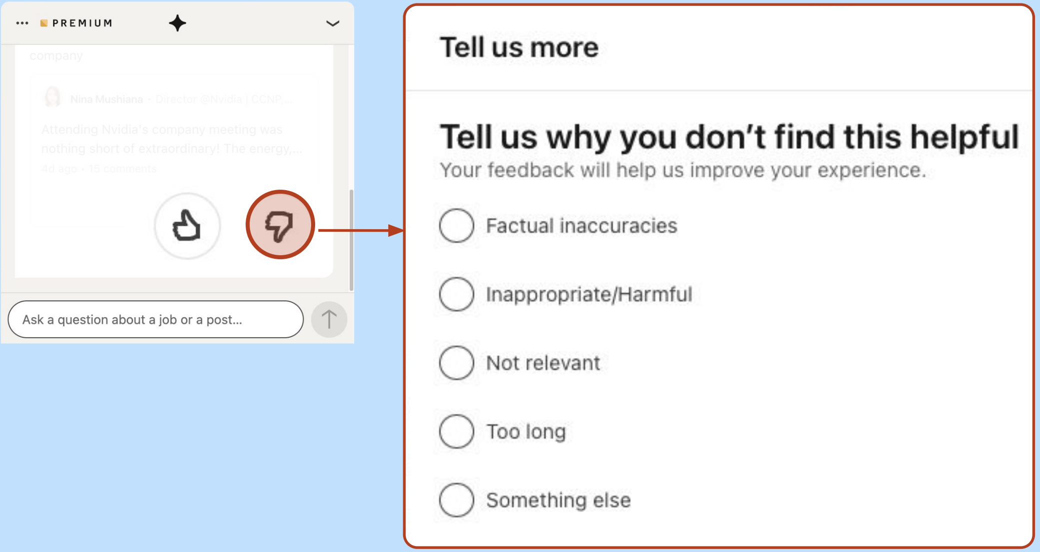


Measurement: Feedback-Based

- User feedback metrics serve as the guardrails for all GenAI products

Common UX & User Feedback

Standardizes UI for users and enables scalable, central oversight of feedback for all GenAI products



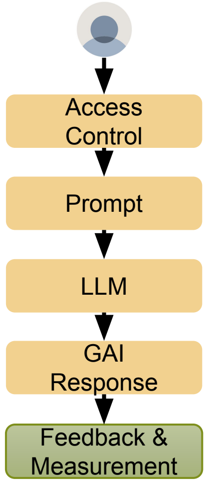
Common Metrics

(Guardrail) Negative Feedback Rate measures the user perception of GenAI response quality

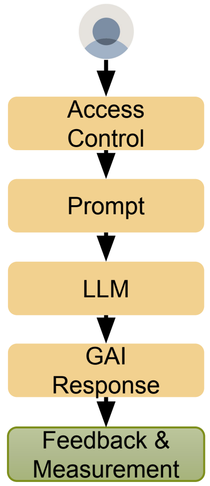
$$\text{NFR} = \frac{\#\text{negative feedback}}{\#\text{ generated response}}$$

(Guardrail) Trust Feedback Rate measures the user perception of GenAI response harmfulness

$$\text{TFR} = \frac{\#\text{factual inaccuracies} + \#\text{inapproriate}}{\#\text{ generated response}}$$



Measurement: Bias/Stereotypes



Incidental

Biases/stereotypes that are evident in any single piece of GenAI content. Examples:

- Straightforward: Every junior engineer should discuss his career goals with his manager.
- Subtle: New mothers ought to pause their career and stay home with their babies for 2 years.

Representational

Skewness in demographic distribution across GenAI for an industry or a product that does not innately cater to a specific demographic group. Examples:

- Text: Articles on science careers only give examples of male scientists
- Images: doctors are always white

- **Challenges**
  - Bias/stereotypes can be very subjective and varies widely
  - Difficult to collect positive examples to train a high-quality model

- **Solutions**
  - Leverage LLMs, which are trained on large amount of data and have a general understanding of bias/stereotypes (high cost)
  - Leverage two-stage measurement flow to balance cost, precision, and recall
  - Retrain models with balanced data
  - Prompt re-writing strategies

Other Uses of Generative AI in Trust

The Revolution of Generative AI

- Revolutionary technology advances opportunity, but also brings new and increased risks

LinkedIn **faces threats** by bad actors using GenAI to carry out harm through inauthentic accounts and harmful content

LinkedIn **uses** GenAI to improve its ability to measure, prevent, and mitigate abuse

LinkedIn **builds** GenAI products that need to be trustworthy and safeguarded from misuse

AI-Assisted Sampling

Abuse Measurement

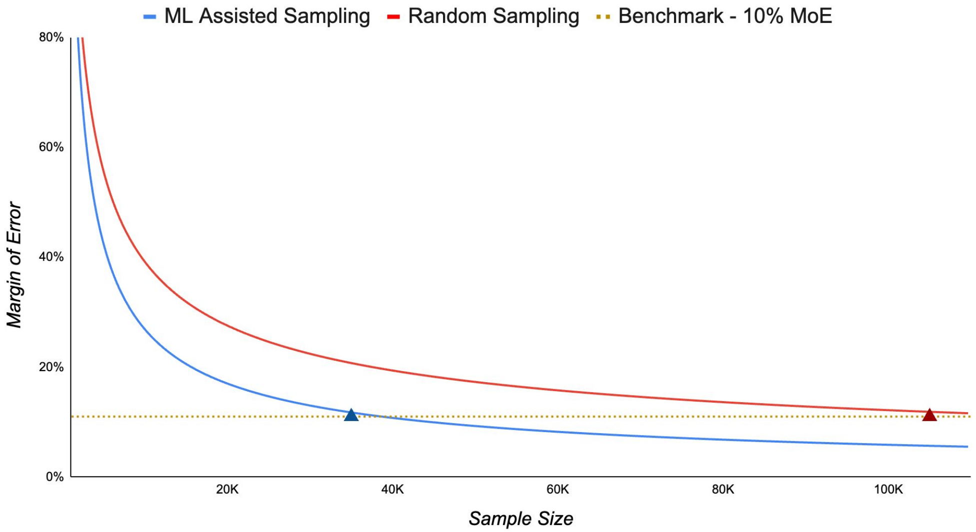
Easily up to 95% is unknown

Sampling with human labeling is expensive

- Less than 1% of samples are typically positive (high class imbalance)

ML-assisted sampling can cut costs by 70% with tighter error bars

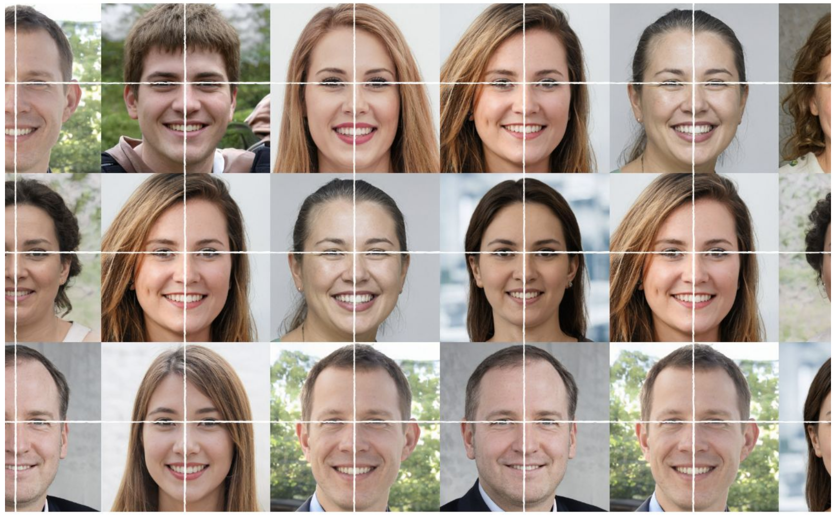
- Unbiased estimator using an ML model biasing our sampling



Deepfakes

TECHNOLOGY

That smiling LinkedIn profile face might be a computer-generated fake



Motivation

Emerging trend of large fake account attacks using automated generation and deepfake profiles photos

Solution

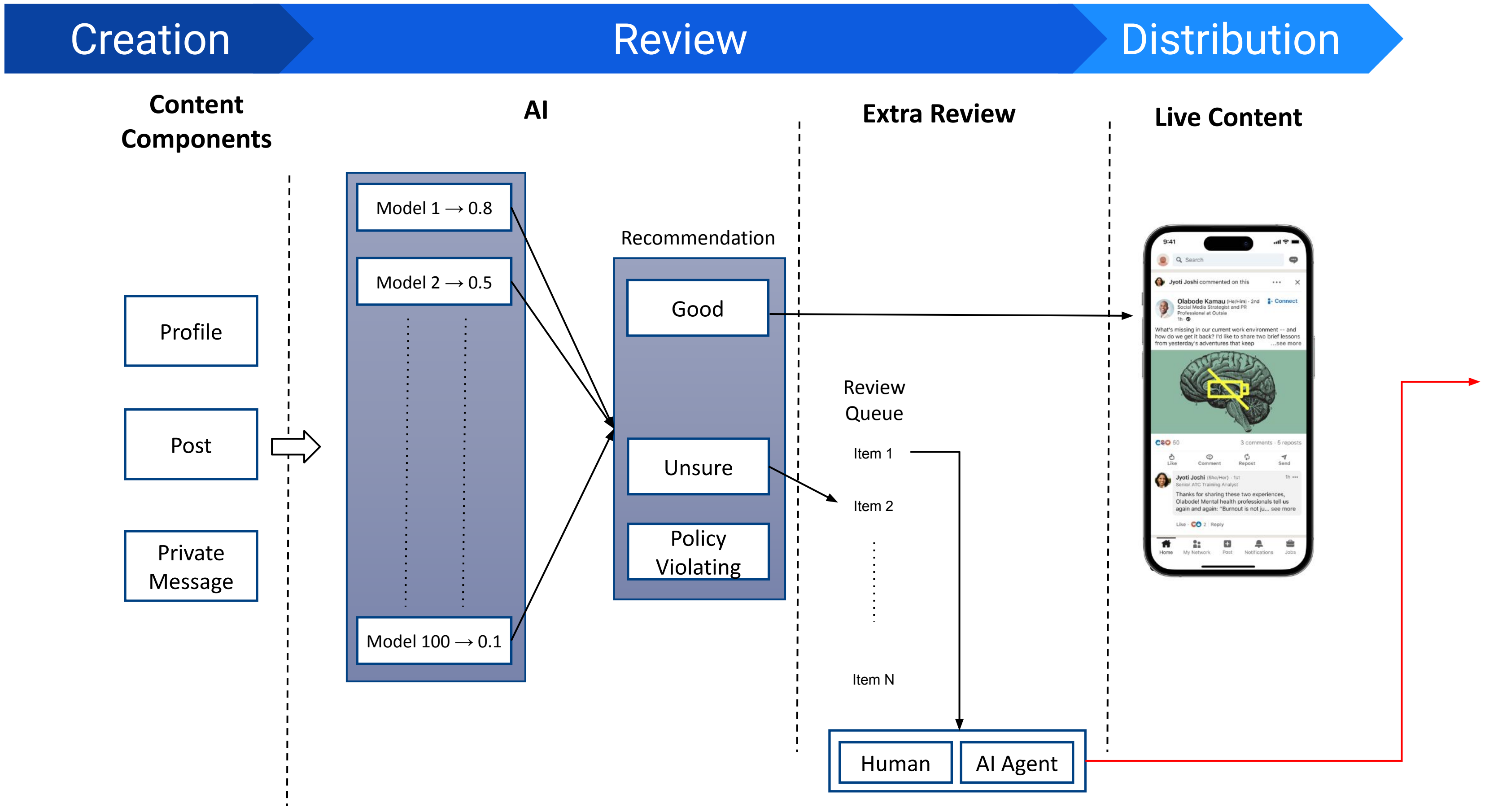
Detector for AI-generated (deepfake) profile photos

- *Precision = 99.3% and recall = 85.4%*
- *Catches GAN-generated (StyleGAN1, StyleGAN2, StyleGAN3, EG3D) and diffusion-based (Stable Diffusion v1, Stable Diffusion v2, and DALL-E) images*

Engineering Blogpost: [New Approaches For Detecting AI-Generated Profile Photos]

Scale Up Content Review

Creating content gets easier and cheaper, requiring trust teams to scale up their review



Opportunity:
- Higher scalability
- Better decision quality
- Human wellbeing improvement
- More cost efficient
- Easier scaling up and down

Thank you!

danielo@linkedin.com