Ethics of Generative AI Systems

Jeff Brill, Max Liu, Melina O'Dell

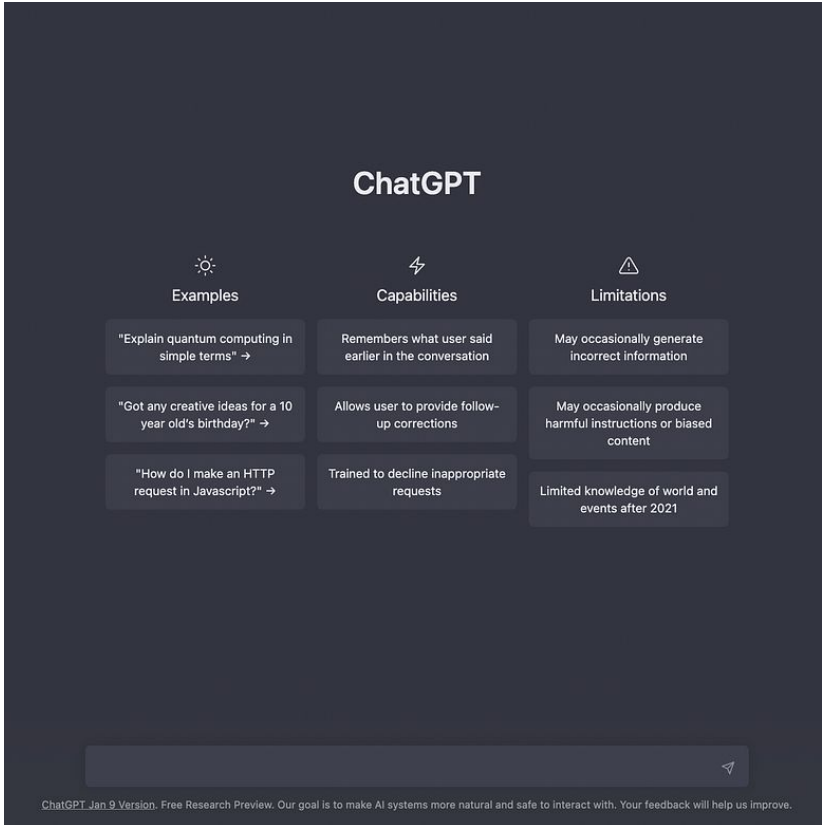
Sociotechnical Safety Evaluation of Generative AI Systems

Google DeepMind

Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, William Isaac

Generative AI

- Generative AI has taken the world by storm across many domains
  - Medical, education, news, art, music, etc.
- Systems are increasingly multimodal, but primarily focus on single modalities (text or images)



Generative AI introduces risks

Image: Cath Virginia / The Verge, Getty Images

SPOTIFY

Not even Spotify is safe from AI slop

How fake music targets real artists.

By **Elizabeth Lopatto**, a reporter who writes about tech, money, and human behavior. Sh
joined The Verge in 2014 as science editor. Previously, she was a reporter at Bloomberg.
Nov 14, 2024, 10:15 AM EST

19 Comments (19 New)

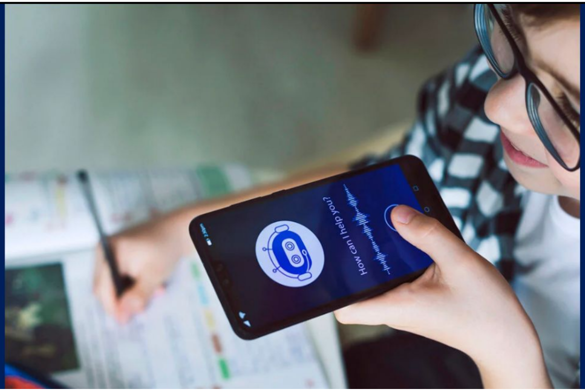
Middlebury Institute of International Studies / Academics / Centers and Initiatives /
Center on Terrorism, Extremism, and Counterterrorism / CTEC Publications

The Dangers of Generative AI and Extremism

Without Guardrails, Generative AI Can Harm Education

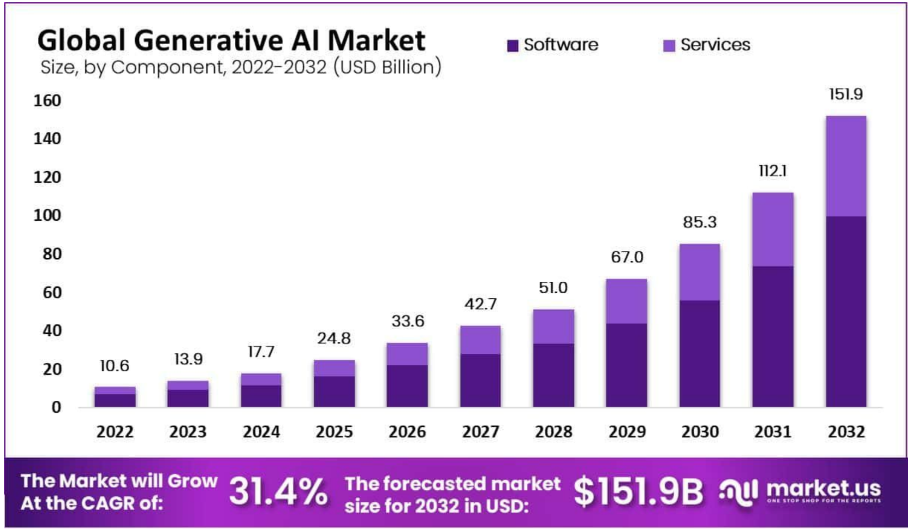
August 27, 2024 • 4 min read

Students who rely on generative AI to help them learn may be missing out on basic skills, according to a paper co-authored by Wharton's Hamsa Bastani.



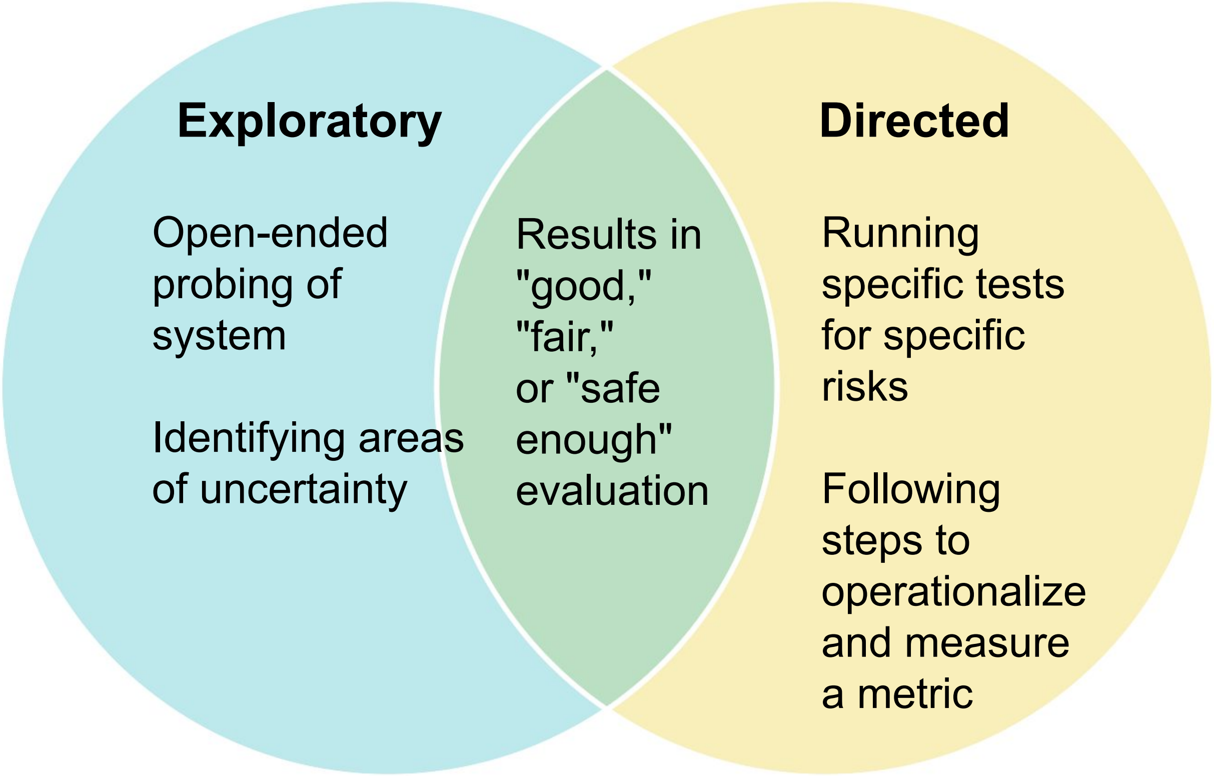
Evaluation is a growing priority

- Growing use of GenAI makes evaluation easier and more important
- Risks are a public safety concern
- Evaluation should be performed by developers, policy makers, and regulators



https://market.us/report/generative-ai-market/

Exploratory vs. directed evaluation



Evaluation is never neutral

- Evaluation "rests on interwoven technical and normative decisions"
  - Deciding what to evaluate
  - How to measure it
  - What results indicate "good" AI system performance

=> Outcome varies based on who is performing the evaluation

Main Contributions

1. **Sociotechnical framework for safety evaluation**
    - 3 layers: capability, human interaction, systemic

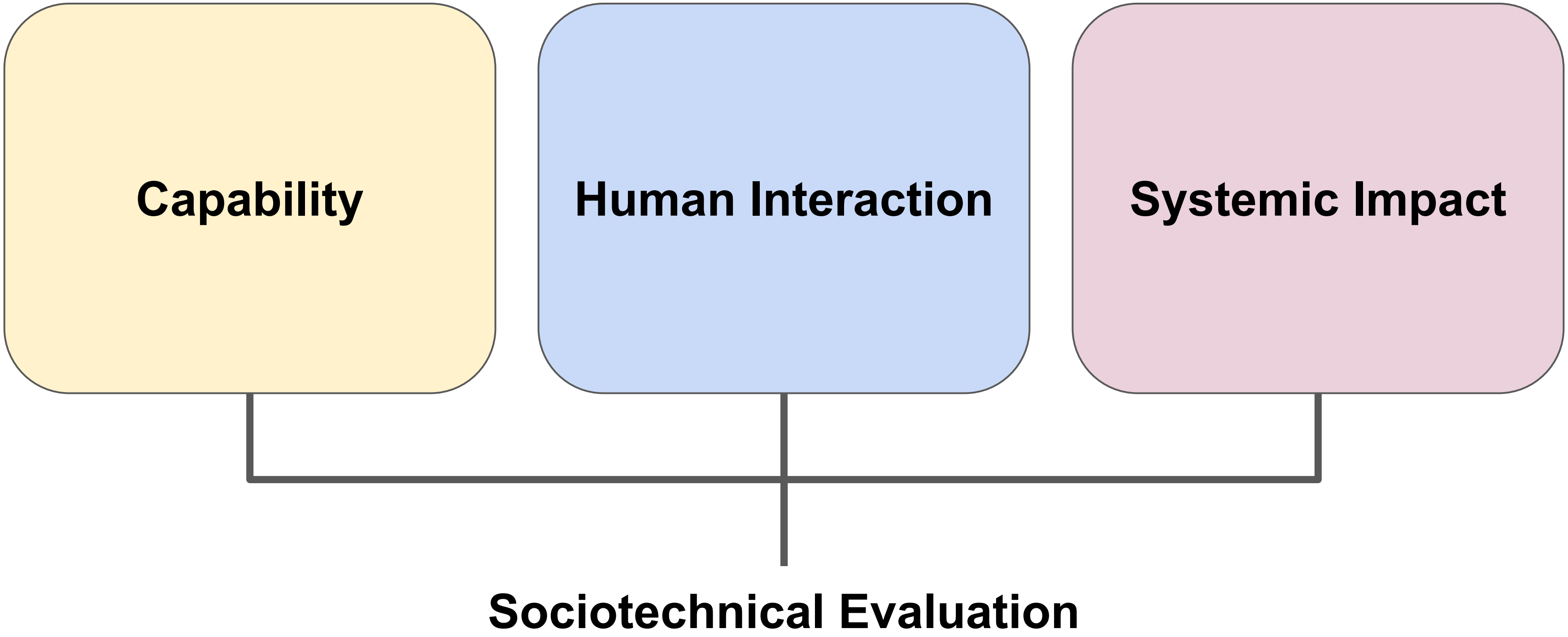
2. **Empirical assessment of current safety evaluation landscape**
    - Investigation of gaps in existing evaluations
    - Proposed steps for closing the gaps

Sociotechnical framework

- AI systems are **sociotechnical** systems
  - "Humans and machines are necessary to make the technology work"

- Evaluations have a **sociotechnical** gap
  - Evaluate only the technical components
  - Ignores the human and systemic factors that contribute to harm
  - Need an approach that encompasses all factors

Three evaluation layers



Capability evaluation

- Targets AI systems and their technical components
    - Testing behavior in response to certain tasks
- Indicates if an AI system is likely to cause downstream harm

- **Examples:**
    - Checking the extent that a model reproduces harmful stereotypes
    - Tracking metrics that indicate potential for environmental impact

Human interaction evaluation

- Looks at the experience of people interacting with an AI system
  - Does the AI system perform its intended function?
  - Do experiences differ between user groups?
- Indicates if an AI system is likely to cause unintended effects on the people interacting with or exposed to the outputs

- **Examples:**
  - A behavioral study on the overreliance or overtrust of AI system output
  - Does frequent exposure to AI increase feelings of social isolation?

Systemic impact evaluation

- Targets the impact of an AI system on the broader system it is used in
  - Society, economy, natural environment
- Some effects may only emerge if the AI system is deployed at scale

- **Examples:**
  - Observing widespread adoption and perception of AI systems in academic environments
  - Data collection for broader environmental impacts on ecosystems

Case study: Misinformation harms

Capability

Is the system likely to produce factually incorrect output?

Human Interaction

Will the misinformation produced influence public knowledge or beliefs? Will it erode trust?

Systemic Impact

Will there be organizations to confirm fact-check? How will the expectations of public information sharing change?

Assessment of current safety evaluation landscape

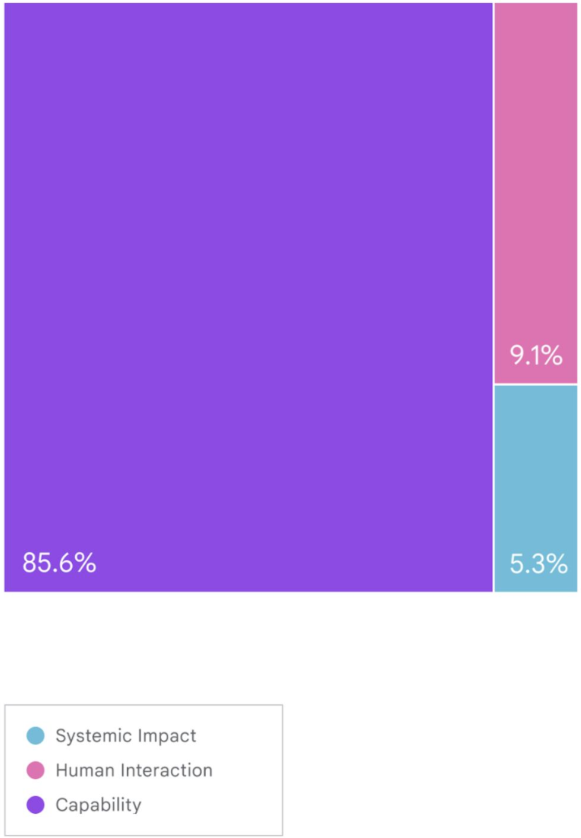
1. Three categories of gaps in the present state of safety evaluations
    - Coverage gap
    - Context gap
    - Multimodal gap
2. Practical steps to improve the safety evaluation landscape
    - Operationalising risk
    - Model-driven evaluation
    - Repurposing existing evaluations
    - Transcribing non-text outputs
3. Limits and next steps

Coverage gap

- Evaluations for many risk areas are lacking
  - Example: Environmental harm
- Many areas with evaluations are not comprehensively covered
  - Example: Representation harm
  - Heavily focused on gender and race
  - Leaves out age, religion, social class, etc.

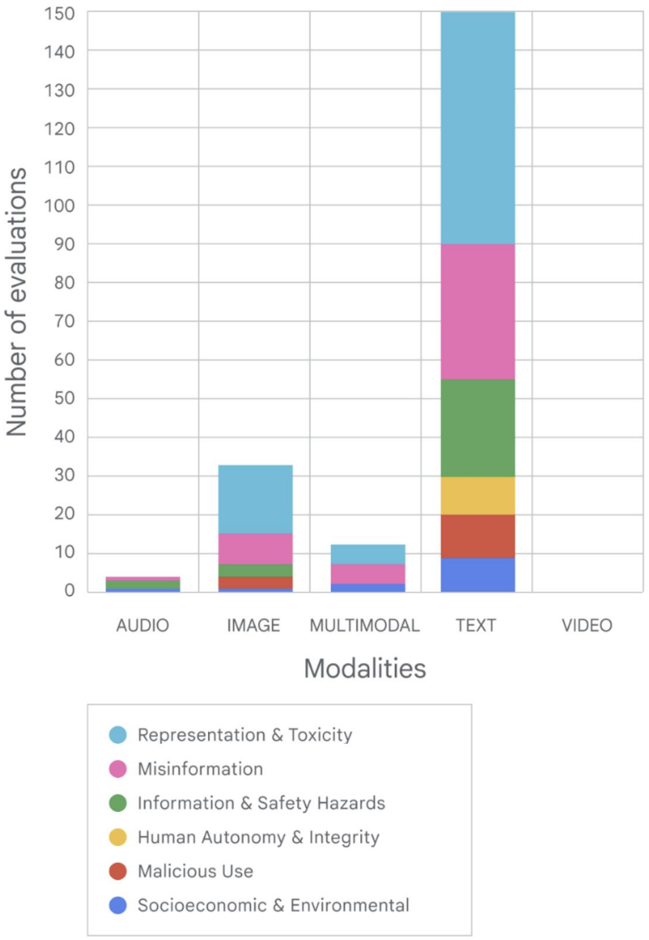
Context gap

- Current evaluations focus on technical abilities of systems
- Lack of evaluations considering human interactions and systemic impacts
  - More difficult to test than capabilities
  - … but they are just as important!



Multimodal gap

- Majority of evaluations are purely text-based
- Certain risks are more pronounced in other modalities
    - Example: violent content



Improvement: Operationalising Risks

- **Operationalisation**: turning complex and abstract risks into measurable metrics
- **Example:** Risk of building a bio-weapon with AI

Capability

Assess properties of model related to output of harmful biological information

Human Interaction

Likelihood of people following instructions to assemble bombs

Systemic Impact

Modeling potential distribution mechanisms for such created biohazards

Improvement: Model-Driven Evaluation

- Perform evaluation using pre-trained models
  - Can replace human efforts
  - Provide easy method for filling gaps in evaluation
- Downsides
  - Limited by biases of the evaluator model
  - Not accessible to all evaluators

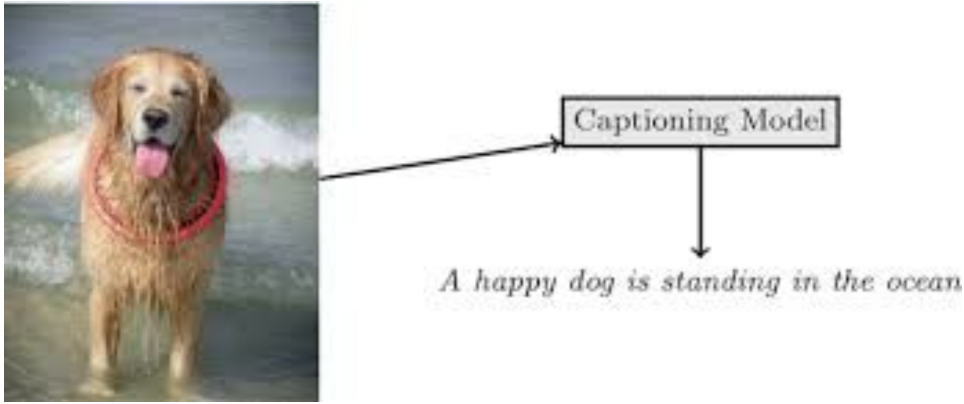
Improvement: Repurposing Existing Evaluations

- Existing text-based benchmarks can be repurposed for other modalities
  - Important to note limitations when switching contexts

- **Example:** Winogender (2018), a benchmark for gender bias in text-based LLMs, is now used to evaluate video LMs like DALLE2

Improvement: Transcribing Non-Text Outputs

- Transcribe other modalities to text
  - Automated Speech Recognition for audio output
  - Captioning for image and video output
- Allows text-based metrics to evaluate new modalities
- Downsides
  - Transcription is often lossy
  - May introduce other errors



Roles and Responsibilities in Evaluation

- AI developers
  - Capability evaluations and iterative improvements
- Application developers
  - Human interaction evaluations and real-world use testing
- Third party stakeholders
  - Independent system impact evaluations
- Public sector and civil society
  - Governance and regulation

Limits of Evaluation

- These evaluations are inherently incomplete
  - Impossible to predict the future of AI
  - “General Purpose” AI systems
- Complementary government mechanisms
  - Post-deployment risk monitoring
  - Swift intervention when necessary

Steps Forward

- Develop evaluations where they don’t exist yet
- Integrate evaluations into standard development process
- Give real importance to evaluations
- Move to a shared framework for AI safety
  - Convergence of risk domains
  - Dynamic risk mapping
  - Collaboration between research areas

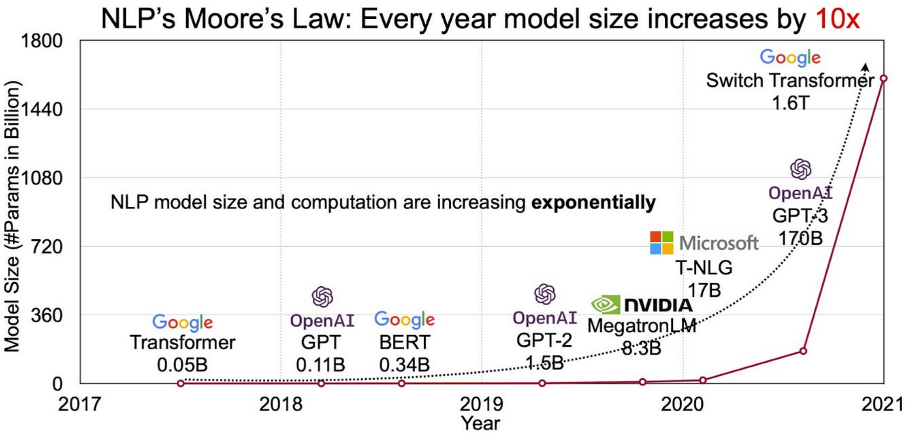
On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜

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

Motivation

The NLP community has been developing models with **increasing** size (number of parameters and size of training data).

Need to come up with methodologies for risk assessment and mitigation of those risks.



Background of language model (LM)

Before neural models:

N-gram models uses large amount to data for machine translation from another source language to English when direct translation examples are scarce.

Type of tasks is very limited in scope

Background of language model (LM)

The adaptation of word embeddings for labeling and classification

Word embeddings are pretrained representations of the distribution of words usually coming from well known datasets (word2vec)

Reduced the amount of labeled data necessary for training. Prompts for faster convergence and smaller training data size.

Background of language model (LM)

Modern transformer models have benefited from larger architectures and larger quantities of data.

As long as the increase in model size is correlated with increase in performance, we expect this trend to continue.

- LMs have shown strictly increasing performance in fluency and coherence scores (BLEU score)

But at what cost?

family	model	size (number of parameters in billion)	E2E	ViGGo	WikiTableText	DART	WebNLG
BART	BART-base	0.1	0.399	0.281	**0.421**	**0.423**	0.481
	BART-large	0.4	**0.403**	**0.283**	0.419	0.413	**0.503**
T5	T5-base	0.2	0.398	0.268	0.408	0.461	0.527
	T5-large	0.7	**0.411**	**0.302**	**0.431**	**0.479**	**0.546**
OPT	OPT-2.7B	2.7	0.350	0.262	0.421	0.441	0.521
	OPT-6.7B	6.7	**0.369**	**0.269**	**0.426**	0.448	0.538
	OPT-13B	13.0	0.347	0.269	0.412	**0.463**	**0.549**
BLOOM	BLOOM-1.1B	1.1	0.374	0.255	0.411	0.437	0.491
	BLOOM-3B	3.0	**0.380**	0.260	0.396	**0.446**	0.520
	BLOOM-7B	7.0	0.379	**0.274**	**0.423**	0.444	**0.530**
Llama 2	Llama2-7B	7.0	**0.419**	0.248	0.436	0.494	0.532
	Llama2-13B	13.0	0.408	**0.288**	**0.451**	**0.51**	**0.563**

An Overview Of Costs & Risks

- Environmental costs
- Financial costs
- Risk of substantial harm
- Opportunity cost in research

Environmental Cost

Energy consumption for training and inference is huge.

The majority of cloud computing providers still rely on fossil fuels

Consumption	CO_2e (lbs)
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

Consumer	Renew.	Gas	Coal	Nuc.
China	22%	3%	65%	4%
Germany	40%	7%	38%	13%
United States	17%	35%	27%	19%
Amazon-AWS	17%	24%	30%	26%
Google	56%	14%	15%	10%
Microsoft	32%	23%	31%	10%

Environmental Cost

Renewable sources are still costly to build infrastructure (ex. make space for wind/solar farms)

Climate change is impacting the world's marginalized communities

4 typhoons have hit the Philippines in just the past 10 days

It's the most active November on record after a slow start to the 2024 Pacific typhoon season that's now in hyperdrive.

Financial Cost

Little literature in the research community promotes measure of **efficiency** as primary contribution.

From estimations:

- An increase in 0.1 BLEU score in machine translation scores results in an increase of $ 150,000
- Amount of compute has increased 300,000x in 6 years

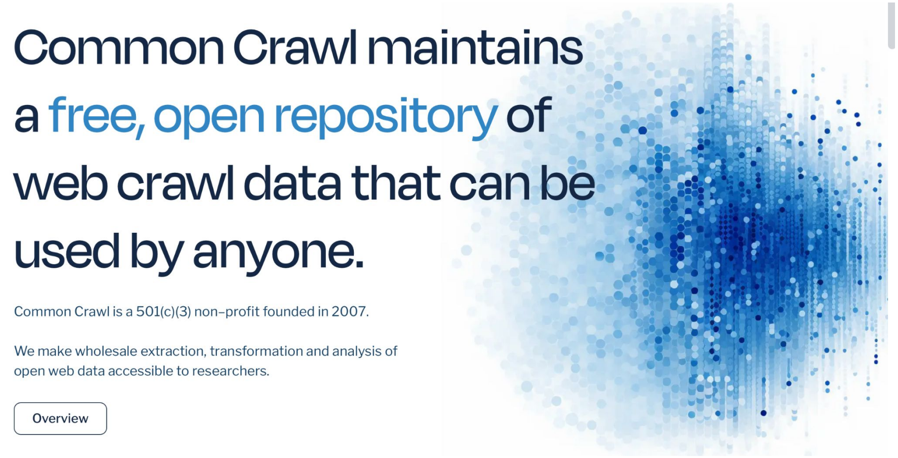
Trade-off between (energy) cost and performance needs to be weighed carefully to reduce negative environmental impact and inequitable access to resources.

Risk of Substantial Harm - Hegemonic LM

How are the current language models trained?

GPT3 uses Common Crawl (collected over 8 years of web crawling)

The training data **lacks representation** of the general population



Problematic Training Data

Web data has a narrow participation

- Access to Internet is not evenly distributed
- Subcommunities on web can be structurally biased
- Ineffective content moderation could limit participation
- Niche online communities could be omitted in web crawl

Reddit users and news users more likely to be male and young

% of U.S. adults, Reddit users and Reddit news users who are ...

	U.S. adults %	Reddit users %	Reddit news users %
Men	49	67	71
Women	51	33	29
18-29	22	64	59
30-49	34	29	33
50-64	25	6	7
65+	19	1	<1
College degree	28	42	48
Some college	31	40	43
High school or less	41	18	9
White non-Hispanic	65	70	74
Black non-Hispanic	12	7	8
Hispanic	15	12	7
Other non-Hispanic	8	11	10

Problematic Training Data

What is the result?

Dominant viewpoints perpetuates

Increase power imbalances

What we can do?

Don't aim **solely for scale**

Be thoughtful of what to include in training dataset

hegemony noun

he·ge·mo·ny hi-ˈje-mə-nē -ˈge- ˈhe-jə-ˌmō-nē

Synonyms of *hegemony* >

1 : preponderant influence or authority over others : **DOMINATION**
battled for *hegemony* in Asia

2 : the social, cultural, ideological, or economic influence exerted by a dominant group

Changing Social Views

Old data used to train LM may misrepresent movements and misaligns social value.

Poor documented movements may lose in the process.

Dealing with Encoded Bias

The reflection of training data characteristics may include…

- Stereotypical association towards specific group
- Effects of intersectionality of bias towards certain identity

These biases are learnt by language models

Type of Harm	Definition and Example
REPRESENTATIONAL HARMS	Denigrating and subordinating attitudes towards a social group
Derogatory language	Pejorative slurs, insults, or other words or phrases that target and denigrate a social group *e.g.,* `"Whore"` *conveys hostile and contemptuous female expectations* (Beukeboom and Burgers 2019)
Disparate system performance	Degraded understanding, diversity, or richness in language processing or generation between social groups or linguistic variations *e.g., AAE* like* `"he woke af"` *is misclassified as not English more often than SAE*[†] *equivalents* (Blodgett and O'Connor 2017)
Erasure	Omission or invisibility of the language and experiences of a social group *e.g.,* `"All lives matter"` *in response to* `"Black lives matter"` *implies colorblindness that minimizes systemic racism* (Blodgett 2021)
Exclusionary norms	Reinforced normativity of the dominant social group and implicit exclusion or devaluation of other groups *e.g.,* `"Both genders"` *excludes non-binary identities* (Bender et al. 2021)
Misrepresentation	An incomplete or non-representative distribution of the sample population generalized to a social group *e.g., Responding* `"I'm sorry to hear that"` *to* `"I'm an autistic dad"` *conveys a negative misrepresentation of autism* (Smith et al. 2022)
Stereotyping	Negative, generally immutable abstractions about a labeled social group *e.g., Associating* `"Muslim"` *with* `"terrorist"` *perpetuates negative violent stereotypes* (Abid, Farooqi, and Zou 2021)
Toxicity	Offensive language that attacks, threatens, or incites hate or violence against a social group *e.g.,* `"I hate Latinos"` *is disrespectful and hateful* (Dixon et al. 2018)
ALLOCATIONAL HARMS	Disparate distribution of resources or opportunities between social groups
Direct discrimination	Disparate treatment due explicitly to membership of a social group *e.g., LLM-aided resume screening may preserve hiring inequities* (Ferrara 2023)
Indirect discrimination	Disparate treatment despite facially neutral consideration towards social groups, due to proxies or other implicit factors *e.g., LLM-aided healthcare tools may use proxies associated with demographic factors that exacerbate inequities in patient care* (Ferrara 2023)

*African-American English; [†]Standard American English

Dealing with Encoded Bias

GPT-2’s training data may include up to **272**k documents from unreliable news sites and **63**k from banned Reddit threads.

Challenges on evaluating encoded bias?

- Response varies across specific demographic group
- Requirement of *a priori* knowledge of social category to be evaluated
- Operationalizing new definitions into algorithms is political

Curation, Documentation, Accountability

Curation of dataset: **selecting** and **collecting** training data from online corpus

Documentation of dataset: description of research goals, value, and motivations **underlying** data selection and collection process

Why is this important?

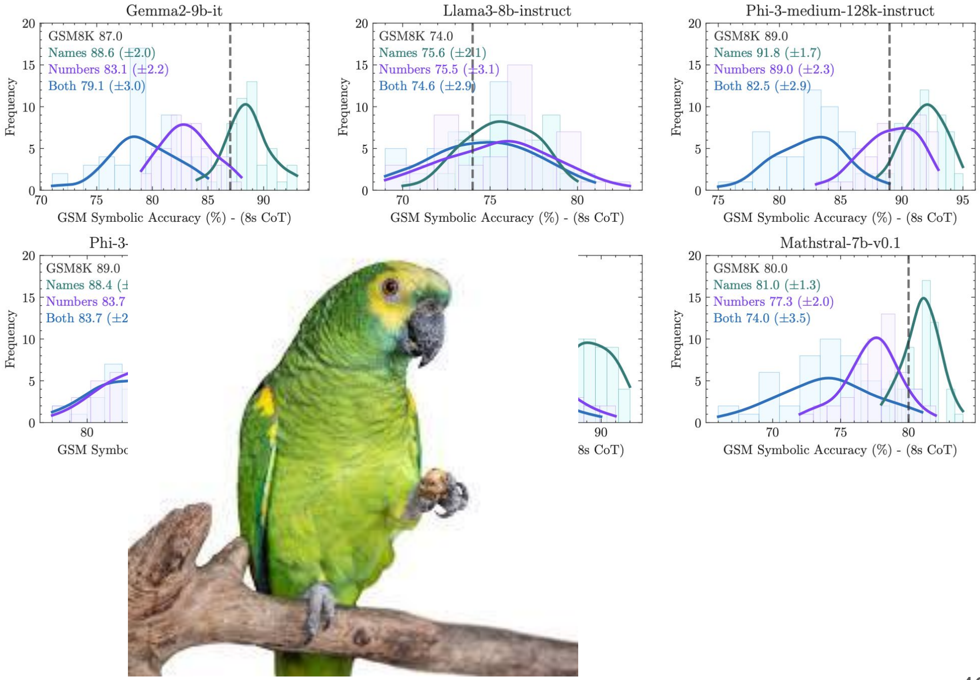
- Help hold accountability
- Help people understand training data characteristics

Opportunity Cost in Research - Stochastic Parrots

Are language models actually good at natural language understanding?

LM stitches together sequence of linguistic forms observed in training data

… like stochastic parrots



Stochastic Parrots

Why is this a problem?

LM generated text is **not** grounded in communicative intent (understanding the subject's intentions within context). But for most of the time… human may mistake LM output for **meaningful** text

More time can be spent on dataset curation and applying meaning capturing approaches.

Risks and Harms for LM in a Nutshell

A hegemonic worldview that encompasses encoded biases

- Negative reinforcement for for underrepresented and marginalized groups
- Potential amplification of biases and stereotypes
- Allocational and reputational harm when affecting system decisions

Bad actors with a biased system

- Seemingly coherent texts can be used to deceive the general public

Elicitation of sensitive training data

Misalignment of communicative intention

Solutions & Looking Forward

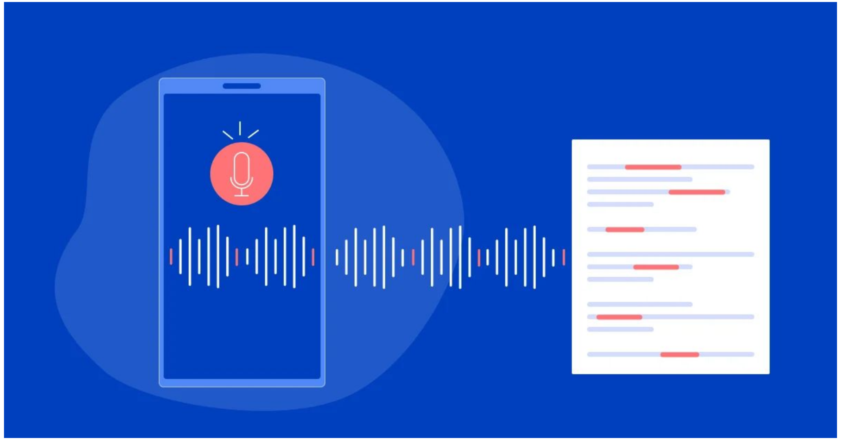
Rethinking language models to prioritize efficiency and inclusivity

- Document environmental, social, and use-case impacts upfront
- Careful curation of dataset with documentation on selection and collection and make note of users & stakeholders
- **Pre-mortem analysis** of systems to reverse engineer previously unanticipated causes and explore alternative paths

Solutions & Looking Forward

Human-centered LM development

- **Value sensitive design** ensures communication with stakeholders early on and align systems with their values
- Recognize **synthetic human behavior** modeling as a critical ethical boundary
- Be mindful to **dual use problems** and think about the downstream effect of LM development early on



Thank you!