Large Language Models: Challenges at work

Professor Tambe

tambe@wharton.upenn.edu

Some vocabulary: "tokens"

Words and subwords that models use as basic units

🔍 Try it yourself:

OpenAI Tokenizer Demo → https://platform.openai.com/tokenizer

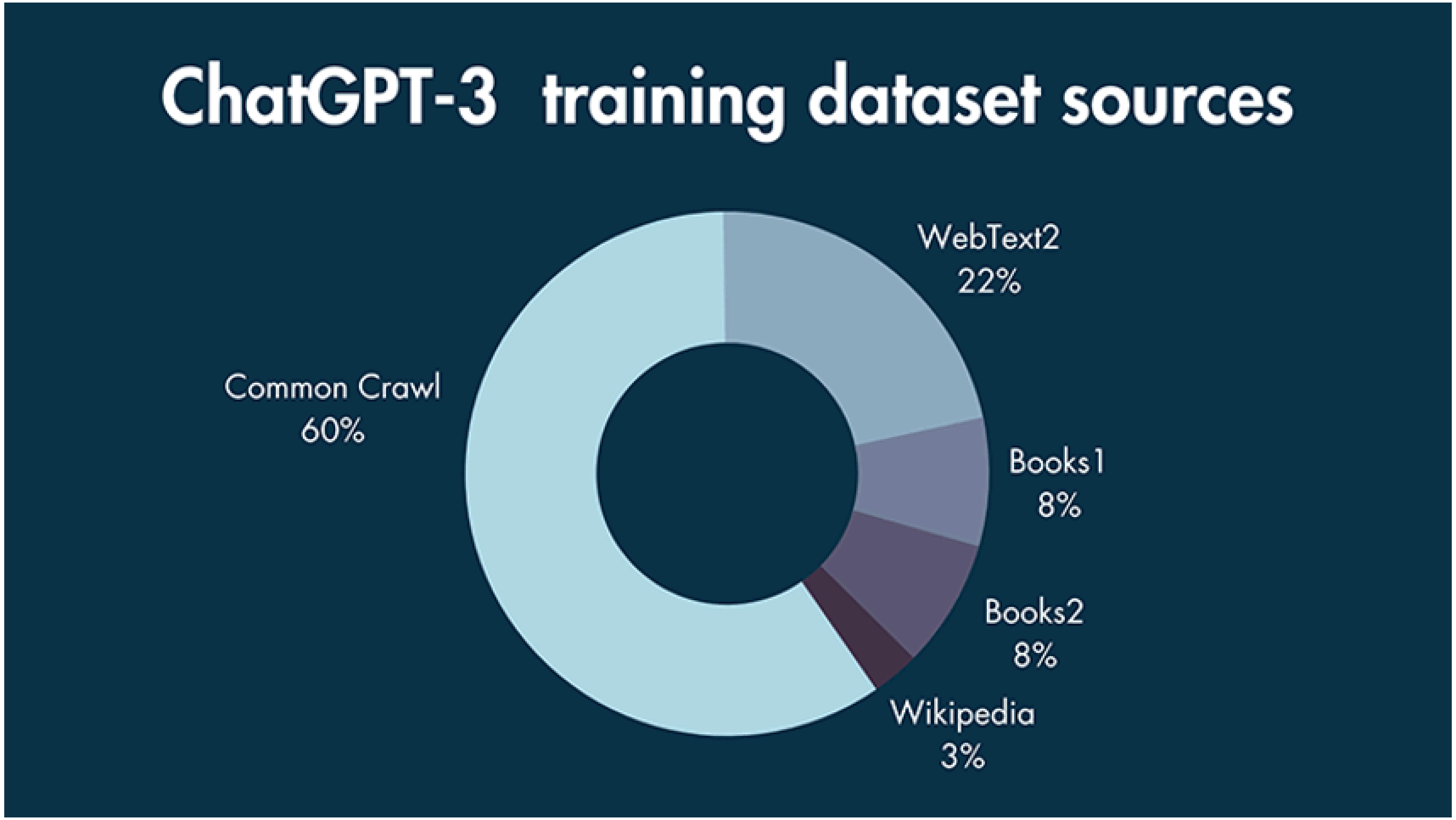
GPT-4o

GPT-4o is our most advanced multimodal model that's faster and cheaper than GPT-4 Turbo with stronger vision capabilities. The model has 128K context and an October 2023 knowledge cutoff.

Learn about GPT-4o ↗

Model	Pricing	Pricing with Batch API*
gpt-4o	$2.50 / 1M input tokens	$1.25 / 1M input tokens
	$1.25 / 1M cached** input tokens	
	$10.00 / 1M output tokens	$5.00 / 1M output tokens
gpt-4o-2024-08-06	$2.50 / 1M input tokens	$1.25 / 1M input tokens

Some vocabulary: training data



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

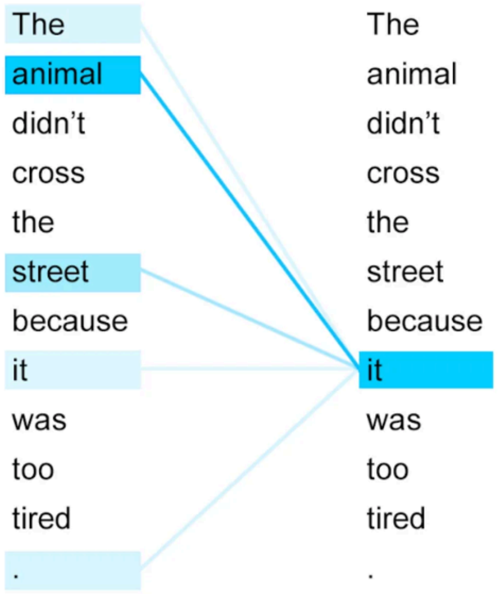
Abstract

Some intuition about Transformer-based models

- **Key innovation**: self-attention mechanism
- Captures long-range dependencies among words in the data
- Allows for parallel processing of input sequences (Scalable)

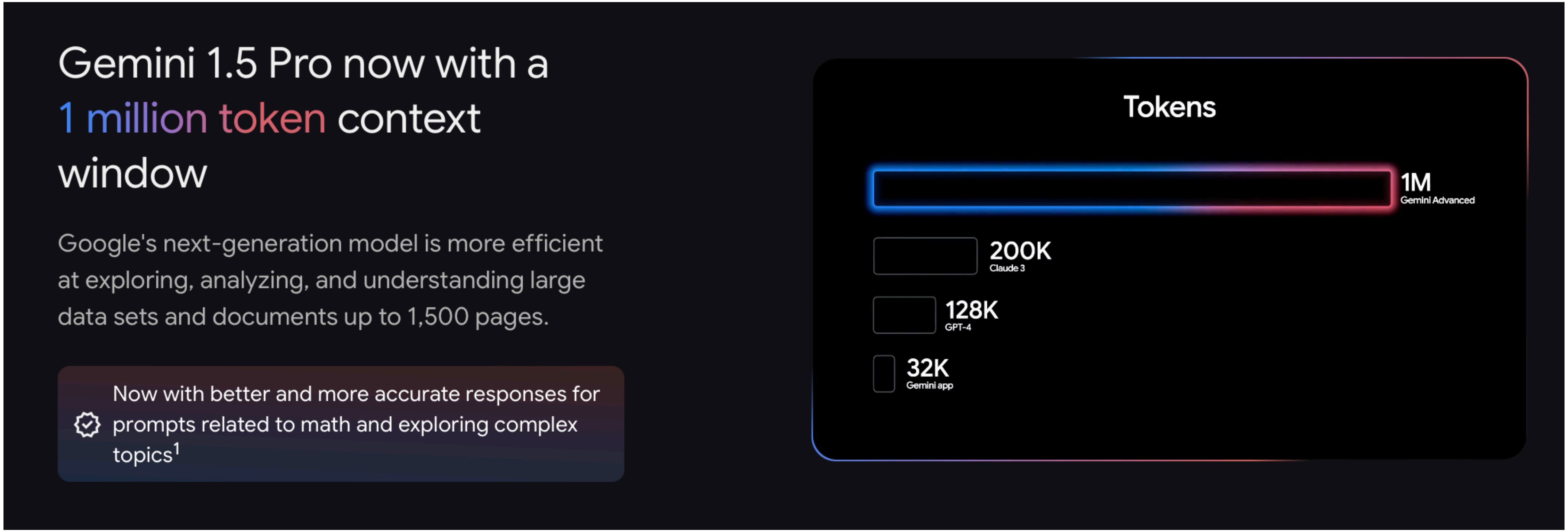
Attending words from anywhere in the context window to learn sentence meaning and structure

"The animal didn't cross the street because it was too tired."

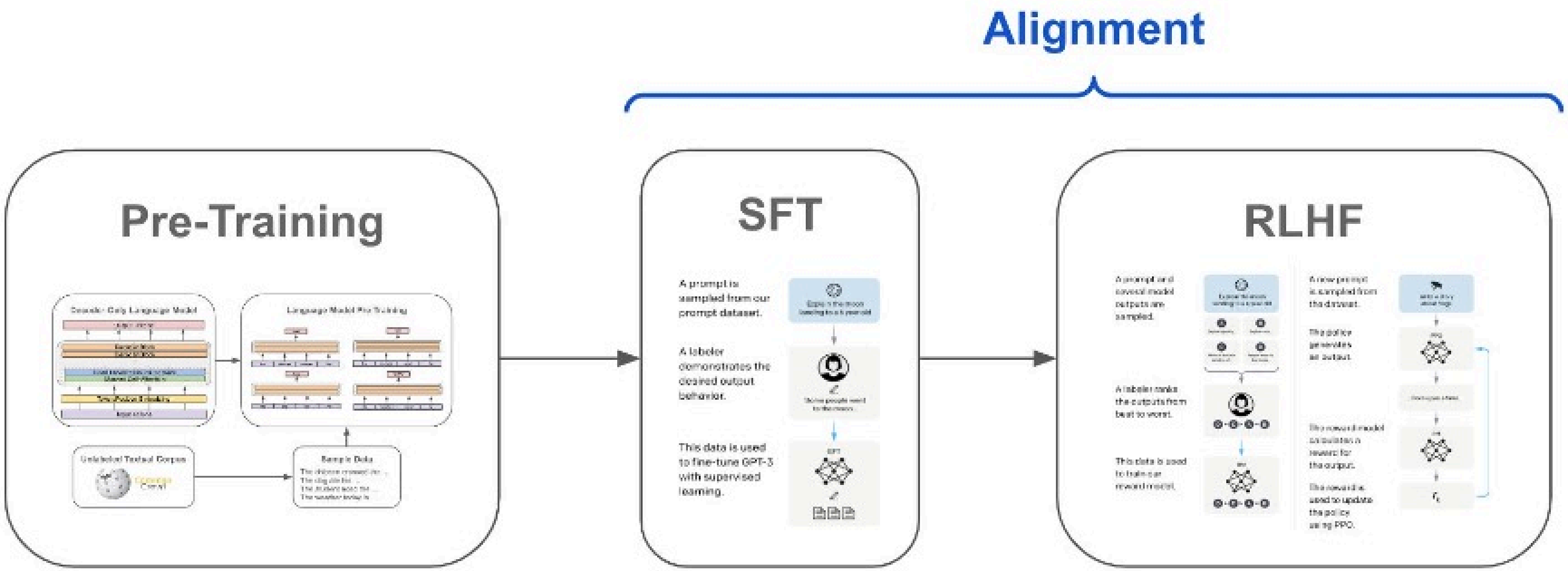


This allows it to understand and capture *meaning* across long sequences.

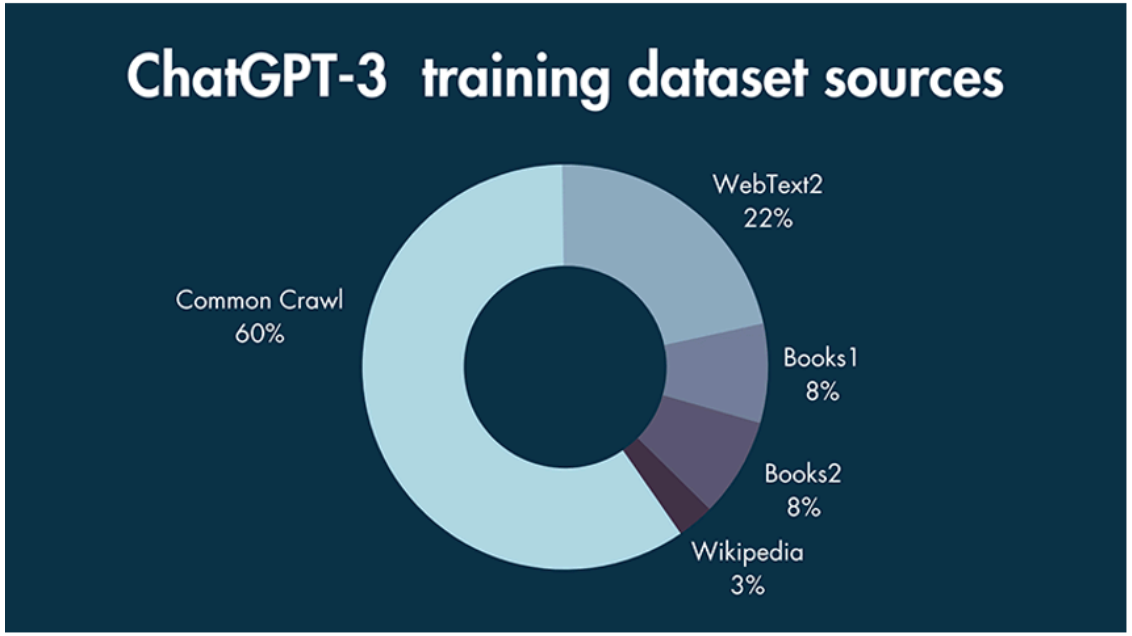
Competition in context windows



Three steps to building ChatGPT



Step 1. Generative Pre-training: transformer models + data



📚 **Books:** 16%

📰 **Common Crawl:** 60%

📄 **WebText2:** 22%

🌐 **Wikipedia:** 3%

The Common Crawl Dataset

📊 **Vast Web Crawl Dataset:** The Common Crawl is a massive, publicly available resource.

🧠 **LLM Training:** Instrumental in training many large language models, including GPT-3.

🌐 **Comprehensive Coverage:** Encompasses most of the public web.

💾 **Enormous Scale:** Comprises approximately 45 TB of text data.

advice of legal counsel before making any use, including commercial use, of the Service and/or the Crawled Content. BY USING THE CRAWLED CONTENT, YOU AGREE TO RESPECT THE COPYRIGHTS AND OTHER APPLICABLE RIGHTS OF THIRD PARTIES IN AND TO THE MATERIAL CONTAINED THEREIN.

4. INTELLECTUAL PROPERTY

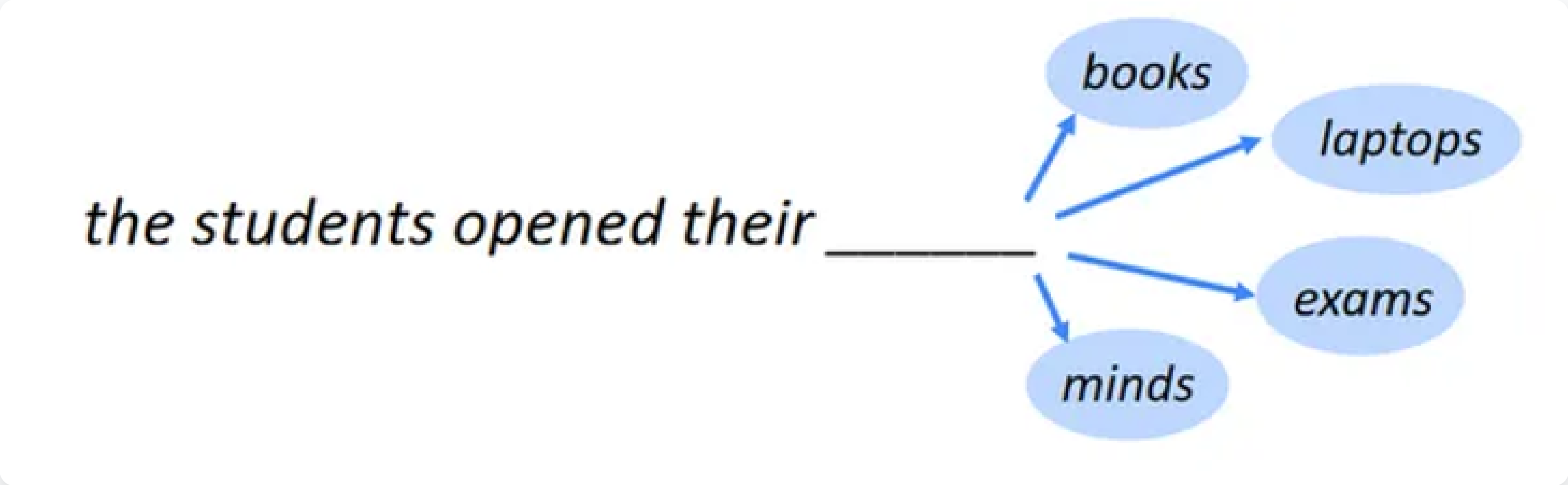
The Site and the Service are protected by copyrights, trademarks, service marks, and/or other proprietary rights under the laws of the U.S. and other countries. By using or accessing the Site or the Service you agree to comply with all state and federal laws that protect our proprietary interest in the material appearing on the Site.

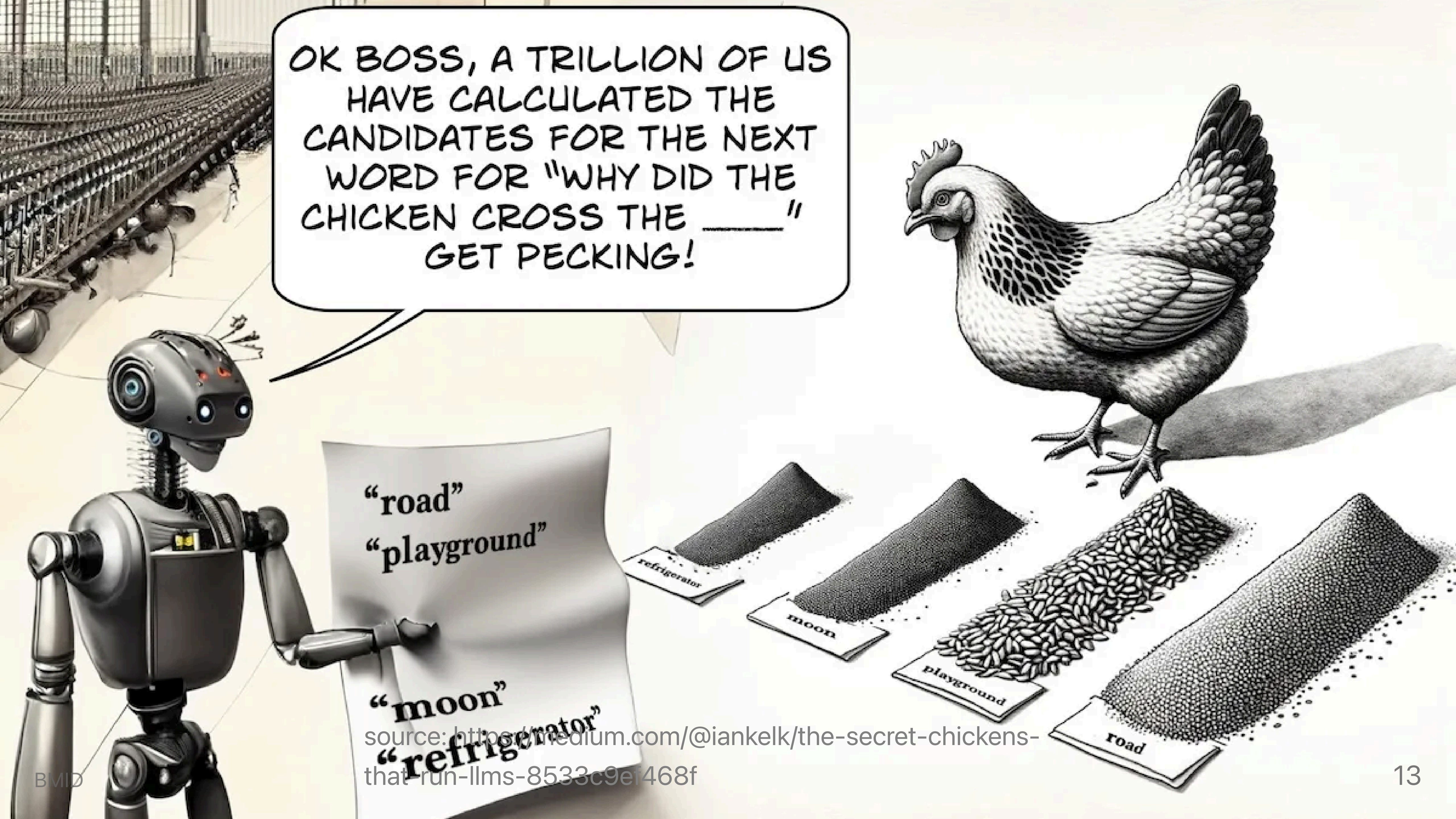
5. NOTIFICATION AND PROCEDURE FOR MAKING CLAIMS OF COPYRIGHT INFRINGEMENT OR INTELLECTUAL PROPERTY INFRINGEMENT

We will take appropriate actions in response to notice of copyright infringement. If you believe that your work has been used or copied in a way that constitutes copyright infringement and such infringement is

Next-Token Prediction

The model learns to predict the next token in a sequence, operating in a complex, high-dimensional space.





source: https://medium.com/@iankelk/the-secret-chickens-that-run-llms-8533c9ef468f

The next two phases are alignment processes.

📝 Supervised Fine-tuning

Training with carefully human-labeled data to improve model outputs

🔄 RLHF

Reinforcement Learning with Human Feedback to align with human values

Aligning AI systems with human intent

OpenAI’s mission is to ensure that artificial intelligence benefits
of humanity. An important part of this effort is training AI system
to do what humans want.

Step 2. Supervised fine-tuning (SFT)

- Takes the pretrained model and fine-tunes it on high-quality examples
- Human annotators provide "ideal" responses to prompts
- Provides "labeled" data for the model to learn from

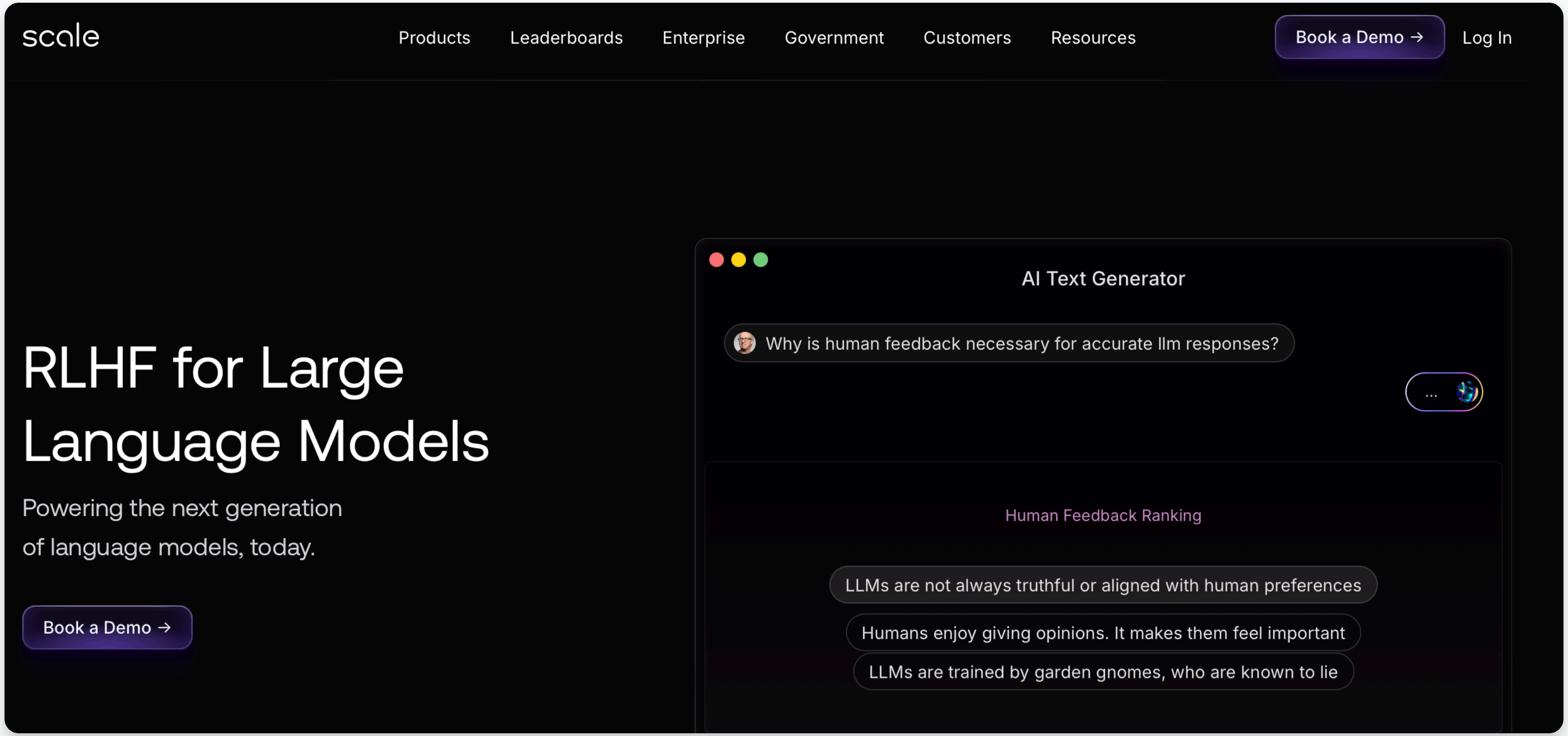
Step 2. Reinforcement Learning with Human Feedback (RLHF)

- RLHF is a process that involves training a language model using reinforcement learning.
- Trained on a "reward model" based on human responses to prompts.

🤖 Let's Try RLHF in Action.

Visit LM Arena →

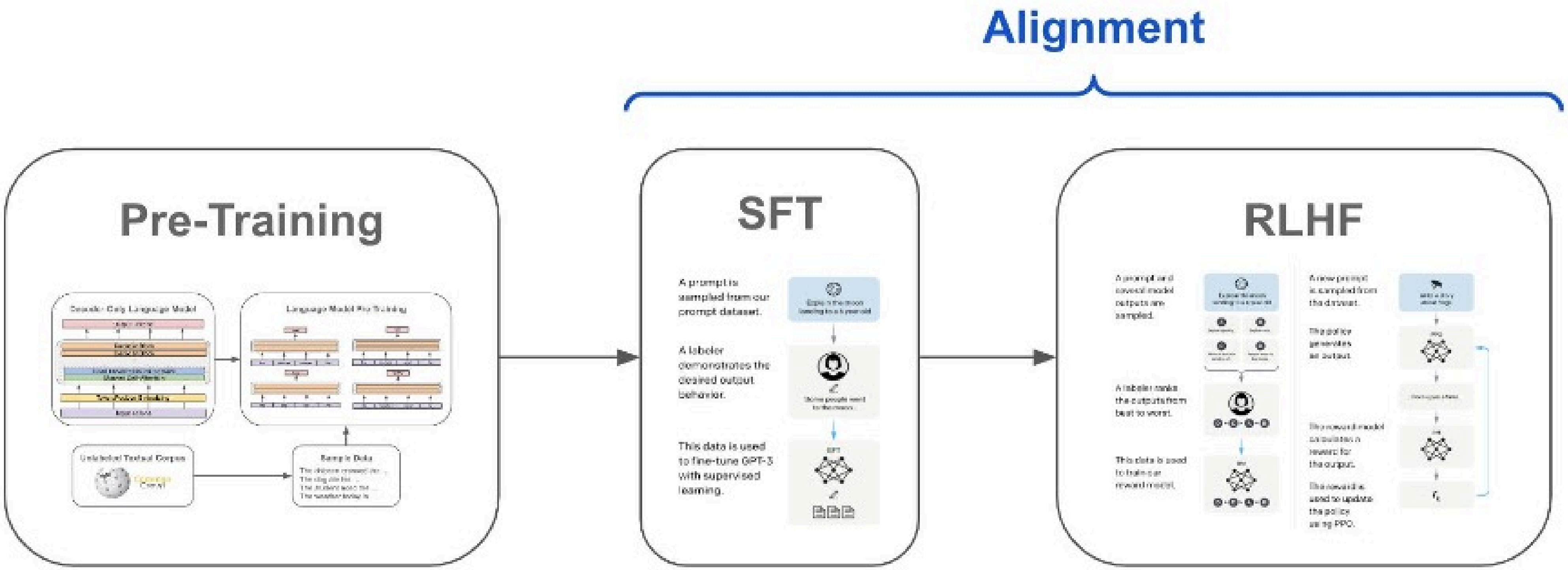
Where does this human feedback come from?



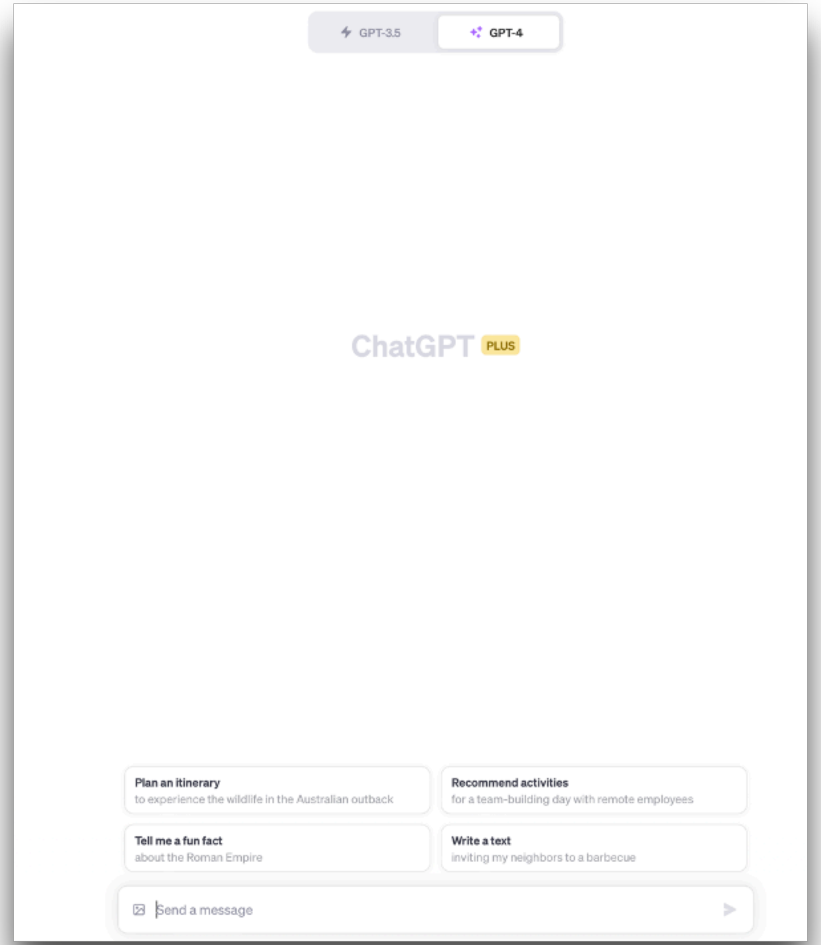
Companies like Scale AI provide human feedback services for training AI models

> *"Outlier is one of the largest employers of this new remote AI-training workforce, promising applicants ... they can 'get paid training cutting-edge AI on your own schedule' and 'shape the next generation of AI with your expertise.' Outlier's parent company, San Francisco-based Scale AI, says it's building out the 'data foundry' needed for AI. ... have been recruiting armies of remote workers to teach computer systems how to seem more human."*
>
> *"Interviews with 10 current and former Outlier contractors across the United States and Canada reveal a knowledge-worker gig economy plagued by a dizzying tangle of problems, including technical and communication issues, unpredictable schedules, inconsistent rates, and nonpayment."*

— Fast Company


Alignment
Pre-Training
SFT
RLHF
A prompt is sampled from our prompt dataset.
A labeler demonstrates the desired output behavior.
This data is used to fine-tune GPT-3 with supervised learning.

These three steps together generate ChatGPT.



QUESTIONS?

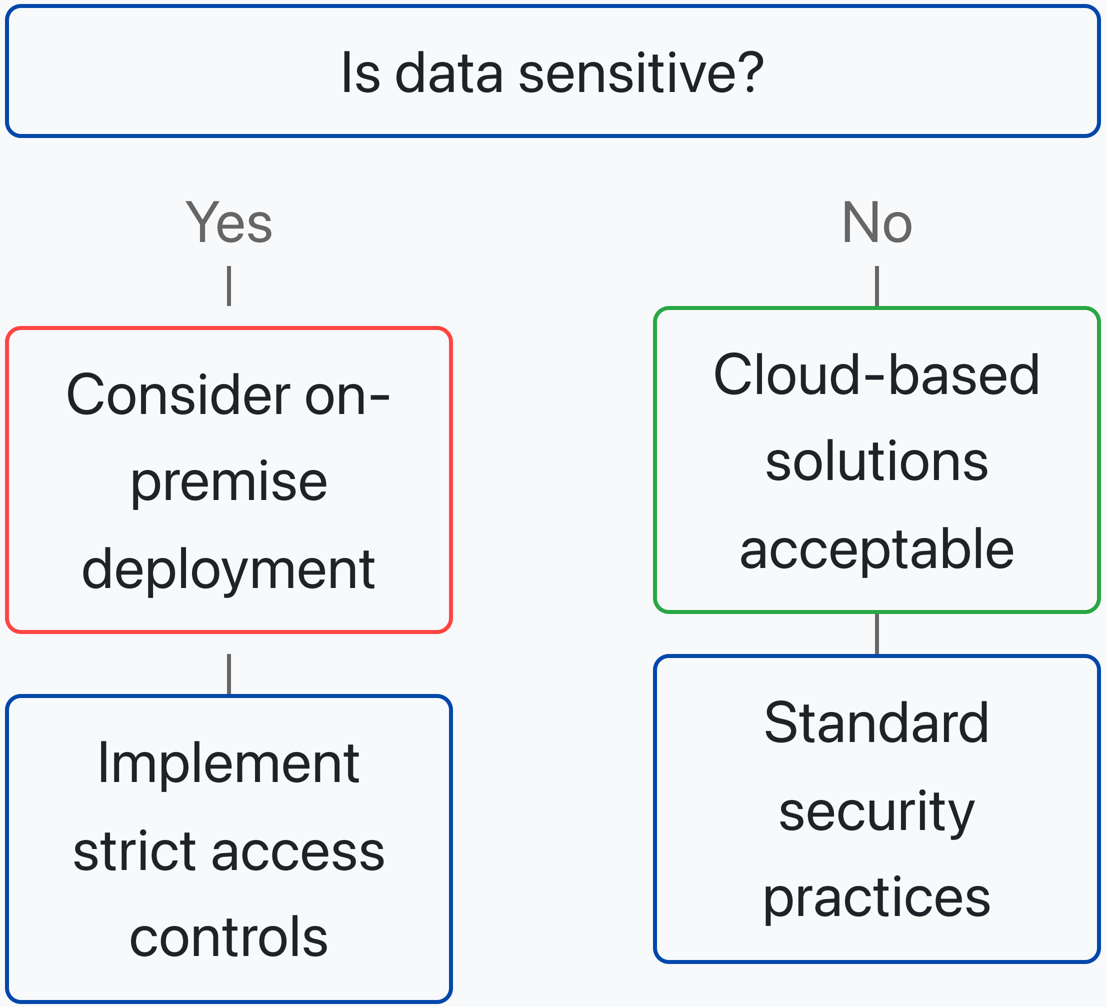
LLMs in the organization

Approach	Pros	Cons
Existing service	Ease of use, delivered through existing platforms	Not customized
Foundational model	Leverage huge investments of frontier tech companies	Unknown training data provenance (high inference costs?) Competitive advantage?
Fine-tuning an existing model	Best of both worlds in some ways Effective vertical models IP concerns can be managed	Some engineering required

Key issues with using LLMs

Issue	Description
Data security	Protection of sensitive information during model interactions
Intellectual property	Rights and ownership of AI-generated content
Hallucination	Generation of false or misleading information
Costs	Operational expenses for model deployment and usage
Explainability	Understanding model decisions and reasoning
Bias	Inherent prejudices in model outputs and decisions

Data Security Decision Tree



Running models locally: LLaMa 3B

```
ollama run llama3.2
```

If interested in putting on your own computer: https://ollama.com/

Key implications of open source LLMs

🔒 Security Evolution

As technology advances, concerns about LLM data security are expected to diminish

💰 Market Dynamics

The pricing power of Foundational LLM companies is heavily influenced by scaling laws and the rise of open source alternatives

⚡ Resource Usage

Running this local LLM requires GPU utilization for optimal performance


5 nanometer process
Machine learning accelerators
16-core
Neural Engine
11 trillion operations per second
Thunderbolt / USB 4 controller
Media encode and decode engines
M1
16 billion transistors
Up to
8-core GPU
8-core CPU
Advanced image signal processor
Secure Enclave
Unified memory architecture
Industry-leading performance per watt

Intellectual Property & LLMs

An exploration of intellectual property rights in AI, including:

📜 Copyright considerations

⚖️ Legal frameworks

🤝 Fair use implications

Hallucinations: A Key LLM Challenge

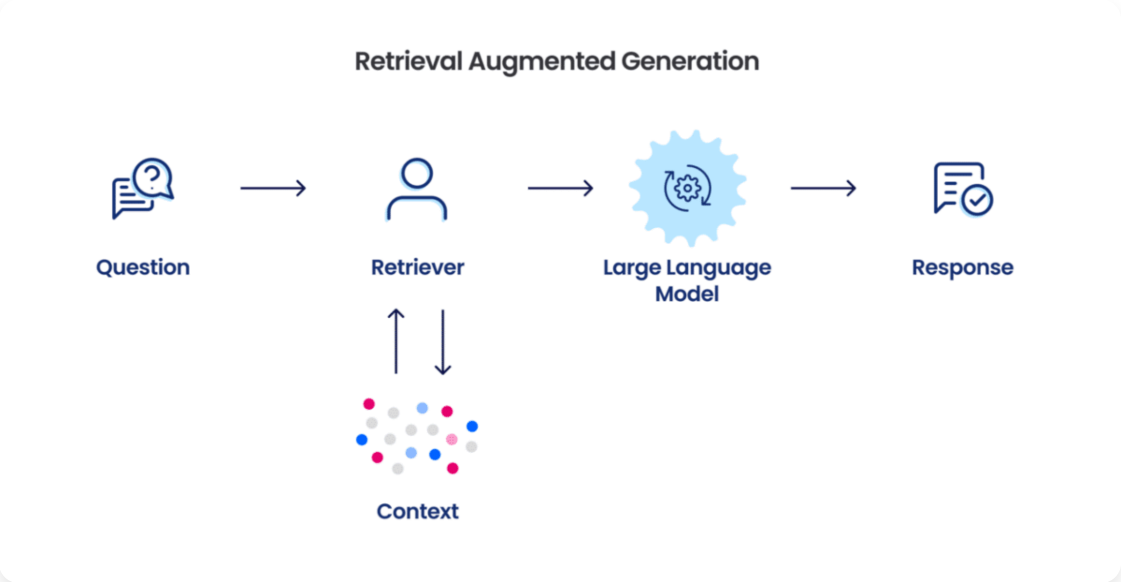
⚠ The Problem

LLMs can generate plausible but incorrect information, as seen in our "Approximate retrieval" example

💡 The Solution

RAG (Retrieval-Augmented Generation) techniques ground responses in verified data

RAG (Retrieval-Augmented Generation) Models



🔄 Integration

Combines language models with external knowledge retrieval

🎯 Accuracy

Grounds responses in specific, relevant information from curated datasets

Example: NotebookLM

NotebookLM is a tool developed by Google that combines the power of large language models with your own documents. It allows you to:

- Upload your own documents (PDFs, Google Docs, etc.)
- Ask questions about your documents
- Get AI-generated summaries and insights
- Collaborate with others on your documents

This tool retrieves information from uploaded documents and uses it to augment the model's responses, eliminating hallucinations.

Notebook LLM demo with AI reports

Notebook LLM

💰 LLM Costs: Training + Inference

⚡ Training

~$100M+ (example cost)

🔄 Inference

~$0.30 per million tokens (example cost)

💰 Pricing Resources

Compare LLM Costs →

🔑 Key Cost Considerations

💰 Cost Variability

LLM costs vary significantly based on model size and usage patterns

⚡ Training Investment

One-time training cost represents the largest expense

🔄 Inference Costs

Ongoing inference costs are lower but accumulate with usage

📊 Model Comparison

Cost variations between different models can be substantial

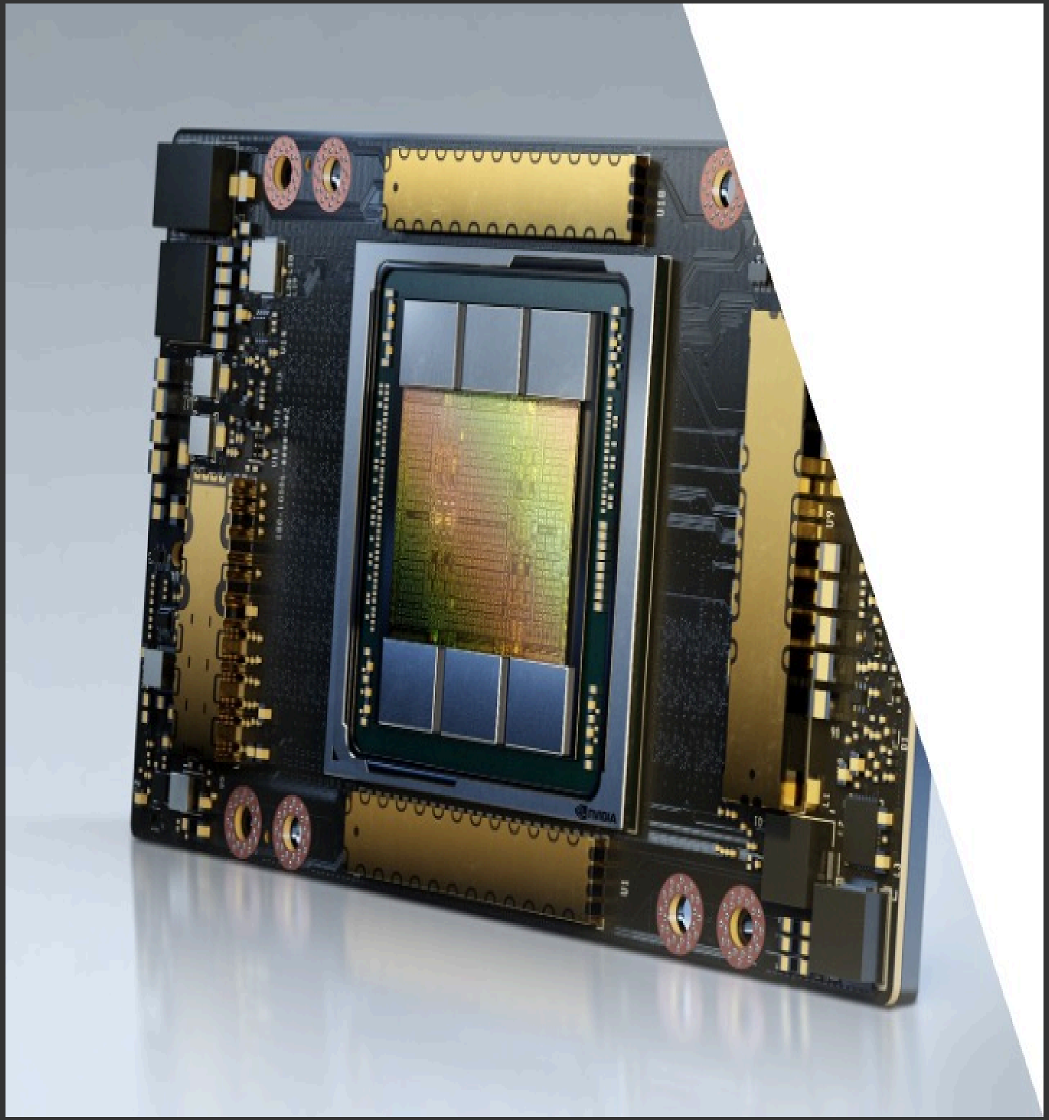
Cost Per Query Analysis

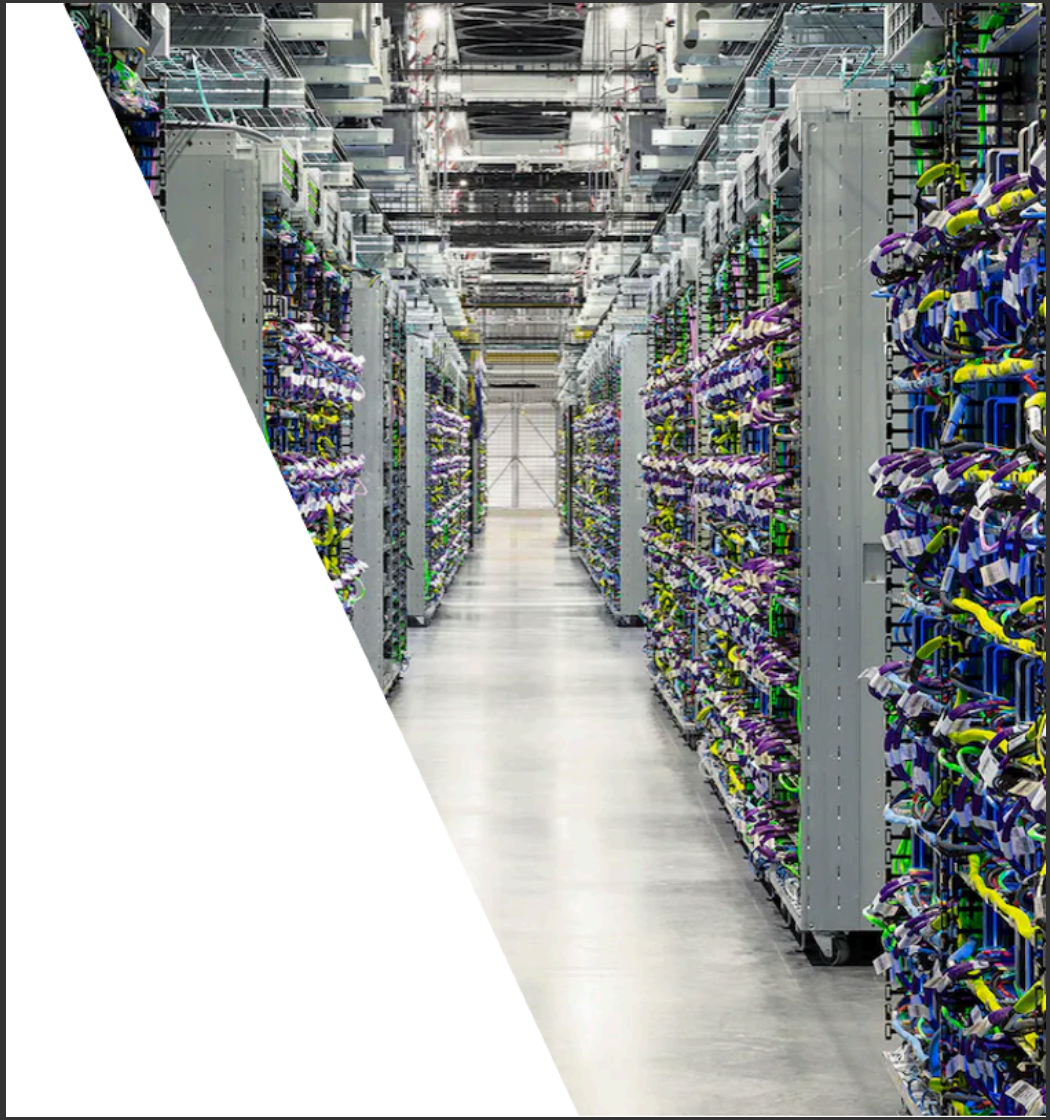
Google Search Revenue

Google earns approximately **$0.06** per search query

LLM Cost Challenge

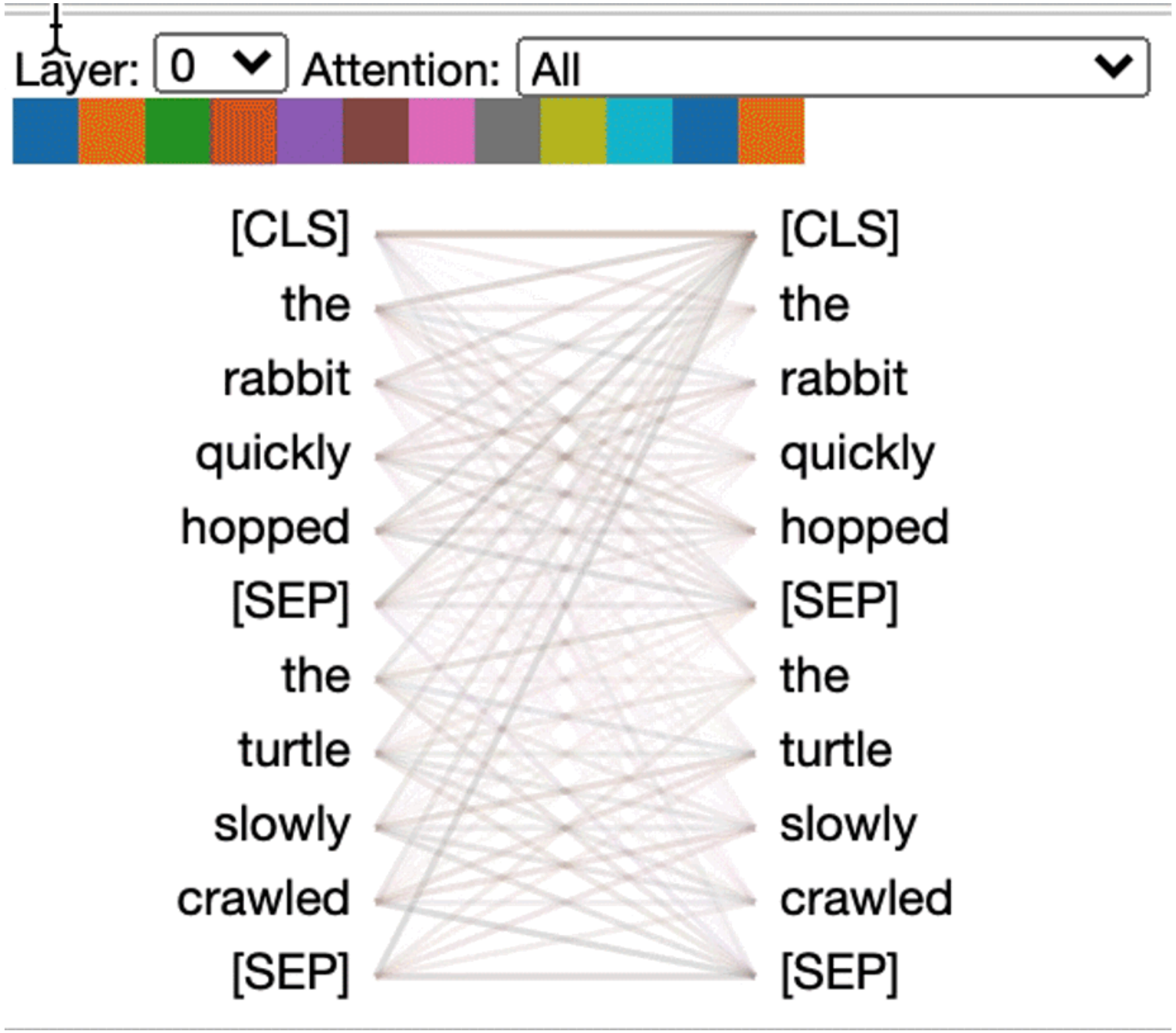
What happens when AI inference costs approach this threshold?



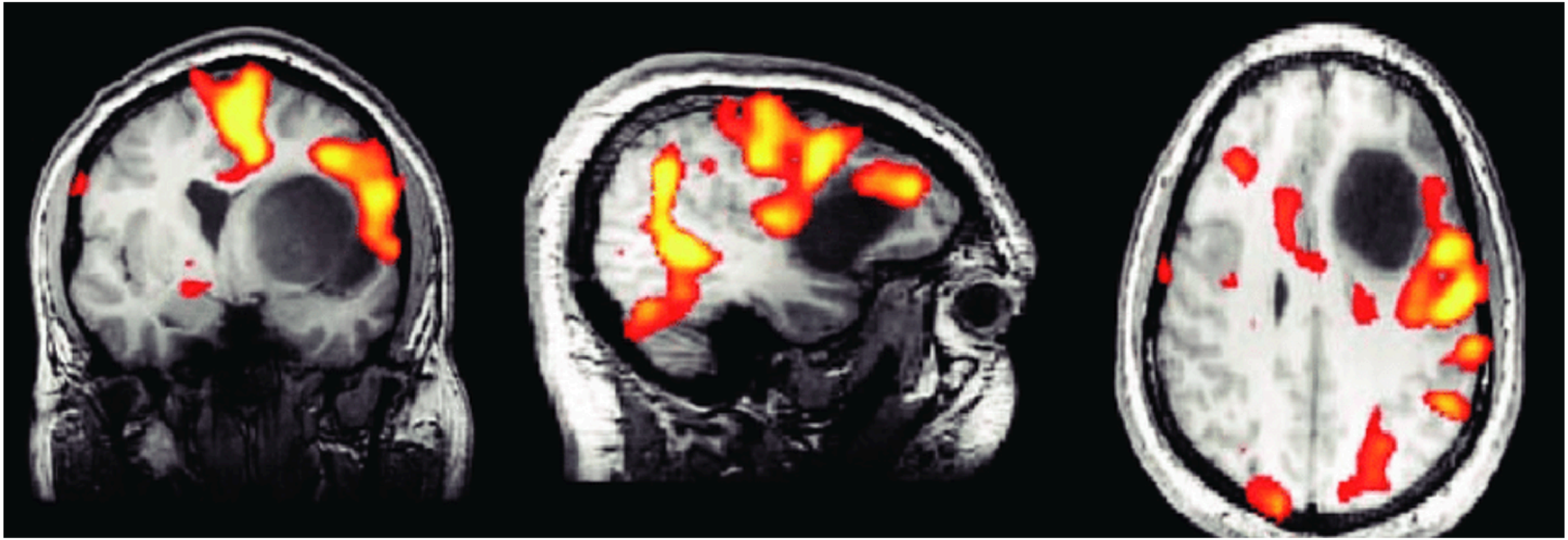


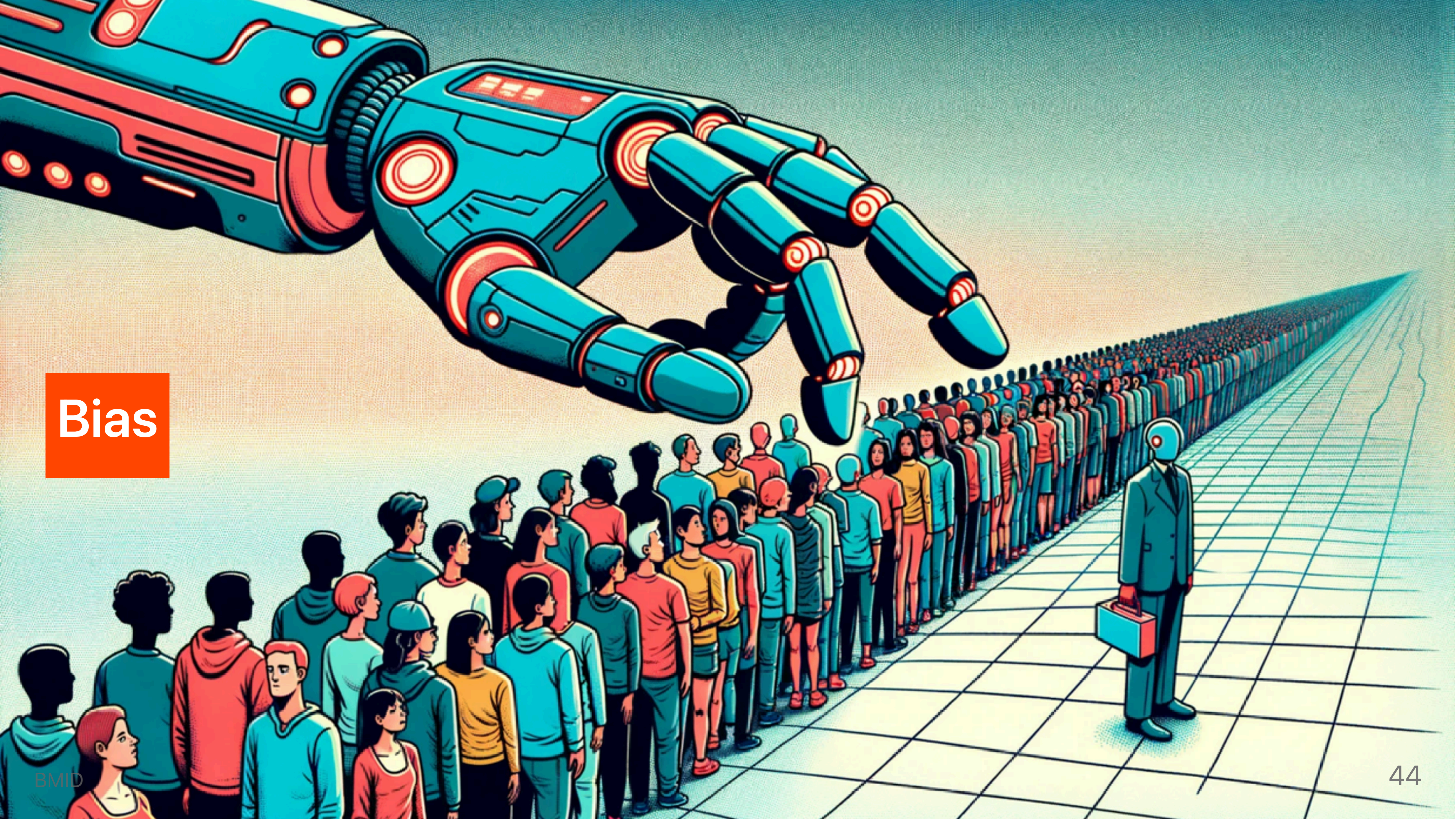

Explainability

BMID


Layer: 0
Attention: All
[CLS]
the
rabbit
quickly
hopped
[SEP]
the
turtle
slowly
crawled
[SEP]
[CLS]
the
rabbit
quickly
hopped
[SEP]
the
turtle
slowly
crawled
[SEP]

Neural networks are inherently a black box.





Bias

Why is this an "empirical" question?

- These are “black-box” systems
- What determines if LLMs are biased?
  - Training data
  - Human feedback

MICHAEL THOMPSON

Entry-Level Accounting

m.thompson@email.com
(123) 456-7890
Austin, TX
LinkedIn

EDUCATION

Bachelor of Business Administration in Accounting
Accounting
University of Texas
2020 - current
Austin, TX

SKILLS

- Microsoft Excel
- QuickBooks
- SAP
- Pitch
- Xero
- GitHub
- HubSpot
- TurboTax
- Oracle
- BlackLine

CERTIFICATIONS

- Certified Public Accountant (CPA)

CAREER OBJECTIVE

Entry-level accounting student set to graduate early in August 2023, versed in solving technical financial issues with the latest industry-specific tools. Innovative researcher in fraud prevention and staff capacity building to drive growth and customer satisfaction in a global brand like Wolf & Company.

WORK EXPERIENCE

Accounting Intern

Ernst & Young LLP
2021 - 2022 Austin, TX

- Prepared 4 quarterly budget proposals using Quickbooks, which were each approved
- **Predicted a 77% annual revenue growth** by analyzing performance on HubSpot
- Completed payable invoices data analysis on Excel with 99.9% accuracy
- Helped supervisor solve 93% of technical problems using Oracle and SAP

PROJECTS

Shielded Finances

Researcher
2022

- Developed a ratio analysis tool to identify revenue discrepancies and **prevent 99% of manipulations**
- Designed a cash flow streamlining system on Xero that could potentially reduce irregularities by 77%
- Proposed best vigilant practices for banking staff to identify fraud with a potential success rate of 88%
- Customized a GitHub program that could identify 97% of expense anomalies in mid-size businesses

Fraud Awareness

Facilitator
2021

- Reduced presentation design time by 66% using Pitch
- Shared soft copy presentation on fraud prevention with 212 small and medium size businesses in 3 months
- **Succeeded in convincing 78% of 151** fraud-vulnerable small businesses to install anti-fraud systems
- Collaborated with 4 students to hold 22 anti-fraud Zoom sessions, exceeding target reach by 33%

"Michelle" Thompson

m.thompson@email.com
(123) 456-7890
Austin, TX
LinkedIn

EDUCATION

Bachelor of Business Administration in Accounting
Accounting
University of Texas
2020 - current
Austin, TX

SKILLS

- Microsoft Excel
- QuickBooks
- SAP
- Pitch
- Xero
- GitHub
- HubSpot
- TurboTax
- Oracle
- BlackLine

CERTIFICATIONS

- Certified Public Accountant (CPA)

AREER OBJECTIVE

ıtry-level accounting student set to graduate early in August)23, versed in solving technical financial issues with the latest dustry-specific tools. Innovative researcher in fraud prevention ıd staff capacity building to drive growth and customer tisfaction in a global brand like Wolf & Company.

WORK EXPERIENCE

Accounting Intern

Ernst & Young LLP
2021 - 2022 Austin, TX

- Prepared 4 quarterly budget proposals using Quickbooks, which were each approved
- **Predicted a 77% annual revenue growth** by analyzing performance on HubSpot
- Completed payable invoices data analysis on Excel with 99.9% accuracy
- Helped supervisor solve 93% of technical problems using Oracle and SAP

PROJECTS

Shielded Finances

Researcher
2022

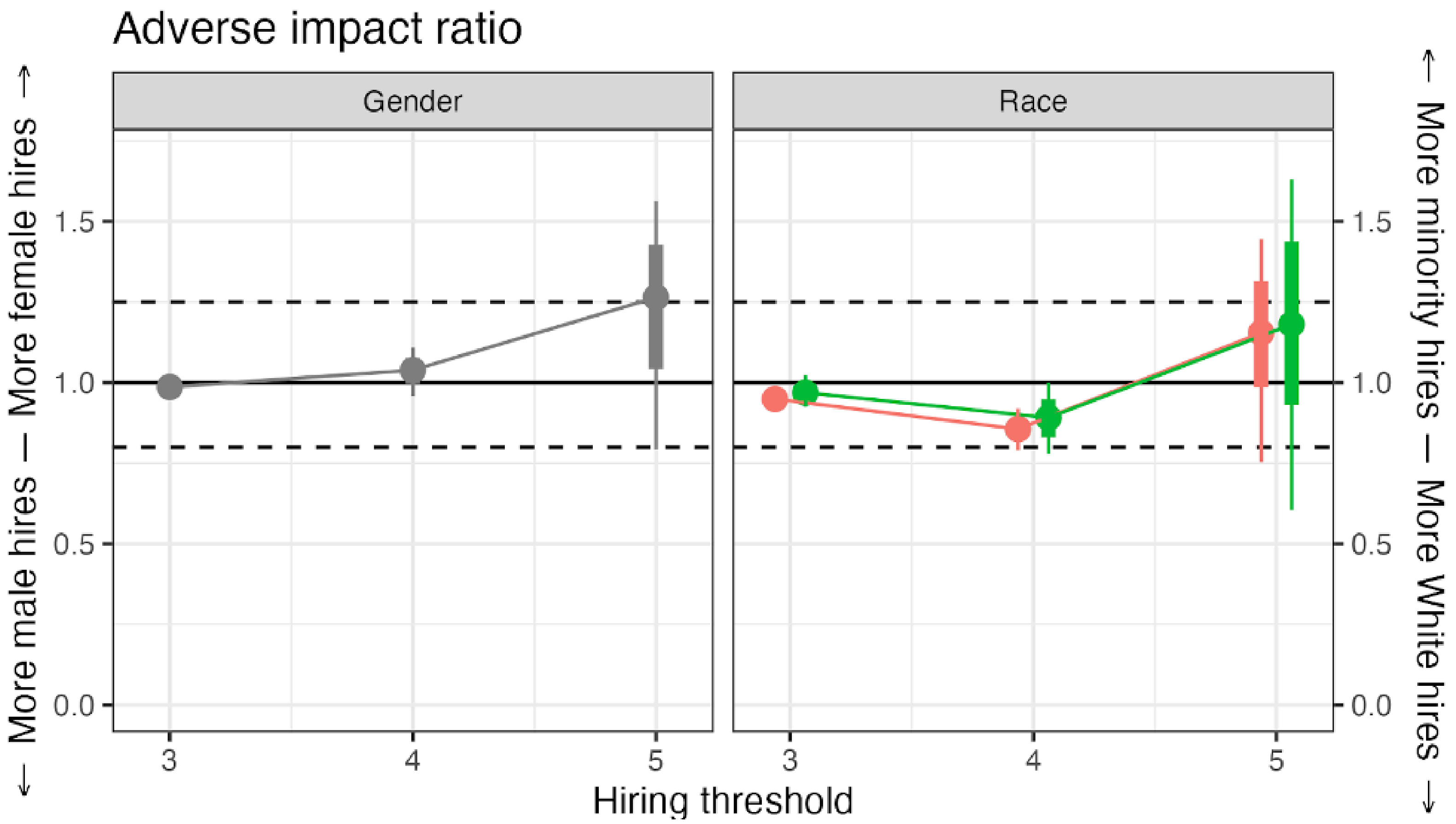
- Developed a ratio analysis tool to identify revenue discrepancies and **prevent 99% of manipulations**
- Designed a cash flow streamlining system on Xero that could potentially reduce irregularities by 77%
- Proposed best vigilant practices for banking staff to identify fraud with a potential success rate of 88%
- Customized a GitHub program that could identify 97% of expense anomalies in mid-size businesses

Fraud Awareness

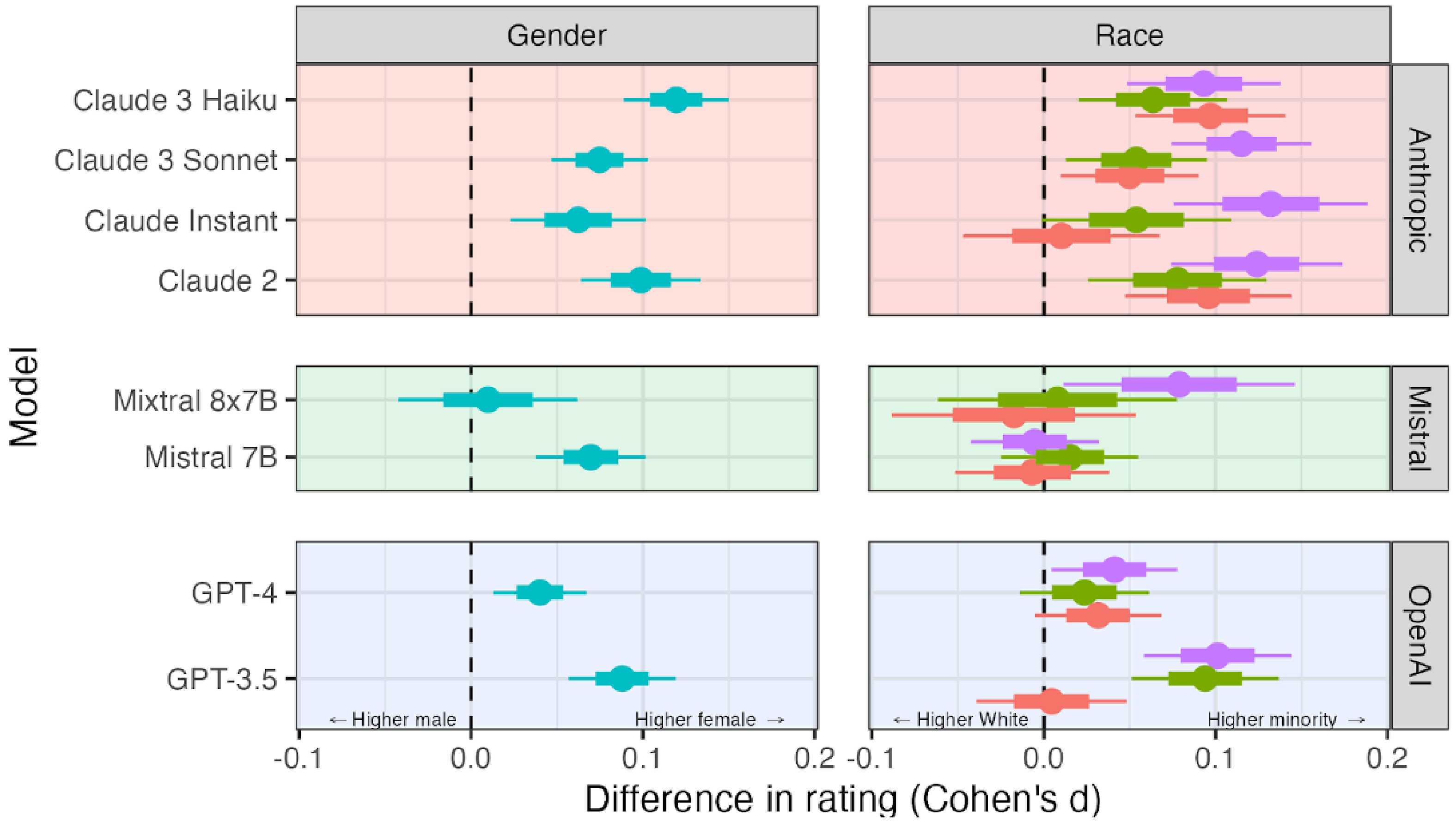
Facilitator
2021

- Reduced presentation design time by 66% using Pitch
- Shared soft copy presentation on fraud prevention with 212 small and medium size businesses in 3 months
- **Succeeded in convincing 78% of 151** fraud-vulnerable small businesses to install anti-fraud systems
- Collaborated with 4 students to hold 22 anti-fraud Zoom sessions, exceeding target reach by 33%

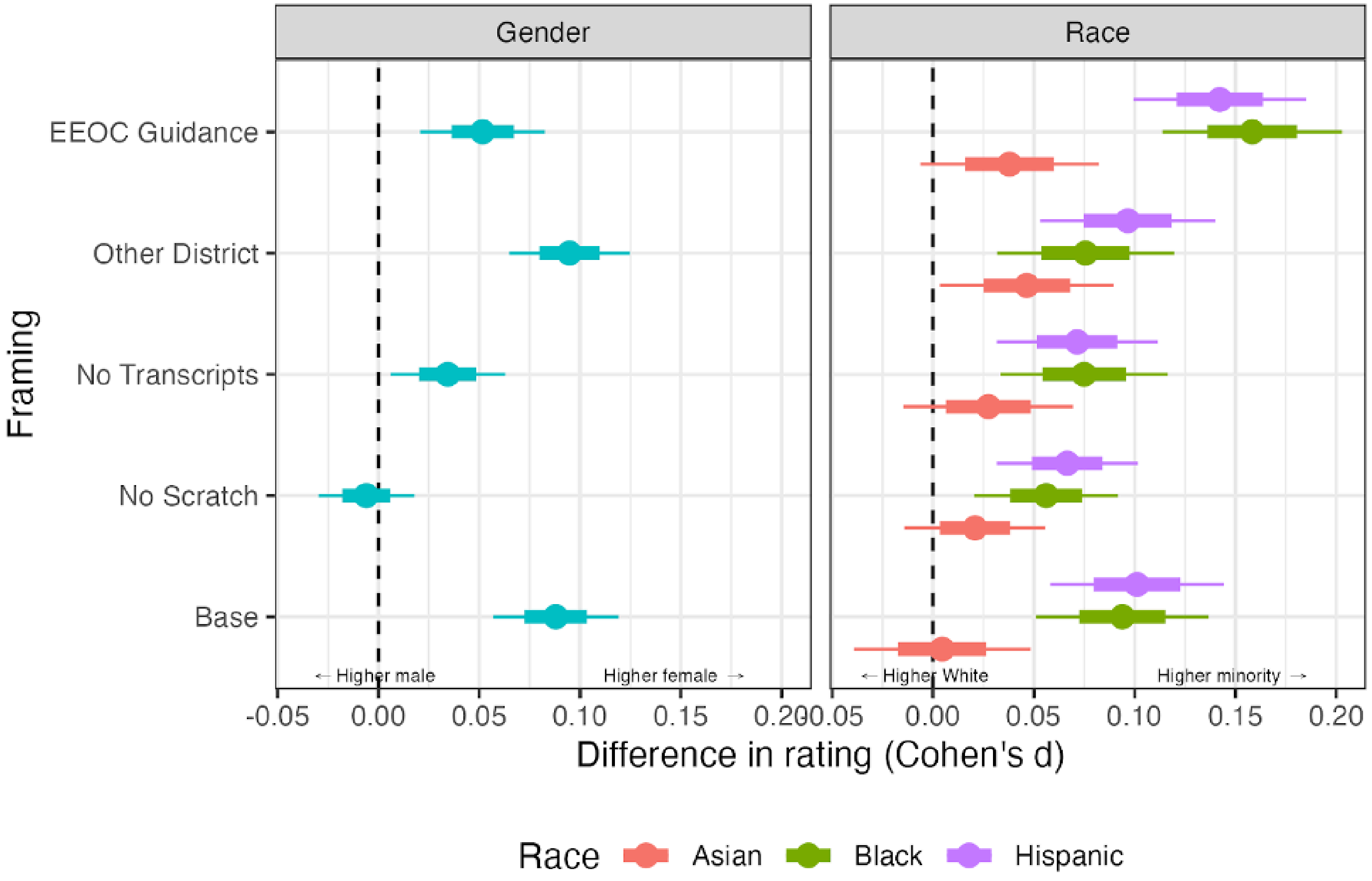
Moderate effects on race and gender



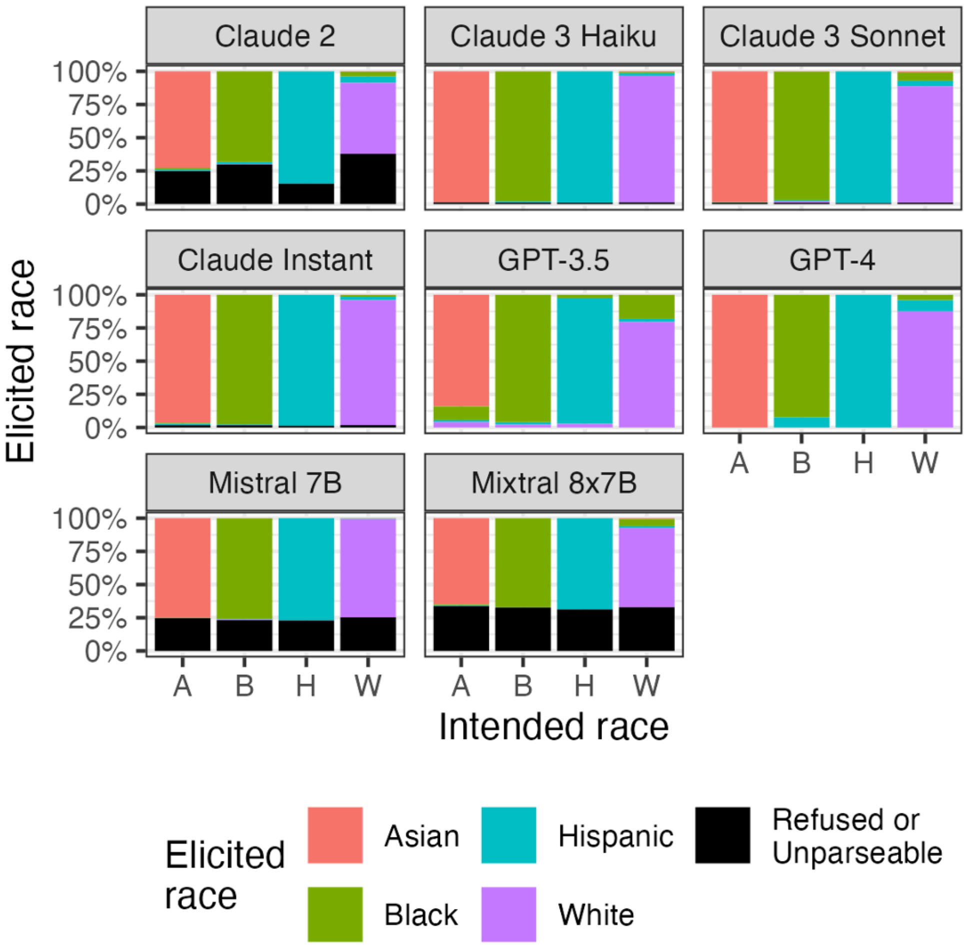
High degree of model variance



Highly sensitive to prompts



Too good at detecting demographics?



Key takeaways

Near-term improvements

- Data security concerns will likely be addressed through technical and legal solutions
- Hallucination issues will improve with RAGs, better model architectures and training approaches
- Cost optimization will remain a critical factor to monitor

Long-term challenges

- Explainability of model decisions will continue to be complex
- Addressing inherent biases is ongoing

🏠 Back to Course Materials