



# CS 435-L1

## Lecture 4

Adnan Masood, PhD.

---

## **Administrivia**

---

Ethics & AI – A Taxonomy

---

Recap Quiz

---

Guest Lecture Heather Dawe

---

Readings

---

Coding Assignment - Debiaser

# Responsible Innovation: Bridging Ethics and LLMs

## Guest Lecture by Heather Dawe

Generative AI has opened up remarkable possibilities for creative output, personalized recommendations, and complex decision-making. At the same time, it introduces critical ethical considerations around bias, fairness, accountability, and transparency. In this session, we will delve into how organizations can develop and deploy large language models (LLMs) and related generative technologies responsibly. By drawing on real-world examples, we will examine the socio-technical implications of AI, the regulatory frameworks that shape its governance, and proven strategies for mitigating risk. Attendees will gain practical insights into why ethical AI matters, how to establish guardrails for transparency and fairness, and how to foster trust in a rapidly evolving digital landscape.



### About the speaker

Heather Dawe leads the UK data science team at UST. In a career that has spanned multiple industries and disciplines, she has consistently sought out and delivered innovative ways to use data and technology to improve people's lives. Away from data and technology, Heather loves mountains. She walks and runs among them, climbs them, writes about them and paints them. In 2021 she sat on the book prize jury at the Banff Mountain Film and Book festival and she is Guest Editor of the 2022 edition of The Himalayan Journal.

**Ethics** is the study of moral principles that guide human behavior and decision-making.

<https://github.com/adnanmasood/cs-435-spring-2025/blob/main/ethical-ai.md>

**Morals** are personal or communal principles about right and wrong, often shaped by culture, upbringing, or belief systems.

<https://github.com/adnanmasood/cs-435-spring-2025/blob/main/ethical-ai.md>

**Ethics** typically refers to external standards or frameworks (e.g., professional codes) guiding behavior.

**Morals** are more subjective, stemming from individual or societal values and personal conscience.

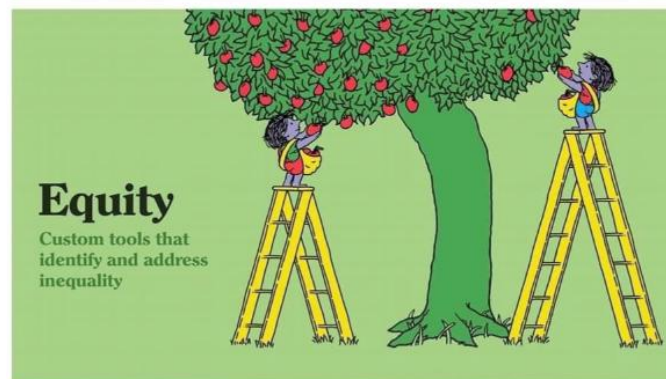
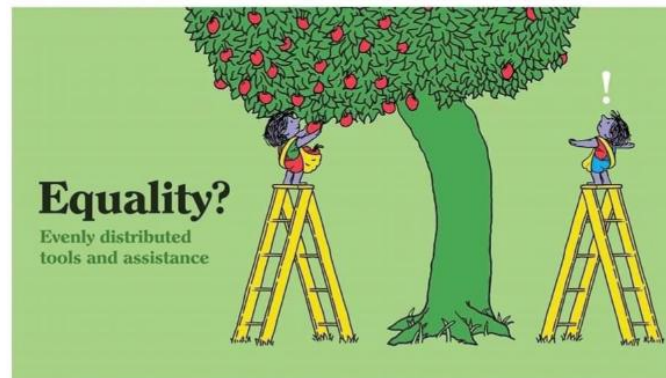
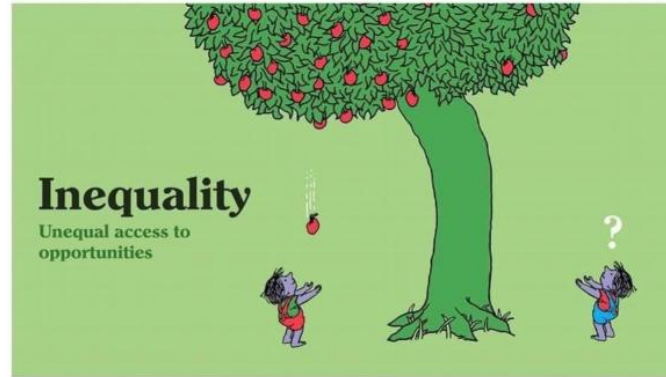
**Ethical AI** is the development and use of AI systems in a manner that upholds **moral principles** (like fairness, transparency, privacy, and accountability) and aims to **benefit** society without causing **harm** or **discrimination**.

<https://github.com/adnanmasood/cs-435-spring-2025/blob/main/ethical-ai.md>

**Fairness** is Ensuring (AI)  
decisions do not  
systematically  
disadvantage any group.

<https://github.com/adnanmasood/cs-435-spring-2025/blob/main/fairness.md>





**Equality** is providing  
uniform (AI) outcomes  
or opportunities to all  
users.

<https://github.com/adnanmasood/cs-435-spring-2025/blob/main/equity-vs-equality.md>

**Equity** is adjusting (AI) systems to address different needs or contexts fairly.

<https://github.com/adnanmasood/cs-435-spring-2025/blob/main/equity-vs-equality.md>

**Trust** is users'  
confidence in (AI)'s  
reliability, integrity, and  
benefit.

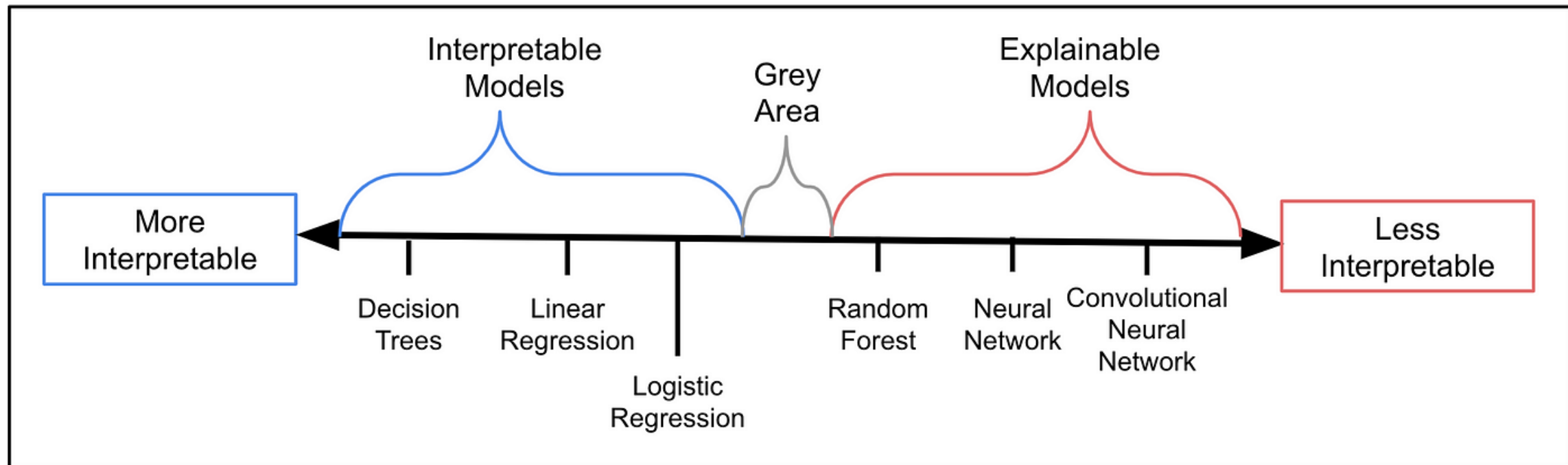
**Accountability** is  
holding (AI) creators and  
operators responsible  
for system impacts.

<https://github.com/adnanmasood/cs-435-spring-2025/blob/main/accountability.md>

**Explainability** is clearly  
describing how (AI)  
arrives at specific  
outputs.

**Interpretability** is  
making (AI's) internal  
logic understandable to  
humans.

<https://github.com/adnanmasood/cs-435-spring-2025/blob/main/interpretability.md>





**Transparency is  
Openness about (AI's)  
purpose, data, and  
processes.**

<https://github.com/adnanmasood/cs-435-spring-2025/blob/main/transparency.md>

**Privacy** is safeguarding  
personal data and  
preventing unauthorized  
(AI) data use.

**Bias is systemic skew in  
AI that unfairly **favors** or  
**harms** certain groups.**

<https://github.com/adnanmasood/cs-435-spring-2025/blob/main/bias.md>

# **Glossary of Gen AI Terms**

<https://github.com/adnanmasood/cs-435-spring-2025/blob/main/llm-glossary.md>

# THE WALL STREET JOURNAL.

English Edition ▾ | Print Edition | Video | Audio | Latest Headlines | More ▾

Adnan Masood ▾

Latest World Business U.S. Politics Economy Tech Markets & Finance Opinion Arts Lifestyle Real Estate Personal Finance Health Style Sports 🔍

CIO JOURNAL

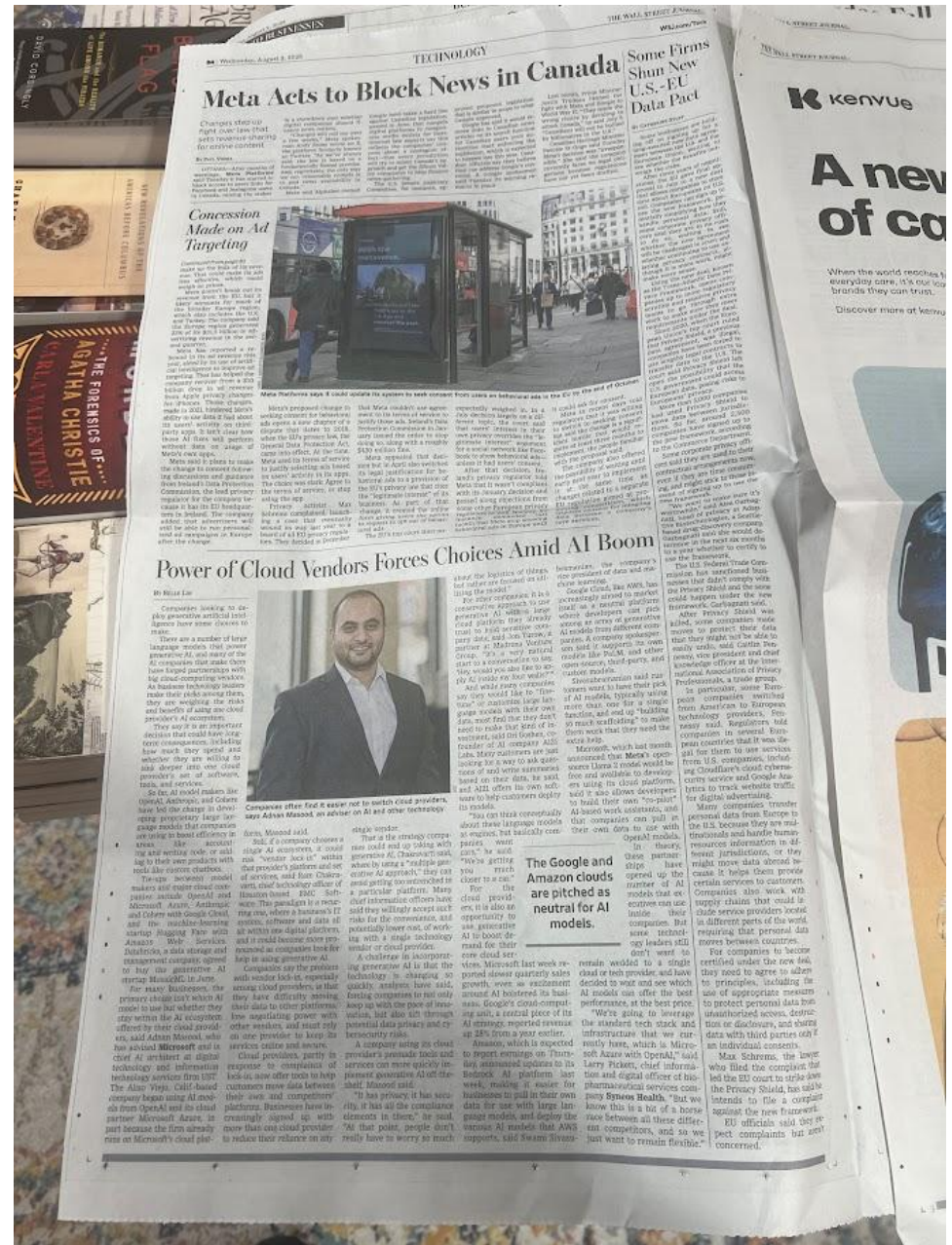
## U.S. Businesses Already Love DeepSeek

The Chinese company's new model promises to lower the cost of AI for enterprises—even amid concerns about cybersecurity and geopolitics

By Isabelle Bousquette [Follow](#) and

Adnan Masood, chief AI architect of digital technology and information-technology services firm UST, said, “The anxiety is that you’re feeding sensitive corporate data into a system that originated from a strategic adversary, no matter how ingenious the engineering.”

“On the flip side, drastically lower costs and advanced capabilities tempt executives to risk these concerns for competitive advantage,” he added.



The image features a central white circle with a thick green border. Inside this circle, the word "Readings" is written in a bold, white, sans-serif font. Surrounding the central circle are several abstract elements: a small orange circle with a white outline to the left; two white wavy lines to the upper left; a small orange circle with a white outline to the upper right; and a grid of small white dots to the lower right.

# Readings

---

# Attention Is All You Need

---

Ashish Vaswani\*  
Google Brain  
avaswani@google.com

Noam Shazeer\*  
Google Brain  
noam@google.com

Niki Parmar\*  
Google Research  
nikip@google.com

Jakob Uszkoreit\*  
Google Research  
usz@google.com

Llion Jones\*  
Google Research  
llion@google.com

Aidan N. Gomez\* †  
University of Toronto  
aidan@cs.toronto.edu

Łukasz Kaiser\*  
Google Brain  
lukaszkaiser@google.com

Illia Polosukhin\* †  
illia.polosukhin@gmail.com

## Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with

- 
- <https://github.com/adnanmasood/cs-435-spring-2025/blob/main/attention-is-all-you-really-need.md>



- <https://github.com/adnamasood/cs-435-spring-2025/blob/main/On-the-Dangers-of-Stochastic-Parrots.md>

## On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

Emily M. Bender\*  
ebender@uw.edu  
University of Washington  
Seattle, WA, USA

Angelina McMillan-Major  
aymm@uw.edu  
University of Washington  
Seattle, WA, USA

Timnit Gebru\*  
timnit@blackinai.org  
Black in AI  
Palo Alto, CA, USA

Shmargaret Shmitchell  
shmargaret.shmitchell@gmail.com  
The Aether


### ABSTRACT

The past 3 years of work in NLP have been characterized by the development and deployment of ever larger language models, especially for English. BERT, its variants, GPT-2/3, and others, most recently Switch-C, have pushed the boundaries of the possible both through architectural innovations and through sheer size. Using these pretrained models and the methodology of fine-tuning them for specific tasks, researchers have extended the state of the art on a wide array of tasks as measured by leaderboards on specific benchmarks for English. In this paper, we take a step back and ask: How big is too big? What are the possible risks associated with this technology and what paths are available for mitigating those risks? We provide recommendations including weighing the environmental and financial costs first, investing resources into curating and carefully documenting datasets rather than ingesting everything on the web, carrying out pre-development exercises evaluating how the planned approach fits into research and development goals and supports stakeholder values, and encouraging research directions beyond ever larger language models.

### CCS CONCEPTS

• Computing methodologies → Natural language processing.

#### ACM Reference Format:

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In *Conference on Fairness, Accountability, and Transparency (FACET '21)*, March 3–10, 2021, Virtual Event, Canada. ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/3442188.3445922>

### 1 INTRODUCTION

One of the biggest trends in natural language processing (NLP) has been the increasing size of language models (LMs) as measured by the number of parameters and size of training data. Since 2018

\*Joint first authors.



This work is licensed under a Creative Commons Attribution International 4.0 License.  
FACET '21, March 3–10, 2021, Virtual Event, Canada  
ACM ISBN 978-1-4503-8309-7/21/03.  
<https://doi.org/10.1145/3442188.3445922>

alone, we have seen the emergence of BERT and its variants [39, 70, 74, 113, 146], GPT-2 [106], T-NLG [112], GPT-3 [25], and most recently Switch-C [43], with institutions seemingly competing to produce ever larger LMs. While investigating properties of LMs and how they change with size holds scientific interest, and large LMs have shown improvements on various tasks (§2), we ask whether enough thought has been put into the potential risks associated with developing them and strategies to mitigate these risks.

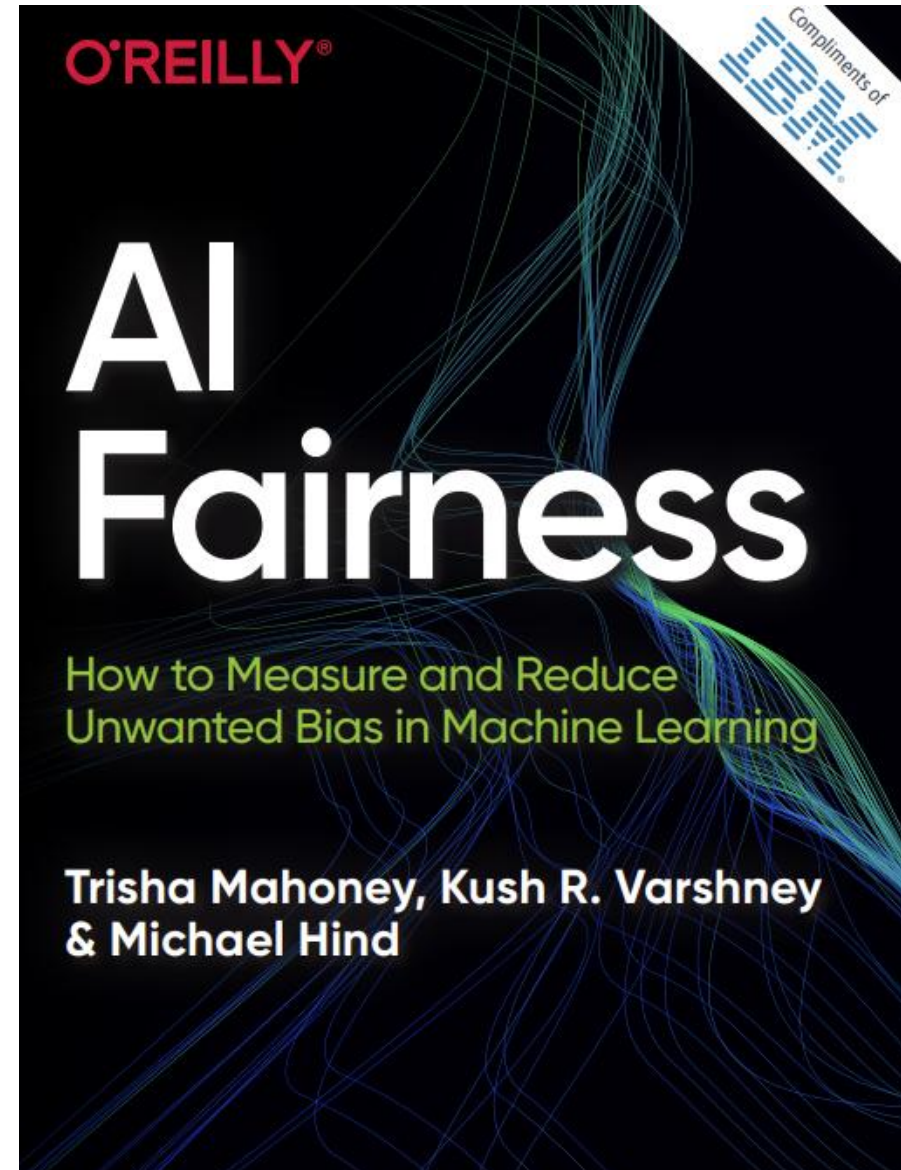
We first consider environmental risks. Echoing a line of recent work outlining the environmental and financial costs of deep learning systems [129], we encourage the research community to prioritize these impacts. One way this can be done is by reporting costs and evaluating works based on the amount of resources they consume [57]. As we outline in §3, increasing the environmental and financial costs of these models doubly punishes marginalized communities that are least likely to benefit from the progress achieved by large LMs and most likely to be harmed by negative environmental consequences of its resource consumption. At the scale we are discussing (outlined in §2), the first consideration should be the environmental cost.

Just as environmental impact scales with model size, so does the difficulty of understanding what is in the training data. In §4, we discuss how large datasets based on texts from the Internet overrepresent hegemonic viewpoints and encode biases potentially damaging to marginalized populations. In collecting ever larger datasets we risk incurring documentation debt. We recommend mitigating these risks by budgeting for curation and documentation at the start of a project and only creating datasets as large as can be sufficiently documented.

As argued by Bender and Koller [14], it is important to understand the limitations of LMs and put their success in context. This not only helps reduce hype which can mislead the public and researchers themselves regarding the capabilities of these LMs, but might encourage new research directions that do not necessarily depend on having larger LMs. As we discuss in §5, LMs are not performing natural language understanding (NLU), and only have success in tasks that can be approached by manipulating linguistic form [14]. Focusing on state-of-the-art results on leaderboards without encouraging deeper understanding of the mechanism by which they are achieved can cause misleading results as shown



- 
- <https://github.com/adnanmasood/cs-435-spring-2025/blob/main/aif360.md>



## DeepSeek-V3 Technical Report

DeepSeek-AI

research@deepseek.com

### Abstract

Introducing DeepSeek-V3, a strong Mixture-of-Experts (MoE) language model with 671B total parameters with 37B activated for each token. To achieve efficient inference and cost-effective training, DeepSeek-V3 adopts Multi-head Latent Attention (MLA) and DeepSeekMoE architecture, which were thoroughly validated in DeepSeek-V2. Furthermore, DeepSeek-V3 pioneers a zero-loss-free strategy for load balancing and sets a multi-token prediction training for stronger performance. We pre-train DeepSeek-V3 on 14.8 trillion diverse and high-quality tokens, followed by Supervised Fine-Tuning and Reinforcement Learning stages to assess its capabilities. Comprehensive evaluations reveal that DeepSeek-V3 outperforms open-source models and achieves performance comparable to leading closed-source models. Despite its excellent performance, DeepSeek-V3 requires only 2.788M H800 GPU hours for training. In addition, its training process is remarkably stable. Throughout the entire process, we did not experience any irrecoverable loss spikes or perform any rollbacks. Checkpoints are available at <https://github.com/deepseek-ai/DeepSeek-V3>.

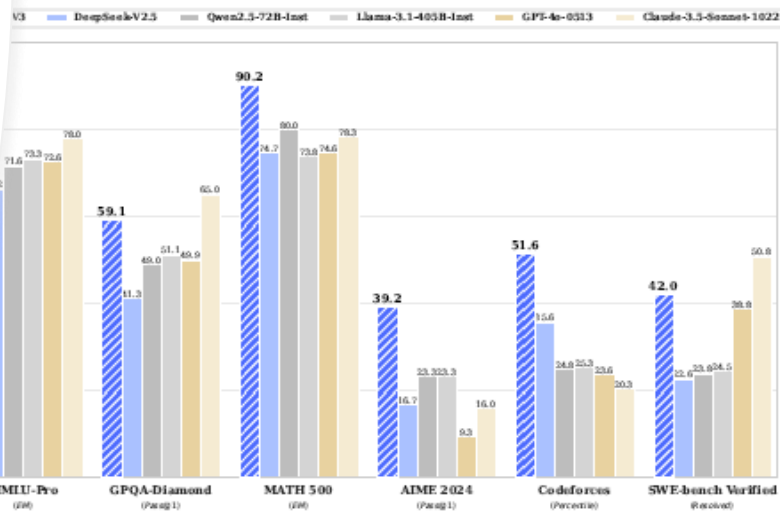


Figure 1 | Benchmark performance of DeepSeek-V3 and its counterparts.

## DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

DeepSeek-AI

research@deepseek.com

### Abstract

We introduce our first-generation reasoning models, DeepSeek-R1-Zero and DeepSeek-R1. DeepSeek-R1-Zero, a model trained via large-scale reinforcement learning (RL) without supervised fine-tuning (SFT) as a preliminary step, demonstrates remarkable reasoning capabilities. Through RL, DeepSeek-R1-Zero naturally emerges with numerous powerful and intriguing reasoning behaviors. However, it encounters challenges such as poor readability, and language gaps. To address these issues and further enhance reasoning performance, we introduce DeepSeek-R1, which incorporates multi-stage training and cold-start data before RL. DeepSeek-R1 achieves performance comparable to OpenAI-o1-1217 on reasoning tasks. To support the research community, we open-source DeepSeek-R1-Zero, DeepSeek-R1, and six dense models (7B, 8B, 14B, 32B, 70B) distilled from DeepSeek-R1 based on Qwen and Llama.

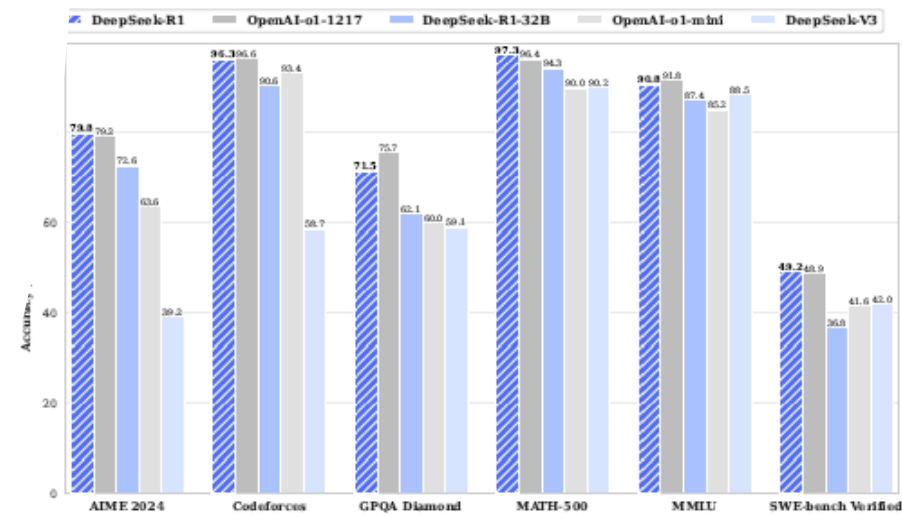
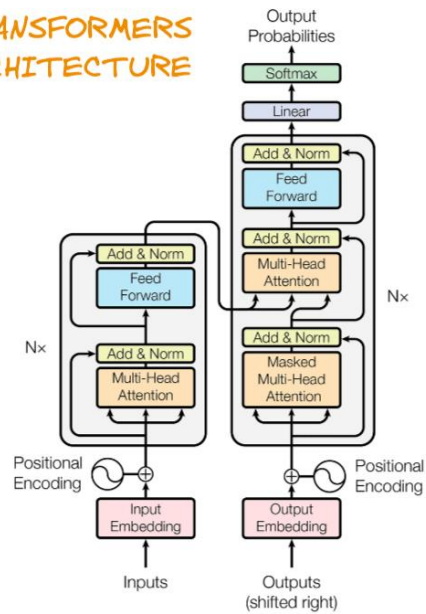


Figure 1 | Benchmark performance of DeepSeek-R1.

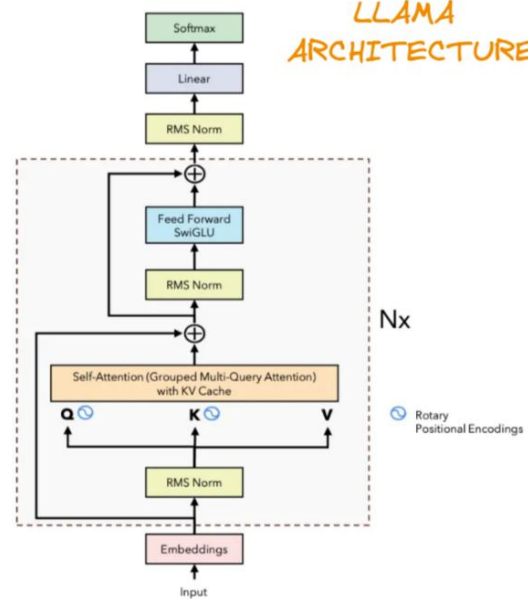
- <https://github.com/adnanmasood/cs-435-spring-2025/tree/main>

## TRANSFORMERS ARCHITECTURE



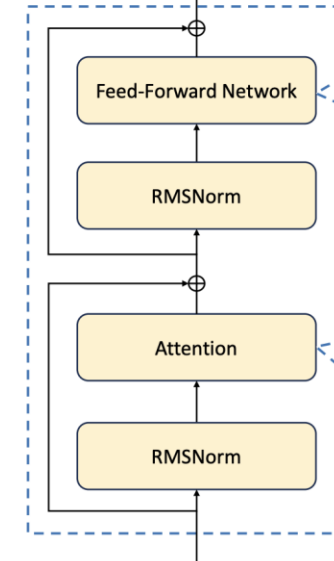
VS.

## LLAMA ARCHITECTURE

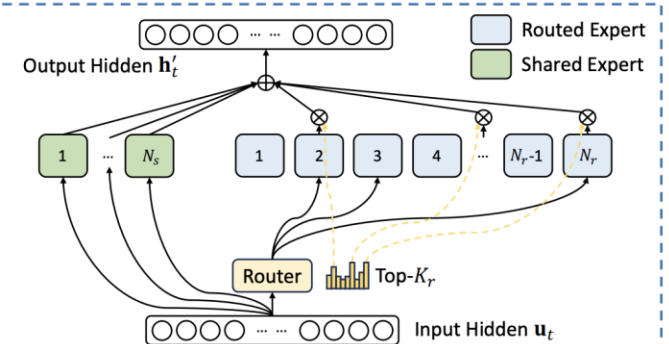


Zoumana K.

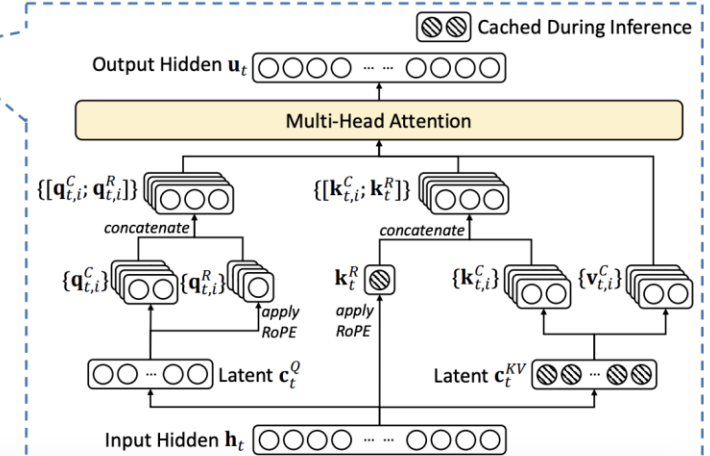
Transformer Block  $\times L$



## DeepSeekMoE



## Multi-Head Latent Attention (MLA)



Aspect	Vanilla Transformer	LLaMA Architecture	DeepSeek Architecture	Rationale / Notes
Feed-Forward Layer	Single dense FFN per layer.	Single dense FFN per layer with <b>SwiGLU</b> activation (instead of ReLU/GeLU).	<b>Mixture-of-Experts (MoE)</b> in <i>DeepSeekMoE</i> : multiple parallel experts + a router that picks top-k experts per token.	<b>•Vanilla &amp; LLaMA</b> : One FFN scales with model dimension. <b>•DeepSeek</b> : Gains capacity cheaply by only activating a few experts at a time, reducing compute cost.
Attention Type	Standard Multi-Head Self-Attention.	Multi-Head with grouped <b>Multi-Query</b> or “Grouped Query” in LLaMA.	<b>Multi-Head Latent Attention (MLA)</b> : compress K/V into smaller latent dimension + small separate decoupled positional part.	<b>•Vanilla</b> : Q/K/V each dimension $\sim d_{\text{model}}$ . <b>•LLaMA</b> : reduces overhead with “grouped queries.” <b>•DeepSeek</b> : shrinks Key/Value cache memory using MLA.
Normalization	LayerNorm typically after Add.	<b>RMSNorm</b> instead of LayerNorm (lighter, stable).	<b>RMSNorm</b> as well, but often multiple RMSNorm calls per block (before/after certain ops).	<b>•RMSNorm</b> is cheaper, uses only root-mean-square statistics. Minimally different but often better for training stability.
Positional Encoding	Often learned positional embeddings or sinusoids.	<b>Rotary Positional Embedding (RoPE)</b> .	Same <b>Rotary</b> style, but the MLA block has a special “decoupled” rotational dimension ( $k^R$ and $q^R$ ).	<b>•Vanilla</b> : either sinusoidal or learned offset. <b>•LLaMA</b> : RoPE helps with extrapolation, especially for longer contexts. <b>•DeepSeek</b> : extends that for MLA compression.
Parameter Scaling	Must scale entire layer as model size grows (dense).	Also a dense approach, but uses some weight tweaks like SwiGLU, RMSNorm to be more parameter-efficient.	<b>MoE</b> approach with large <i>total</i> params, but smaller <i>activated</i> params per token (top-k experts).	<b>•Vanilla / LLaMA</b> : All tokens pass all parameters. <b>•DeepSeek</b> : Gains parameter count cheaply, but each token only uses a fraction of them, saving compute.
Caching in Inference	Each head stores K/V $\sim \text{hidden\_dim} \times \text{num\_heads}$ .	Similar to standard, but can have fewer “heads” that store unique K/V due to Multi-Query or Grouped Query.	<b>MLA</b> heavily compresses K/V into a smaller latent dimension + a small decoupled vector.	<b>•DeepSeek</b> significantly reduces memory overhead for long context, making it easier to handle big batch or 128K context lengths.
Multi-Token Prediction	Usually 1-token-ahead cross-entropy.	Also 1-token-ahead, though LLaMA might be used in speculative decoding.	<b>Multi-Token Prediction (MTP)</b> : the model has extra modules that predict t+2, t+3, etc. in parallel.	<b>•Vanilla &amp; LLaMA</b> : Standard “predict next token.” <b>•DeepSeek</b> : Gains more training signal, can also accelerate inference by speculative decoding.
Quantization / Precision	FP32 or BF16.	LLaMA often uses BF16 or 8-bit “load” for inference.	<b>FP8</b> “fine-grained” quantization in training (plus partial-sum accumulation in higher precision).	<b>•DeepSeek</b> invests heavily in advanced FP8 kernels and specialized scaling $\rightarrow$ big memory & speed gains in training large MoE.
Load Balancing	N/A (no MoE).	N/A (no MoE).	<b>Aux-loss-free</b> or minimal auxiliary penalty to ensure experts are balanced (no single-expert collapse).	<b>•DeepSeek</b> must handle “routing collapse” when tokens pile into one expert. They do so by adjusting gating bias automatically, without large penalty on main loss.
Block Structure	“Add & Norm” after each sub-layer (or “Pre-LN” variant).	“RMSNorm” + feed-forward with SwiGLU. Usually a simpler “pre-normalization.”	<b>RMSNorm</b> + a “DeepSeekMoE feed-forward” + MLA-based attention. Usually multiple RMSNorm calls inside.	<b>•Vanilla</b> : Typically “(Attn $\rightarrow$ Add+Norm) $\rightarrow$ (FFN $\rightarrow$ Add+Norm).” <b>•LLaMA</b> : Has a “lighter” RMSNorm approach. <b>•DeepSeek</b> : Adds MoE and MLA in that pattern, plus more norms.
Typical Model Sizes	Anywhere from $\sim 100\text{M}$ to tens of billions.	LLaMA family from 7B to 70B (and rumors of 400B).	DeepSeek-V3 is $\sim 671\text{B}$ total parameters, but only 37B “activated” per token.	<b>•DeepSeek</b> can “scale up” total parameters thanks to MoE while controlling inference cost.
Key Advantage	Classic “universal” standard.	Strong efficient variant of standard Transformer with better training stability.	Extremely large scale while being <i>economical</i> at inference/training (MoE + MLA + MTP + FP8).	Summarizes each approach’s main benefit.



# GPT-4 Technical Report

OpenAI\*

## Abstract

We report the development of GPT-4, a large-scale, multimodal model which can accept image and text inputs and produce text outputs. While less capable than humans in many real-world scenarios, GPT-4 exhibits human-level performance on various professional and academic benchmarks, including passing a simulated bar exam with a score around the top 10% of test takers. GPT-4 is a Transformer-based model pre-trained to predict the next token in a document. The post-training alignment process results in improved performance on measures of factuality and adherence to desired behavior. A core component of this project was developing infrastructure and optimization methods that behave predictably across a wide range of scales. This allowed us to accurately predict some aspects of GPT-4’s performance based on models trained with no more than 1/1,000th the compute of GPT-4.

## 1 Introduction

This technical report presents GPT-4, a large multimodal model capable of processing image and text inputs and producing text outputs. Such models are an important area of study as they have the potential to be used in a wide range of applications, such as dialogue systems, text summarization, and machine translation. As such, they have been the subject of substantial interest and progress in recent years [1–34].

One of the main goals of developing such models is to improve their ability to understand and generate natural language text, particularly in more complex and nuanced scenarios. To test its capabilities in such scenarios, GPT-4 was evaluated on a variety of exams originally designed for humans. In these evaluations it performs quite well and often outscores the vast majority of human test takers. For example, on a simulated bar exam, GPT-4 achieves a score that falls in the top 10% of test takers. This contrasts with GPT-3.5, which scores in the bottom 10%.

On a suite of traditional NLP benchmarks, GPT-4 outperforms both previous large language models and most state-of-the-art systems (which often have benchmark-specific training or hand-engineering). On the MMLU benchmark [35, 36], an English-language suite of multiple-choice questions covering 57 subjects, GPT-4 not only outperforms existing models by a considerable margin in English, but also demonstrates strong performance in other languages. On translated variants of MMLU, GPT-4 surpasses the English-language state-of-the-art in 24 of 26 languages considered. We discuss these model capability results, as well as model safety improvements and results, in more detail in later sections.

This report also discusses a key challenge of the project, developing deep learning infrastructure and optimization methods that behave predictably across a wide range of scales. This allowed us to make predictions about the expected performance of GPT-4 (based on small runs trained in similar ways) that were tested against the final run to increase confidence in our training.

Despite its capabilities, GPT-4 has similar limitations to earlier GPT models [1, 37, 38]: it is not fully reliable (e.g. can suffer from “hallucinations”), has a limited context window, and does not learn

\*Please cite this work as “OpenAI (2023)”. Full authorship contribution statements appear at the end of the document. Correspondence regarding this technical report can be sent to [gpt4-report@openai.com](mailto:gpt4-report@openai.com)

Greg Brockman, Shantanu Jain, Kyle Kopic, Michael Petrov, Nikolai Tezak, Amin Tootoonchian, Chelsea Voss, Qiming Yuan

### Distributed training infrastructure<sup>11</sup>

Greg Brockman, Trevor Cai, Chris Hesse, Shantanu Jain, Yongjik Kim, Kyle Kopic, Mateusz Litwin, Jakub Pachocki, Mikhail Pavlov, Szymon Sidor, Nikolai Tezak, Madeleine Thompson, Amin Tootoonchian, Qiming Yuan

### Hardware correctness<sup>11</sup>

Greg Brockman, Shantanu Jain, Kyle Kopic, Michael Petrov, Nikolai Tezak, Amin Tootoonchian, Chelsea Voss, Qiming Yuan

### Optimization & architecture<sup>11</sup>

Igor Babuschkin, Mo Bavarian, Adrien Ecoffet, David Farhi, Jesse Han, Ingmar Kanitscheider, Daniel Levy, Jakub Pachocki, Alex Paino, Mikhail Pavlov, Nick Ryder, Szymon Sidor, Jie Tang, Jerry Tworek, Tao Xu

### Training run babysitting<sup>11</sup>

Suchir Balaji, Mo Bavarian, Greg Brockman, Trevor Cai, Chris Hesse, Shantanu Jain, Roger Jiang, Yongjik Kim, Kyle Kopic, Mateusz Litwin, Jakub Pachocki, Alex Paino, Mikhail Pavlov, Michael Petrov, Nick Ryder, Szymon Sidor, Nikolai Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Chelsea Voss, Ben Wang, Tao Xu, Qiming Yuan

## Long context

### Core contributors<sup>11</sup>

Gabriel Goh *Long context co-lead*  
Łukasz Kaiser *Long context lead*  
Ben Wang *Attention architecture lead*  
Clemens Winter *Long context co-lead*

### Long context research<sup>11</sup>

Mo Bavarian, Gabriel Goh, Heewoo Jun, Łukasz Kaiser, Chak Ming Li, Ben Wang, Clemens Winter

### Long context kernels<sup>11</sup>

Phil Tillet

Long Ouyang, Raul Puri, Pranav Shyam, Tao Xu

### Alignment data<sup>11</sup>

Long Ouyang

### Training run babysitting<sup>11</sup>

Trevor Cai, Kyle Kopic, Daniel Levy, David Mély, Reiichiro Nakano, Hyeonwoo Noh, Mikhail Pavlov, Raul Puri, Amin Tootoonchian

### Deployment & post-training<sup>11</sup>

Ilge Akkaya, Mark Chen, Jamie Kiros, Rachel Lim, Reiichiro Nakano, Raul Puri, Jiayi Weng

## Reinforcement Learning & Alignment

### Core contributors<sup>11</sup>

Greg Brockman *Core infrastructure author*  
Arka Dhar *Human data product manager*  
Liam Fedus *Data flywheel lead*  
Tarun Gogineni *Model creativity*  
Rapha Gontijo-Lopes *Synthetic data*  
Joshua Gross *Data collection engineering co-lead*  
Johannes Heidecke *Refusals & model safety co-lead*  
Joost Huizinga *Initial fine-tuning derisking*  
Teddy Lee *Human data product manager*  
Jan Leike *Alignment co-lead*  
Ryan Lowe *Alignment co-lead*  
Luke Metz *Infrastructure lead, ChatML format lead*  
Long Ouyang *IF data collection lead*  
John Schulman *Overall lead*  
Jerry Tworek *Code lead*  
Carroll Wainwright *IF data infrastructure lead*  
Jonathan Ward *Data collection engineering co-lead*  
Jiayi Weng *RL Infrastructure author*  
Sarah Yoo *Human data operations manager*  
Wojciech Zaremba *Human data lead*  
Chong Zhang *Refusals & model safety co-lead*  
Shengjia Zhao *Reward model lead*  
Barret Zoph *Overall training lead*

### Dataset contributions<sup>11</sup>

<https://arxiv.org/pdf/2303.08774>

# Assignment

# Unveiling Gender Bias: Statistical Analyses of Salary Prediction Data

*By Dr. Adnan Masood*

## Purpose of this Notebook

In this notebook, we will explore salary data and see if there is any difference in salaries for people based on their gender. We will do several simple analyses to see if there is any unfairness in how salaries are distributed.


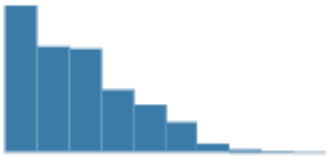

We are using simple explanations so that complex ideas become easier to digest:

- We will break down statistical methods, such as t-tests, into simpler terms.
- We will showcase various data visualizations to spot any inequality or bias.
- We will also look into how to potentially fix or reduce bias (called **debiasing**), with ample detail provided so each step is clear.

<https://github.com/adnanmasood/cs-435-spring-2025/blob/main/gender-bias.ipynb>

Salary\_Data.csv (348.43 kB)

Detail Compact Column

# Age Age of Employee	Gender Gender of Employee	Education Level Education Level of Employee	Job Title Role of Employee	# Years of Experien... Experience of Employee	# Salary Monthly Salary
 2162	Male 55% Female 45% Other (16) 0%	Bachelor's Degree 34% Master's Degree 23% Other (2864) 43%	Software Engineer 8% Data Scientist 7% Other (5733) 86%	 034	 350250k
32	Male	Bachelor's	Software Engineer	5	90000
28	Female	Master's	Data Analyst	3	65000
45	Male	PhD	Senior Manager	15	150000



## Why not just remove the Gender column?

Eliminating the Gender feature (**“fairness through unawareness”**) can be misleading. Other columns—such as certain job titles or zip codes—can be proxies for gender. Thus, the model might still learn gender-based biases indirectly.

# Assignment & Objectives

## Unveiling Gender Bias in Salary Data

**1.Objective:** Examine salary data to determine if gender-based pay differences exist.

**2.Methods:**

- 1.Perform statistical analyses (t-tests, Theil index).
- 2.Visualize distributions (histograms, box plots, scatter plots) for insights on bias.
- 3.Investigate potential unfairness using demographic parity and disparate impact measures.

## **1.Build a Simple Predictive Model**

- 1. Objective:** Train a regression model to predict salaries from factors like Age, Years of Experience, and Gender.
- 2. Experiment:** Compare predicted salaries for identical profiles differing only by Gender, assessing any **learned bias**.

## **2.Debias with a Reweighting Technique**

- 1. Objective:** Adjust training sample weights to counteract observed imbalances.
- 2. Approach:**
  - 1.Compute how frequently each gender appears vs. a “desired” distribution.
  - 2.Assign higher weights to underrepresented groups.
  - 3.Retrain the model, then compare whether male vs. female salary predictions become more aligned.

## Bonus Question

**Suggest another method (besides reweighing) to reduce bias** in your model. For instance, you could use "**post-processing**" techniques (adjusting predictions after the model is trained) or **adversarial debiasing** (where a secondary model is trained to neutralize sensitive information).

Announcing

<https://portal.azure.com/#home>

<https://ml.azure.com>



# Azure AI Foundry



Copilot Studio



Visual Studio



GitHub



Azure AI  
Foundry SDK



## Model Catalog

Foundational models

Open-source models

Task models

Industry models



Azure  
OpenAI Service



Azure  
AI Search



Azure AI  
Agent Service



Azure AI  
Content Safety

Evaluations

Customization

Governance

Monitoring

## Observability