

# Bias and Responsible NLP

**Where bias comes from, and what practice it demands**

Ramaseshan Ramachandran

# BIAS IN LARGE LANGUAGE MODELS

## What is Bias in LLMs?

- ▶ LLMs learn from vast datasets that may contain **societal, historical, or systemic biases**.
- ▶ These biases can be **reflected, amplified, or perpetuated** in generated outputs.

## Types of Bias:

- ▶ **Social Bias** Gender, race, and cultural stereotypes.
- ▶ **Political Bias** Favoring particular ideologies or perspectives.
- ▶ **Confirmation Bias** Reinforcing existing views rather than diversifying perspectives.
- ▶ **Representation Bias** Underrepresentation of minority groups in training data.

# CHALLENGES IN LARGE LANGUAGE MODELS

## Key Challenges:

- ▶ **Bias and Fairness** Ensuring neutral and unbiased responses.
- ▶ **Hallucinations** Generation of **false or misleading** information.
- ▶ **Data Privacy** Risk of memorizing and exposing **sensitive** information.
- ▶ **Computational Costs** High energy consumption and **expensive infrastructure**.
- ▶ **Explainability** Lack of clear interpretability behind model outputs.
- ▶ **Ethical Concerns** Spread of misinformation, deepfakes, and harmful misuse.

## Mitigation Strategies:

- ▶ **Diverse training data** to minimize biases.
- ▶ **Fact-checking / retrieval grounding (RAG)** to detect and reduce hallucinations.
- ▶ **Alignment training** covers instruction tuning and RLHF / RLAIIF, with red-teaming and safety evaluations.
- ▶ **Guardrails and content filters** at deployment time.

## ALIGNING LLMS: SFT AND RLHF

A pre-trained LLM predicts likely next tokens. It does not necessarily give *helpful, honest, or harmless* answers. Alignment closes this gap:

1. **Supervised fine-tuning (SFT)** trains on curated (instruction, ideal response) pairs.
  2. **Reward modeling** has humans rank multiple responses, and a reward model learns to predict these preferences.
  3. **Reinforcement learning (RLHF)** optimizes the policy against the reward model (e.g., PPO), or uses preference-optimization methods such as DPO.
- ▶ RLAIIF replaces some human labels with AI-generated feedback (e.g., Constitutional AI).
  - ▶ Alignment reduces but does not eliminate bias, hallucination, or jailbreaks.

## FAIRNESS AND INCLUSION

Everyone should be treated fairly when using our products and they should work equally well for all people.

# ROBUSTNESS AND SAFETY

- ▶ Should meet high performance standards
- ▶ Should be robust and safe, with rigorous testing and validation to avoid unintended negative consequences

# TRANSPARENCY AND ACCOUNTABILITY

- ▶ Developers and users of the system should be responsible for its actions
- ▶ Rules and regulations governing the use of AI systems should be transparent,
- ▶ Mechanisms should be in place to investigate and punish misuse of the system

# EXPLAINABILITY

- ▶ Should enable users to understand the underlying processes to build trust
- ▶ System should explain its decisions in a clear and understandable way
- ▶ Open for scrutiny

# REFERENCES I