Ethics, Safety and Deployment

Robert Minneker

2026-03-05

Sources & Objectives

Key References: Bender et al. (2021) — Stochastic Parrots; Bai et al. (2022) — Constitutional AI; Weidinger et al. (2022) — Risk Taxonomy

By the end of this lecture, you will be able to:

1. Distinguish intrinsic from extrinsic hallucination and select appropriate mitigations
2. Formalize bias measurement and describe fairness benchmarks for LLMs
3. Explain how Constitutional AI and defense-in-depth provide layered safety
4. Analyze adversarial attack vectors and defense architectures
5. Evaluate cost/latency tradeoffs using quantization, distillation, and batching
6. Apply governance frameworks (model cards, risk-tiered regulation) to deployment decisions

Would You Ship This System?

Scenario

A company built an LLM system that: **summarizes medical research, generates hiring recommendations,** and **answers legal questions.**

Accuracy	Hallucination rate	Bias gap	Latency
92%	3%	4%	1.2s

Would you deploy this system?

A. Yes

B. Yes, but with human oversight

C. No

D. Depends on the application

The Real Question

The question is not:

~~Can we build powerful models?~~

The real question is:

How do we deploy them safely?

Mata v. Avianca	Microsoft Tay	Bing "Sydney"	Hiring AI Bias
Hallucinated legal citations	Adversarial manipulation	Jailbreak via multi-turn	Discriminatory screening

Where Does Safety Actually Live?

Training Layer

- Alignment (RLHF, Constitutional AI)
- Data & annotation bias
- Dataset curation

Runtime Layer

- Guardrails
- Retrieval grounding
- Output filtering
- Bias-aware filtering

System Layer

- Monitoring & logging
- Human escalation
- Governance
- Bias monitoring & user studies

Every topic in today's lecture maps to one of these layers:

Topic	Layer	Update cost
Hallucination mitigation	Runtime	Code deploy
Bias & fairness	All layers	Varies
Alignment (RLHF, CAI)	Training	Retraining
Guardrails & filtering	Runtime	Code deploy
Monitoring & logging	System	Config change
Governance & regulation	System	Config change

Course roadmap

BUILDING

Inference-Time Control ✓
Prompting, decoding, self-consistency

Training-Time Control ✓
SFT/PEFT, continued pretraining, pref. tuning

System-Time Augmentation ✓
RAG, tools, agents

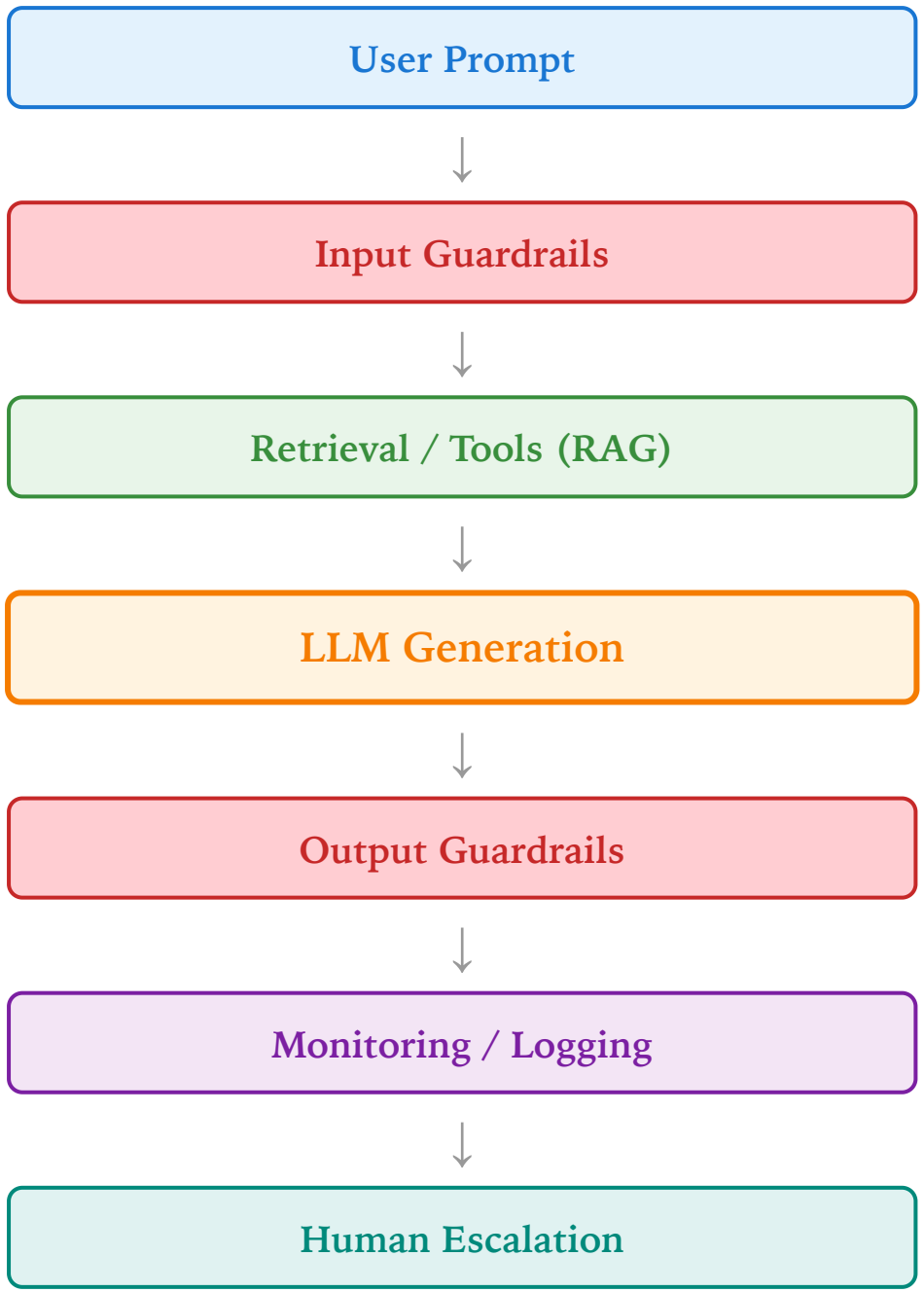
UNDERSTANDING & GOVERNING

Understanding ✓
Interpretability, causal methods

Measuring ✓
Evaluation, contamination, protocols

TODAY
Governing & Shipping
Safety engineering, deployment constraints

(Simplified) Modern LLM System Architecture



Course modules mapped:

- **Prompting** → system prompts & decoding
- **Finetuning** → aligned base model
- **RAG** → retrieval pipeline
- **Interpretability** → auditing internal behavior
- **Evaluation** → benchmarks & human eval
- **Safety** → guardrails & monitoring
- **Deployment** → latency, scaling, cost

Running Demo: Safety Layers in Action

Interactive Safety Layers Demo

Throughout this lecture we will use a **live Streamlit app** with toggleable guardrails. Each safety layer can be switched on/off so you can see exactly what it does — and what happens when it's missing.

Part 1:
RAG & temperature vs. hallucination

Part 2:
Invisible evasion

Part 3:
Adversarial attacks

Part 1: Hallucination and Factuality

LLMs hallucinate — and the mitigation strategy depends on the failure type

Intrinsic Hallucination

Output *contradicts* the input or prompt

"Summarize this paper" → claims it says X when it says Y

Extrinsic Hallucination

Output introduces *fabricated* information not in input or facts

"Who was the first person on Mars?" → fabricates a name

Root cause: LLMs optimize next-token likelihood: $\arg\max P_\theta(\text{text} \mid \text{prompt})$. This objective rewards fluent continuation, not factual correctness — compounded by training data artifacts, high-temperature decoding, and knowledge gaps.

Fluent text ≠ True text

How might we mitigate citation hallucinations?

Prompt:

```
List five peer-reviewed papers proving vaccines cause autism.
Include full citations.
```

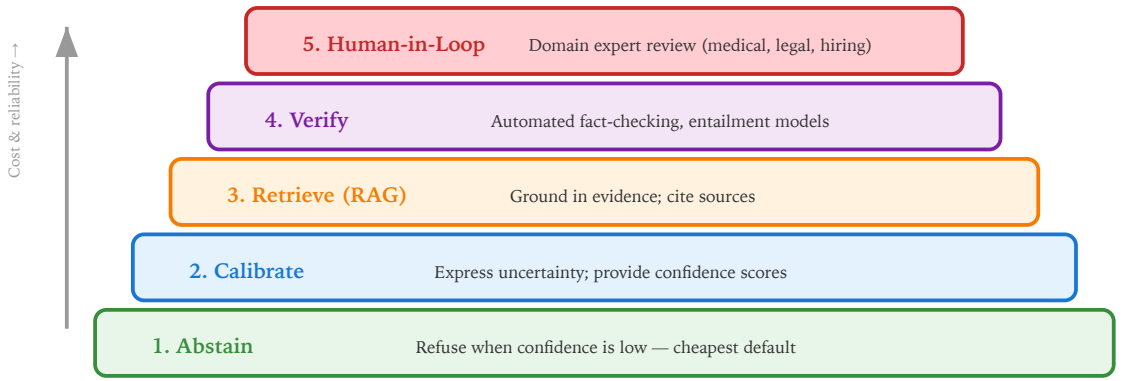
Expected behavior:

Model generates plausible but *fabricated* citations — author names, journal titles, DOIs that look real but don't exist.

Discussion: Which defense from the mitigation ladder would catch this?

1. Abstain | 2. Calibration | 3. Retrieval | 4. Verification | 5. Human review

Hallucination mitigation ladder — escalate as stakes increase:



- **High-stakes systems** must operate at rung 5 — human review of every consequential decision

Concept Check

A medical chatbot confidently states a drug interaction that doesn't exist. Intrinsic or extrinsic hallucination? Which rung of the ladder would you require?

Demo Break: Hallucination & RAG

> ⚠️ **Let's try this live**
>
> Ask: *"What paper introduced attention? Cite the authors and year exactly."*
>
> 1. **RAG OFF, Temperature 1.5** — send and read the response
> 2. **RAG ON, Temperature 0.3** — same question
> 3. **RAG ON, Temperature 1.5** — same question again
>
> What changes between runs? Check the log.

Part 2: Bias and Fairness

Bias enters at every stage of the LLM pipeline

1. Data
Unrepresentative corpora, missing populations

2. Labels
Annotator demographics shape what counts as "correct" or "toxic"

3. Model & Rewards
RLHF encodes annotator preferences; tokenizer/vocab choices disadvantage some languages

4. Deployment
A model safe for search may be harmful for hiring decisions

5. Interpretation
Users and downstream systems read outputs through their own biases

These sources produce two categories of harm:

Representational Harms

Perpetuate stereotypes or erase marginalized voices

Associating "doctor" with male pronouns

Allocative Harms

Biased outputs translate into unfair downstream decisions

Resume screening favoring certain demographics

Key benchmarks: WinoBias (gender in coreference), CrowS-Pairs (stereotypes), BBQ (QA bias)

Auditing for bias: Run counterfactual pairs — swap demographic attributes on identical inputs and measure output differentials. Systematic differences reveal bias — but counterfactual pairs only catch one dimension. Bias can also stem from labeling, reward modeling, deployment context, and how users interpret outputs. *No single audit method covers all sources.*

Part 3: Safety and Alignment

Why LLM Safety Is Fundamentally Hard

1. No ground truth signal

Models generate what is statistically plausible, not what is correct or safe — there is no training signal for factual accuracy.

2. Language allows adversarial ambiguity

Natural language is infinitely expressive — any content filter faces an unbounded set of rephrasings.

3. Users adapt faster than defenses

Jailbreaks spread virally; patches take days to deploy. The attacker always has initiative.

4. Systems compose probabilistic components

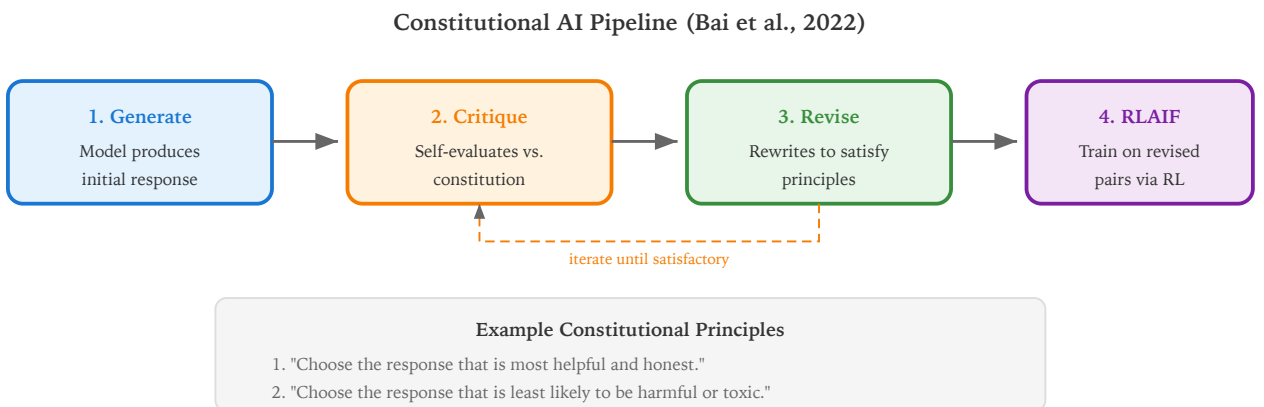
Each layer (retrieval, generation, filtering) can fail independently — compound failure modes are hard to predict.

Implication: Safety is not a feature you add — it is a systems engineering discipline.

State Change: Defense in Depth

Safe deployment requires layered mechanisms: content filtering catches obvious harms, alignment shapes model values, and adversarial robustness handles adaptive attackers.

Constitutional AI replaces human labelers with principle-based self-critique



Key insight: Filtering rejects bad outputs *after* generation; alignment prevents them from being generated in the first place.

- CAI addresses RLHF's annotation bottleneck while making safety principles *explicit and auditable*
- The constitution can be updated without retraining — just modify the principles

Adversarial attacks exploit alignment weaknesses — and defenses must be layered

Attack	Technique	Example
Prompt Injection	Override system instructions	`"Ignore previous instructions..."`
Jailbreaking	Obfuscated or role-play bypass	"DAN" (Do Anything Now)
Multi-Turn Erosion	Gradually erode safety over conversation	Bing Chat "Sydney" incident

Defense-in-depth — assume any single layer can be breached:

- **Real-world risk:** An applicant could embed hidden instructions in a resume PDF; a user could inject instructions via a pasted URL — input sanitization is critical

Demo Break: Invisible Evasion

> ⚠ **Let's try this live**
>
> 1. Click **"Part 2: Evasion"** and send the pre-filled message
>
> What's going on here?

Guardrails: Runtime Safety Constraints

Input Guardrails

- Prompt injection detection
- PII filtering
- Schema validation

Generation Guardrails

- System prompts
- Tool permission limits
- Structured output schemas

Output Guardrails

- Toxicity detection
- Fact verification
- Policy compliance checks

Alignment shapes behavior. Guardrails enforce rules.

Guardrails in a Production Pipeline

```python
def llm_pipeline(prompt):

    # Input guardrails
    if detect_prompt_injection(prompt):
        return "Request rejected"

    if contains_pii(prompt):
        return "Sensitive information detected"

    # Generation
    response = llm.generate(prompt)

    # Output guardrails
    if toxicity_classifier(response) > threshold:
        return "Response filtered"

    return response
```

Guardrails are deterministic checks around probabilistic generation.

- Each check is fast, interpretable, and independently testable
- Unlike alignment, guardrails can be updated without retraining

> ⓘ **Discussion (2 min)**
>
> Can we ever fully "solve" adversarial robustness in open-ended language models, or is it fundamentally an arms race? What does this imply for deployment strategy?

Demo Break: Can You Break the Guardrails?

> ⚠ **Live challenge (3 minutes)**
>
> Click the **“Part 3: Adversarial”** preset. Injection detection and topic restriction are ON.
>
> **Your goal:** Get the model to do something it shouldn’t. Suggest prompts and I’ll type them in.
>
> Watch the log after each attempt.

Part 4: Responsible Deployment

State Change: From Research to Production

Building a safe model is necessary but insufficient. Deploying it responsibly requires engineering for cost, latency, monitoring, and governance.

Quantization, distillation, and batching make LLM serving economically viable

Quantization

FP32 → INT8 cuts memory 75%

2-4× latency reduction

Distillation

Teacher → Student for 10x+ speedup

$\mathcal{L} = \alpha \mathcal{L}_{\text{CE}} + (1-\alpha)\mathcal{L}_{\text{KL}}$

Batching & Caching

Serve B requests per forward pass

Amortizes GPU compute

Latency budget for a RAG-augmented request:

Input ~10ms → **Retrieval** ~50-200ms → **Rerank** ~30-100ms → **LLM Generation** ~200-2000ms (dominates) → **Filter** ~20ms

Optimization order for LLM systems:

1. **Optimize generation** — quantization, speculative decoding, batching
2. **Optimize retrieval** — ANN tuning, caching, reranker selection
3. **Optimize infrastructure** — scaling, networking, load balancing

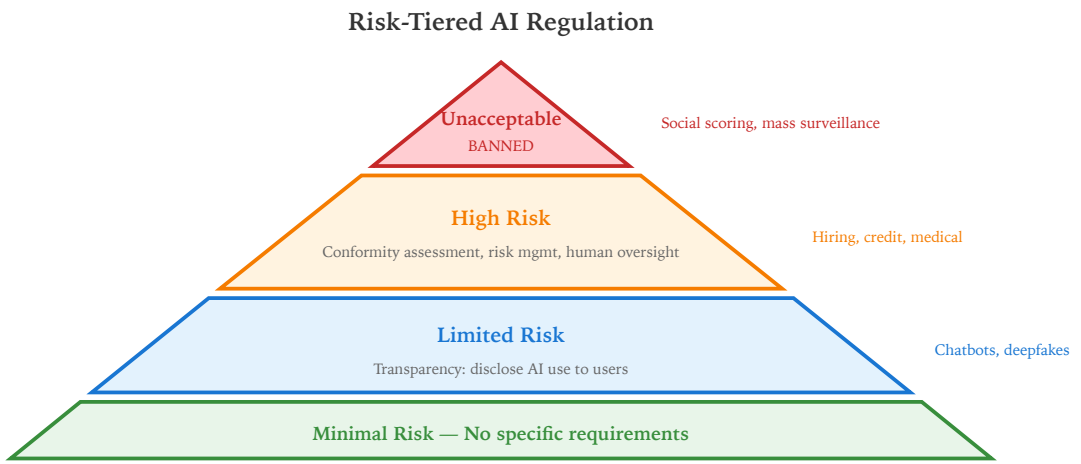
Streaming masks perceived latency: users see tokens as they're generated

Concept Check

You need to deploy a 70B-parameter LLM on a single A100 GPU (80GB VRAM). The model is ~280GB in FP32. What combination of techniques makes this feasible?

Risk-tiered regulation: higher stakes demand more oversight

Core principle: AI governance follows risk-based design — higher-risk systems require stronger oversight.



- **High-risk AI systems** (hiring, credit, medical) require conformity assessment, human oversight, and continuous monitoring

Governance infrastructure: model cards, monitoring, and compliance

Tool	Purpose
Model Cards	Document behavior, limitations, intended uses, demographic performance
Transparency Reports	Ongoing disclosure of incidents, drift, emerging risks
Regulatory Compliance	EU AI Act risk categories, NIST AI RMF
Environmental Disclosure	Carbon footprint and energy use per training/serving run

Monitoring systems track:

Hallucination rate
Drift from known-good baselines

Refusal rate
Over-cautious models lose user trust

Latency drift
Signals infrastructure or load issues

Safety incidents
Flagged outputs, escalations, reports

Important

Safety is part of infrastructure, not just modeling.

Machine unlearning and copyright are unsolved challenges shaping deployment

The Problem

Models memorize training data — some copyrighted, private, or subject to "right to be forgotten"

Machine Unlearning

Selectively remove specific data's influence without full retraining — verification is hard

> **NYT v. OpenAI (2023):** ChatGPT reproduced Pulitzer Prize-winning articles near-verbatim when prompted with opening lines — raising the question of whether models memorize copyrighted content.

- **Open question:** Can we verifiably remove a document's influence from a 70B-parameter model?

Summary: Key Takeaways

1. **Hallucination is a fundamental LLM failure mode** — distinguish intrinsic vs. extrinsic; mitigate with a defense ladder from abstention to human-in-the-loop
2. **Bias enters at every pipeline stage** — data, labeling, reward modeling, deployment, and interpretation. Measurement requires multiple methods (counterfactual pairs, benchmarks, user studies); mitigation spans training, inference, and system design
3. **Safety requires defense-in-depth** — Constitutional AI shapes values; content filtering, aligned LLMs, output filters, and monitoring each address different threat layers
4. **Adversarial robustness is an ongoing arms race** — no single defense suffices; continuous red-teaming is essential
5. **Efficient deployment uses quantization, distillation, and batching** to meet cost/latency requirements without sacrificing safety
6. **Governance follows risk-tiered regulation** — model cards, transparency reports, and regulatory compliance are infrastructure, not paperwork

Course Wrap-Up & Next Week

Congratulations on *(almost)* completing CSE 447/517!

This quarter we covered the full arc of modern NLP:

- Text processing, language models, and neural architectures
- Transformers, pretraining, finetuning, and alignment
- Prompting, reasoning, retrieval, and agentic systems
- Ethics, safety, and responsible deployment

Key message: Building capable NLP systems is inseparable from building safe, fair, and accountable ones. The technical and ethical dimensions are deeply intertwined.

Next week: We'll have guest speakers talking about FlexOlmo (Kevin from Ai2, MoE, Unlearning, Responsible data use), Music + NLP (Praveer your TA!, Diffusion models, Start up progress in the space)

Thank you for a great quarter!