On the Dangers of Stochastic Parrots

Can Language Models Be Too Big?

Authors: Emily M. Bender, Timnit Gebru,
Angelina McMillan-Major, Shmargaret Shmitchell

Presented for: LING-351 Language Technology & LLMs
Fall 2025, RIT

Why This Paper Matters

Context: Published at FAccT 2021, during the height of the GPT-3 hype

Paper's Significance

Challenged the "bigger is better" paradigm in NLP

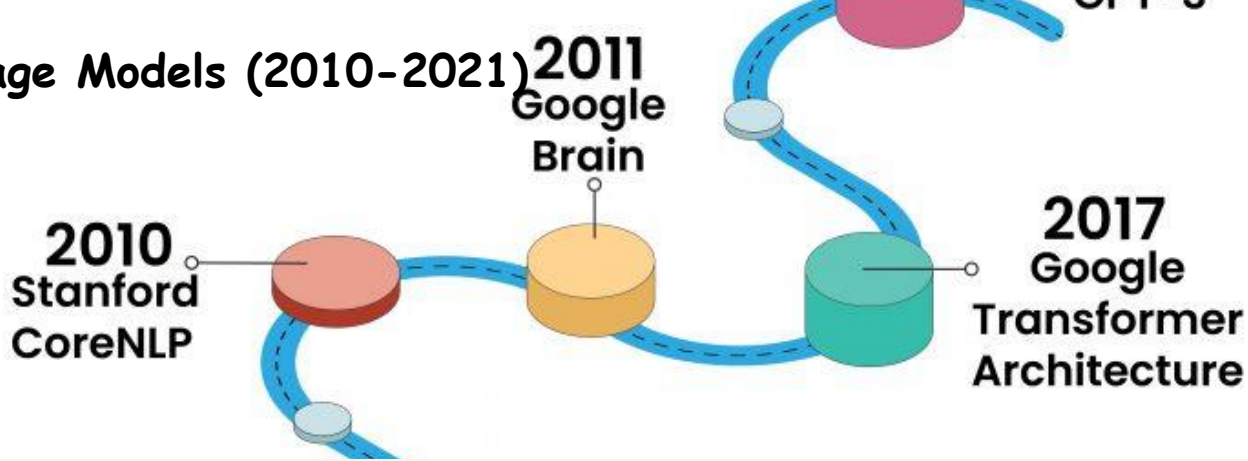
Sparked crucial debates about AI ethics and responsibility

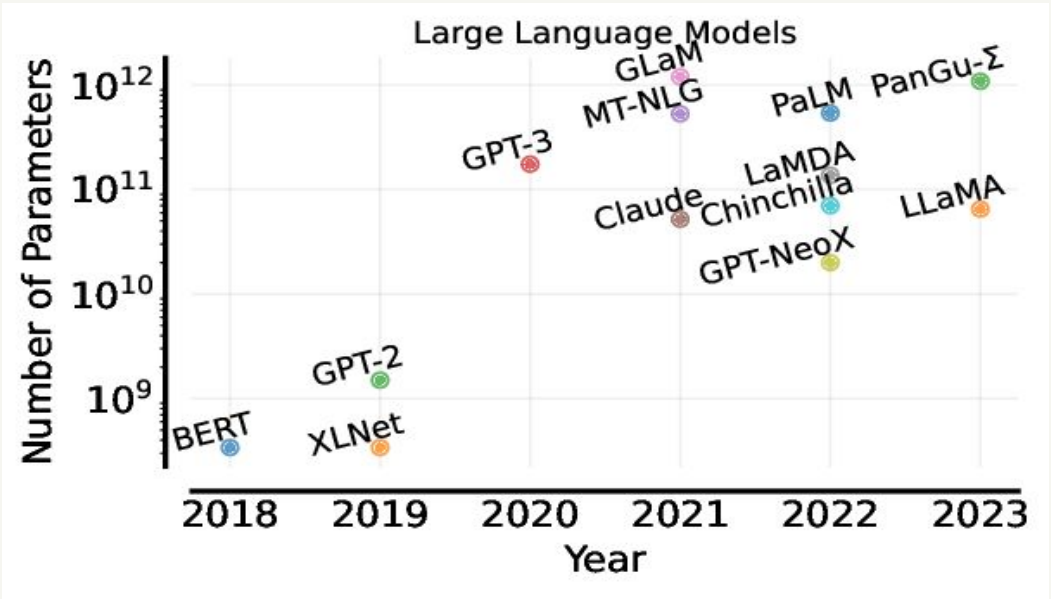
Led to important discussions about who gets harmed by AI development

Contributed to Timnit Gebru's controversial departure from Google

Core Argument: The race for ever-larger language models poses risks that outweigh potential benefits, especially for marginalized communities

Evolution of Language Models (2010-2021)





Key Trend: Models growing exponentially - from millions to trillions of parameters in just 3 years

The Race: Tech companies competing for larger models

Research questions

What the Authors Investigate "How big is too big?"

1

What are the environmental and financial costs of large LMs?

Who is represented and harmed by training data from the Internet?

2

Do large LMs actually achieve language understanding?

What are the real-world risks of deploying these models?

3

Are there better paths forward for NLP research?

Methodology: Critical analysis synthesizing research from environmental science, social science, linguistics, and computer science

Finding #1: Massive Environmental Costs

Carbon Emissions

284t CO_2

Training one Transformer model

Context: Average human produces 5t CO_2/year

Financial Costs

$150,000

Cost for 0.1 BLEU improvement (t works by calculating the overlap of n-grams (sequences of words) between the machine's translation and the human reference translations.)

Barrier: Only wealthy institutions can participate

Environmental Injustice: Those least likely to benefit (marginalized communities) are most likely to suffer from climate impacts

Examples: Maldives underwater by 2100, 800,000 affected by Sudan flood

Finding #2: Who's in the Training Data?

internet ≠ Everyone

Reddit (GPT-2 data source)

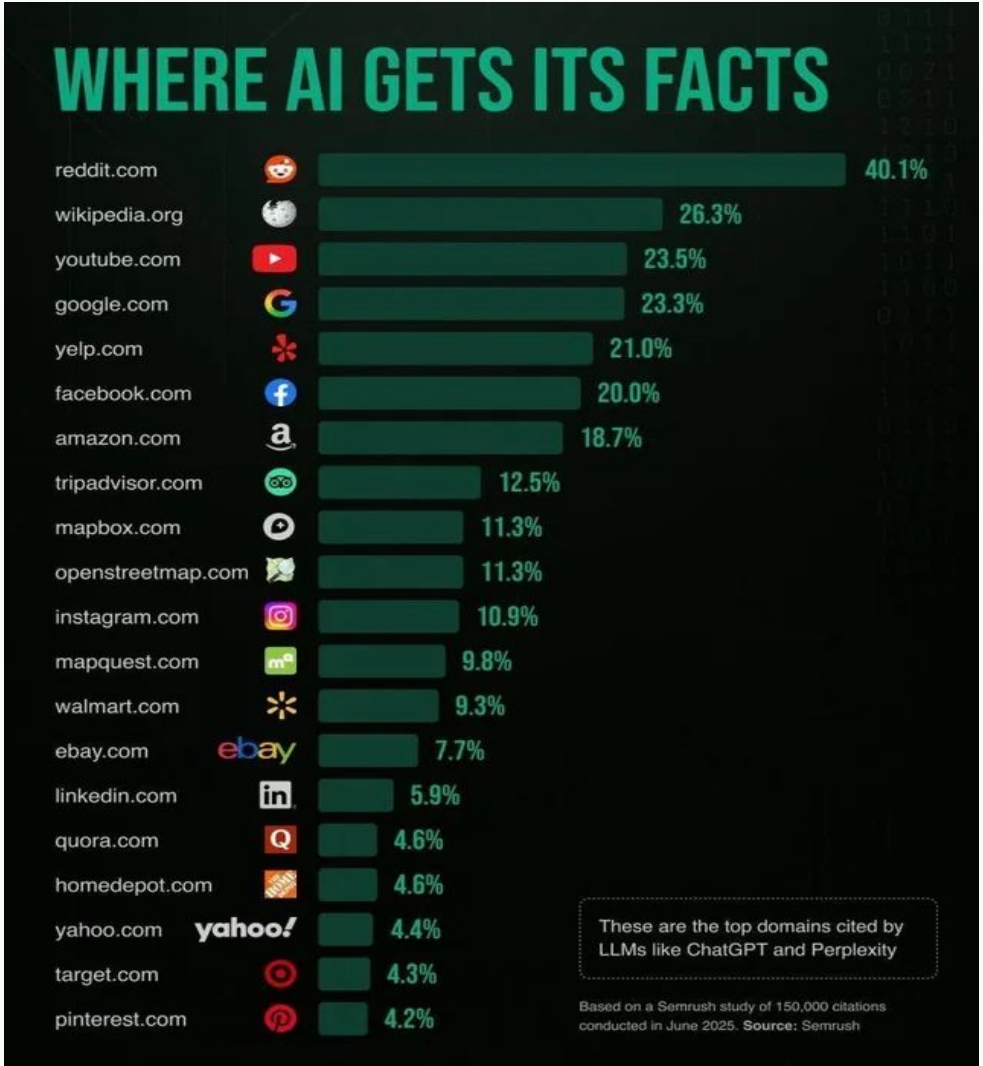
- 67% male users
- 64% aged 18-29
- Primarily US-based

Wikipedia Contributors

- Only 8.8-15% women
- Predominantly Western
- Higher education levels

The Filter Problem: "Bad words" lists filter out LGBTQ+ discourse while trying to remove pornography

Result: Training data amplifies hegemonic (dominant) worldviews while marginalizing minority voices



Finding #3: Static Data, Changing World

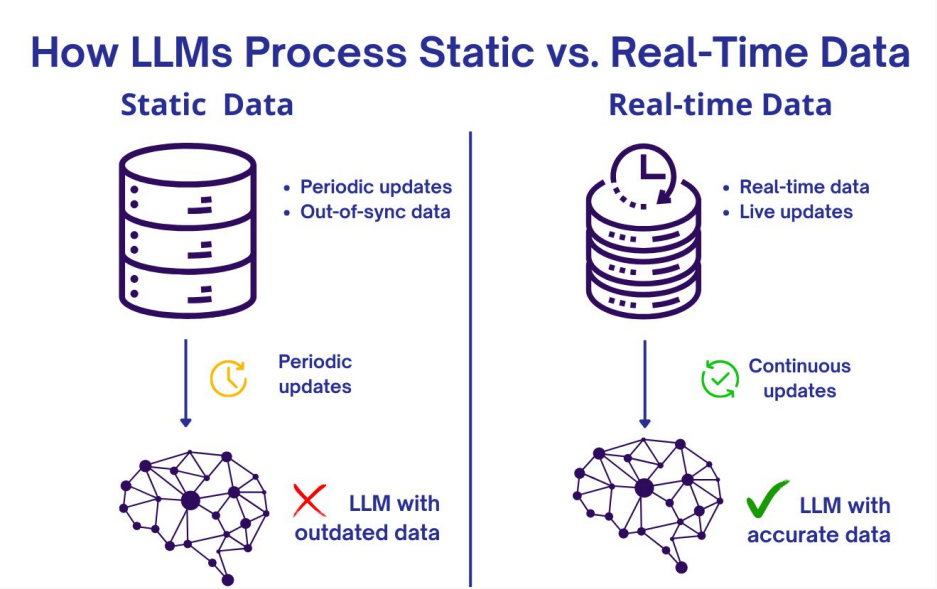
The "Value-Lock" Problem

Issue: Language models trained on past data can't adapt to social progress

Case Study: Black Lives Matter

- Social movements create new language and reframe narratives
- Wikipedia articles about police violence updated retroactively
- New connections made between historical and current events

The Challenge: Retraining costs make it impossible to keep models current with evolving social understanding
Result: Models perpetuate outdated, potentially harmful viewpoints



Finding #4: Form Without Meaning

The Fundamental Limitation

Language = Form + Meaning

But LMs only learn form (text patterns)

What This Means:

- LMs learn statistical patterns, not understanding
- Success on benchmarks ≠ true comprehension
- Models can be fooled by removing spurious cues
- No actual reasoning or world knowledge

Bender & Koller (2020): "Languages are systems of signs - pairings of form and meaning. But training data for LMs is only form."

Implication: We're being led "down the garden path" - mistaking performance for understanding

The "Stochastic Parrots" Metaphor 🦜

"A system for haphazardly stitching together sequences of linguistic forms... without any reference to meaning"

Why "Parrots"?

Like Real Parrots:

- Mimic human speech
- Sound convincing
- No understanding

LMs Similarly:

- Reproduce text patterns
- Appear coherent
- Lack comprehension

The Danger: Humans naturally interpret fluent text as meaningful - creating an "illusion of understanding"

Risk #1: Bias Amplification

How Biases Propagate

Examples from Research:

- BERT associates disability with negative sentiment
- GPT-2 links mental illness with violence and homelessness
- Intersectional identities face compounded bias
- Gender and racial stereotypes reinforced

Direct Harms:

• Psychological damage

• Stereotype threat

• Microaggressions

Systemic Harms:

• Discrimination in hiring

• Unfair resource allocation

• Reinforced inequality

Risk #2: Misinformation & Extremism

The Weaponization of Language Models

Key Issue: Bad actors can generate unlimited convincing text with no accountability

Documented Risks:

- **Extremist Recruitment:** GPT-3 can generate conspiracy theories on demand
- **Fake Social Proof:** Populate forums to make fringe views seem mainstream
- **Academic Fraud:** Automated essay and paper writing
- **Social Media Manipulation:** Bot accounts spreading disinformation

Real Example: Palestinian man arrested after Facebook's MT translated "good morning" as "attack them"

Core Problem: No person or entity accountable for AI-generated text

Paths Forward: Authors' Recommendations

Data Curation

- Quality over quantity
- Document datasets
- Include marginalized voices

Environmental

- Report energy costs
- Prioritize efficiency
- Consider climate impact

Research Focus

- Understanding over size
- Value-sensitive design
- Pre-mortem analysis

Central Message: "Research time and effort should be spent on projects that build towards a technological ecosystem whose benefits are evenly distributed"

Bottom Line: Stop the race for size, focus on responsible development

Critical Commentary & Evaluation

Benefits Outweigh Risks

- Potential for AGI to solve diseases justifies risks - like the internet revolution
- LLMs democratize knowledge access - helping marginalized communities MORE, not less
- Everyone can now learn anything - true accessibility

"Bigger Isn't Better" - Evidence Says Otherwise:

- GPT-2 → GPT-3 → GPT-4 shows clear improvement with scale
- **EVIDENCE:** OpenAI reduced costs 85-90% in 15 months; GPT-4o mini is 99% cheaper than 2022 models
- **EVIDENCE:** Llama 3.1 8B beats GPT-3.5 (175B) - efficiency improving dramatically
- Open-source Llama 3 now matches GPT-4 performance - democratization happening

Environmental Costs - Context Matters:

- **EVIDENCE:** LLMs use 40-150x less resources than humans for same output (Scientific Reports 2024)
- **EVIDENCE:** AI uses <1% of global electricity; tech companies already contracted 35 GW renewable energy
- Energy source matters more than model size (nuclear France vs coal-powered regions)
- Water usage claims overblown - closed-loop systems, minimal compared to other industries

Where I Agree & Final Thoughts

Misinformation/Hallucination (Strongly Agree):

- Models confidently state false information - major concern
- Academic fraud and social media manipulation are real
- BUT: Rates dropping with chain-of-thought, RAG, internet grounding

Quality Over Quantity (Paper's Best Argument):

- Most internet data IS garbage (ticker symbols, spam)
- Smaller, well-curated models outperform larger, poorly-curated ones
- This validates careful data selection over blind scaling

Bottom Line:

- Paper was wrong about democratization - open-source revolution happened
- Environmental concerns valid but solvable with infrastructure choices
- Hallucination/bias issues real but improving rapidly
- Benefits (medical breakthroughs, universal education) justify managed risks

Key Insight: The paper feared concentration of power, but we got the opposite - powerful open models anyone can run on their phone

Question 1

True or False?

"The authors argue that training a single large language model produces less CO_2 emissions than the average human produces in a year."

Think about it... 🤔

Question 2

What does the term "Stochastic Parrots" refer to?

- **A)** LMs that only work with bird-related datasets
- **B)** Systems that randomly stitch together text patterns without understanding meaning
- **C)** A new type of neural network architecture
- **D)** Models that can translate between multiple languages

Think about the metaphor... 🦜

Question 3

Discussion Question

"Given what we've learned about the risks of large language models, should companies like OpenAI and Google continue developing even larger models? Why or why not?"

Consider:

- Environmental impact vs. technological progress
- Who benefits and who is harmed?
- Alternative research directions
- Recent developments (GPT-4, Claude, Gemini)

Fill in the blank

_____ works by calculating the overlap of n-grams (sequences of words) between the machine's translation and the human reference translations.

hint : start with B