Critique of LLMs

Indira Sen, SILLM Lecture 11

Agenda

- Theoretical and Conceptual Critiques
- Practical issues
  - Data contamination
  - Misrepresentation
  - Hallucinations
  - Malicious uses
  - ...?
- Regulating LLMs

Illustration by Vivek Thakker

ANNALS OF ARTIFICIAL INTELLIGENCE

CHATGPT IS A BLURRY JPEG OF THE WEB

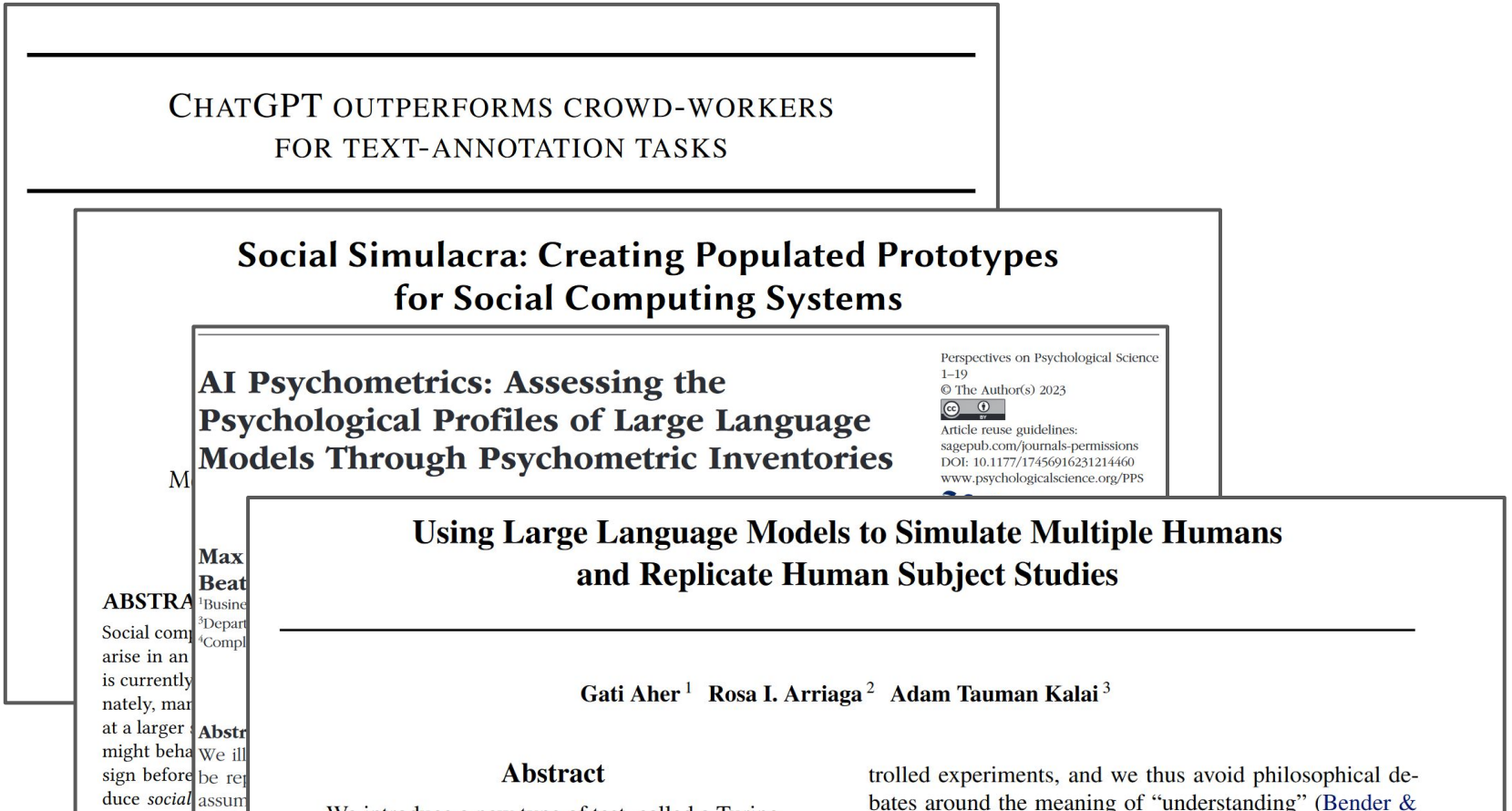
OpenAI's chatbot offers paraphrases, whereas Google offers quotes. Which do we prefer?

By Ted Chiang

February 9, 2023

What can LLMs do well? What can't they?

Recap: The different ways we've seen LLMs being used



CHATGPT OUTPERFORMS CROWD-WORKERS FOR TEXT-ANNOTATION TASKS

Social Simulacra: Creating Populated Prototypes for Social Computing Systems

AI Psychometrics: Assessing the Psychological Profiles of Large Language Models Through Psychometric Inventories

Perspectives on Psychological Science
1–19
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/17456916231214460
www.psychologicalscience.org/PPS

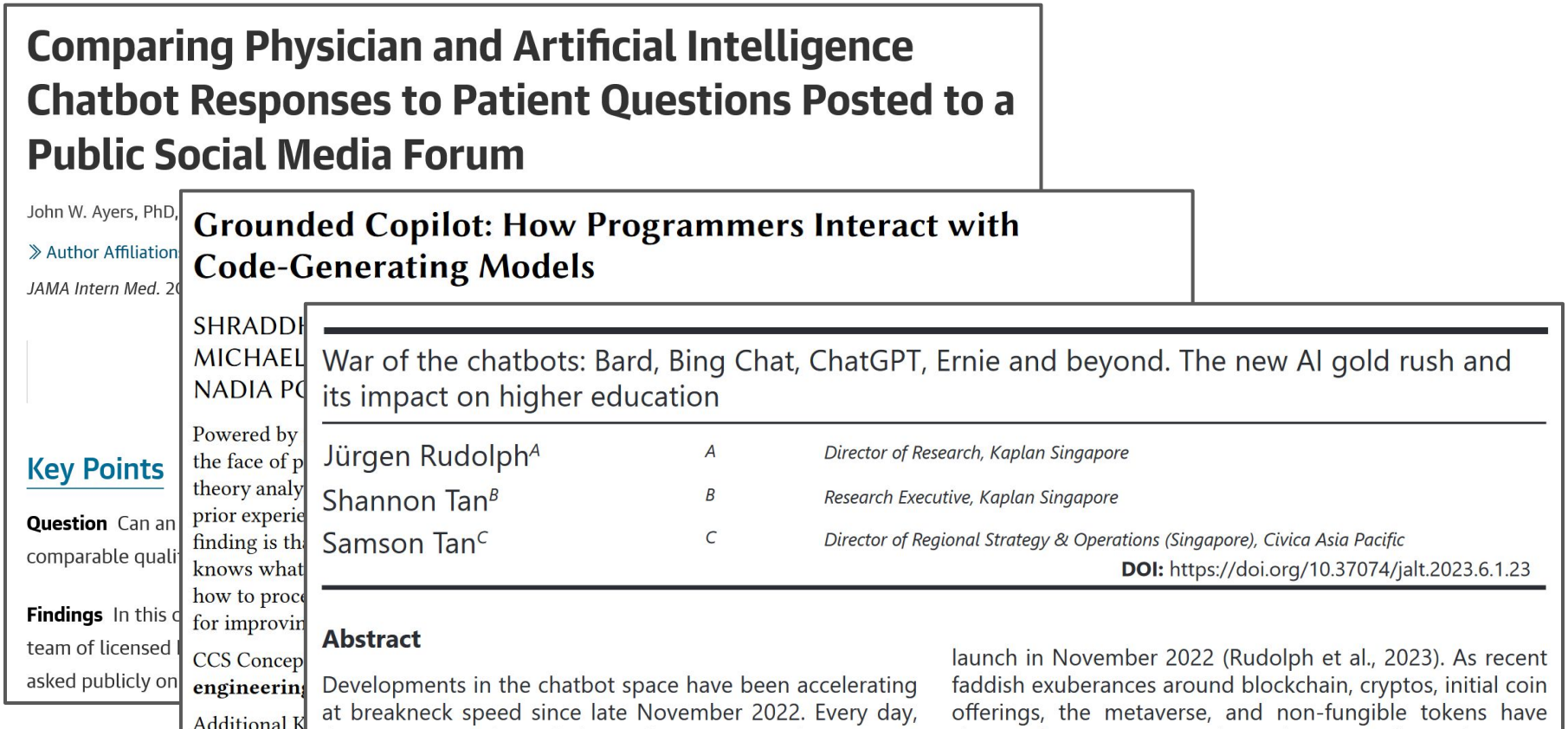
Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies

Gati Aher [1] Rosa I. Arriaga [2] Adam Tauman Kalai [3]

Abstract

trolled experiments, and we thus avoid philosophical de-
bates around the meaning of "understanding" (Bender &

Recap: other ways LLMs are being used



Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum

John W. Ayers, PhD,

» Author Affiliation

JAMA Intern Med. 2

Key Points

Question Can an

comparable quali

Findings In this

team of licensed

asked publicly on

Grounded Copilot: How Programmers Interact with Code-Generating Models

SHRADD

MICHAEL

NADIA P

Powered by

the face of p

theory analy

prior experie

finding is th

knows what

how to proc

for improvin

CCS Concep

engineerin

Additional K

War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education

Jürgen Rudolph[A] A *Director of Research, Kaplan Singapore*

Shannon Tan[B] B *Research Executive, Kaplan Singapore*

Samson Tan[C] C *Director of Regional Strategy & Operations (Singapore), Civica Asia Pacific*

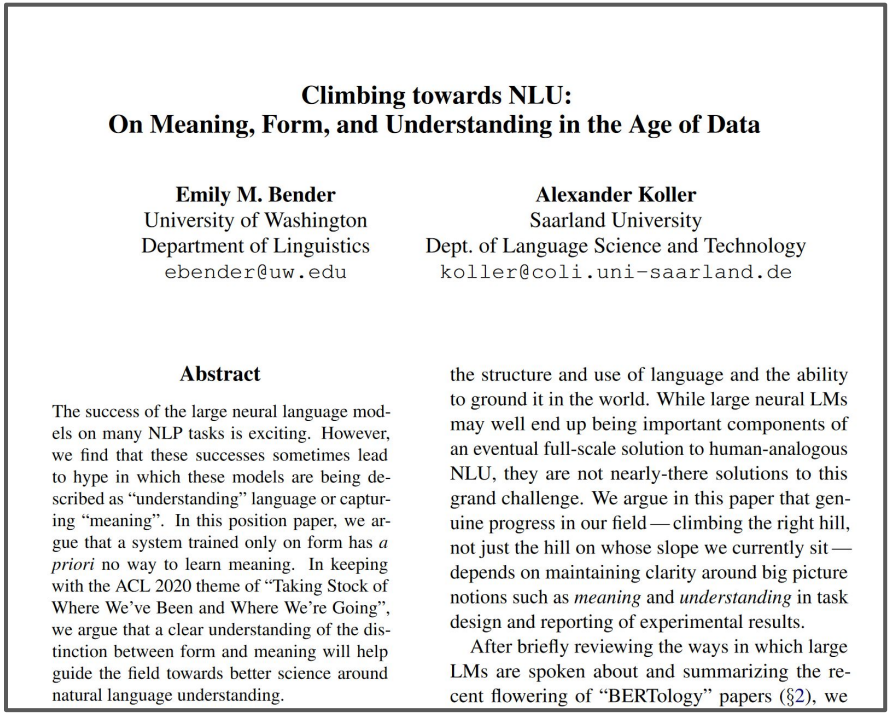
DOI: https://doi.org/10.37074/jalt.2023.6.1.23

Abstract

Developments in the chatbot space have been accelerating at breakneck speed since late November 2022. Every day, launch in November 2022 (Rudolph et al., 2023). As recent faddish exuberances around blockchain, cryptos, initial coin offerings, the metaverse, and non-fungible tokens have

Do LLMs actually ‘understand’ language?

- Form vs. meaning in (L)LMs
  - **Form**: physical manifestations of communications, words on papers, pixels in images, etc
  - **Meaning**: what is actually being communicated
- Form != meaning
  - Pragmatics
  - prosody



Climbing towards NLU:
On Meaning, Form, and Understanding in the Age of Data

Emily M. Bender
University of Washington
Department of Linguistics
ebender@uw.edu

Alexander Koller
Saarland University
Dept. of Language Science and Technology
koller@coli.uni-saarland.de

Abstract

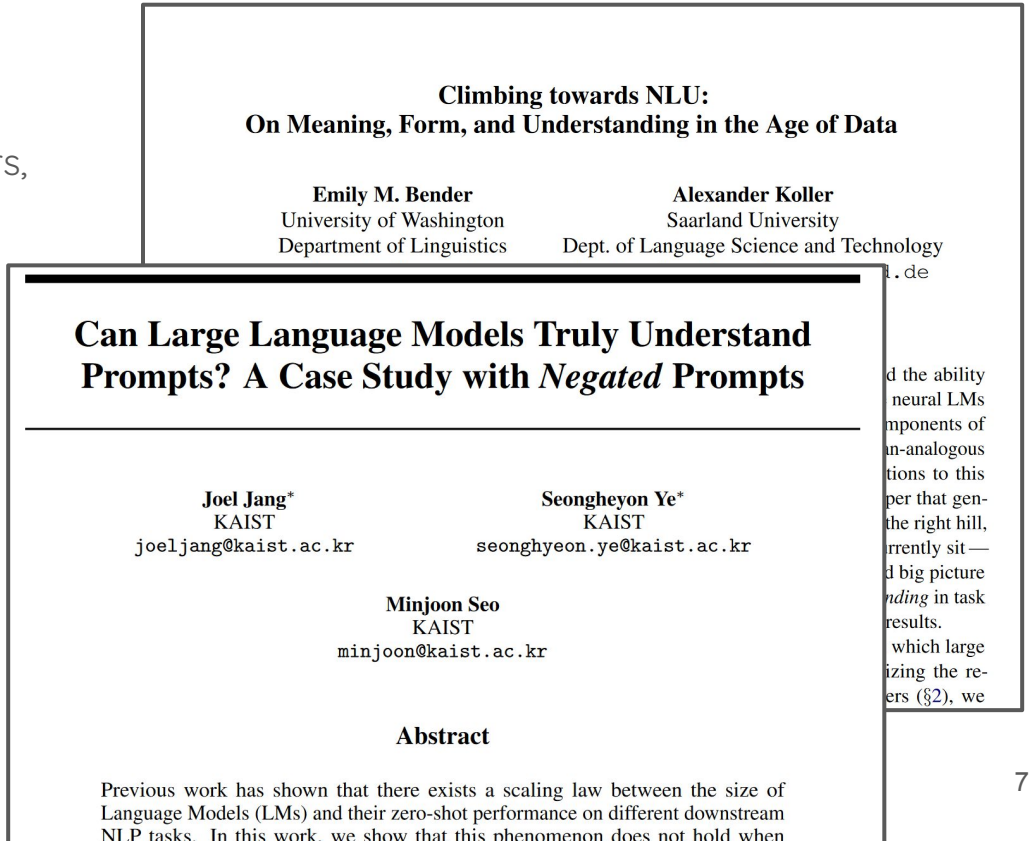
The success of the large neural language models on many NLP tasks is exciting. However, we find that these successes sometimes lead to hype in which these models are being described as “understanding” language or capturing “meaning”. In this position paper, we argue that a system trained only on form has *a priori* no way to learn meaning. In keeping with the ACL 2020 theme of “Taking Stock of Where We’ve Been and Where We’re Going”, we argue that a clear understanding of the distinction between form and meaning will help guide the field towards better science around natural language understanding.

the structure and use of language and the ability to ground it in the world. While large neural LMs may well end up being important components of an eventual full-scale solution to human-analogous NLU, they are not nearly-there solutions to this grand challenge. We argue in this paper that genuine progress in our field — climbing the right hill, not just the hill on whose slope we currently sit — depends on maintaining clarity around big picture notions such as *meaning* and *understanding* in task design and reporting of experimental results.

After briefly reviewing the ways in which large LMs are spoken about and summarizing the recent flowering of “BERTology” papers (§2), we

Do LLMs actually ‘understand’ language?

- Form vs. meaning in (L)LMs
  - **Form**: physical manifestations of communications, words on papers, pixels in images, etc
  - **Meaning**: what is actually being communicated
- Form != meaning
  - Pragmatics
  - Prosody
- LLMs sometimes fail, e.g., Negated prompts



Climbing towards NLU:
On Meaning, Form, and Understanding in the Age of Data

Emily M. Bender
University of Washington
Department of Linguistics

Alexander Koller
Saarland University
Dept. of Language Science and Technology

Can Large Language Models Truly Understand Prompts? A Case Study with *Negated* Prompts

Joel Jang*
KAIST
joeljang@kaist.ac.kr

Seongheyon Ye*
KAIST
seonghyeon.ye@kaist.ac.kr

Minjoon Seo
KAIST
minjoon@kaist.ac.kr

Abstract

Previous work has shown that there exists a scaling law between the size of Language Models (LMs) and their zero-shot performance on different downstream NLP tasks. In this work, we show that this phenomenon does not hold when

Critiques of LLMs' practical uses

General issues

- **Multilinguality**: mBERT does not perform equally well for all languages (Wu and Drezde, 2020)
- Risks and benefits to different communities is not equitably distributed
- Large corpora used to train models are poorly understood

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

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

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

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, 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 (FAccT '21), March 3–10, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3442188.3445922

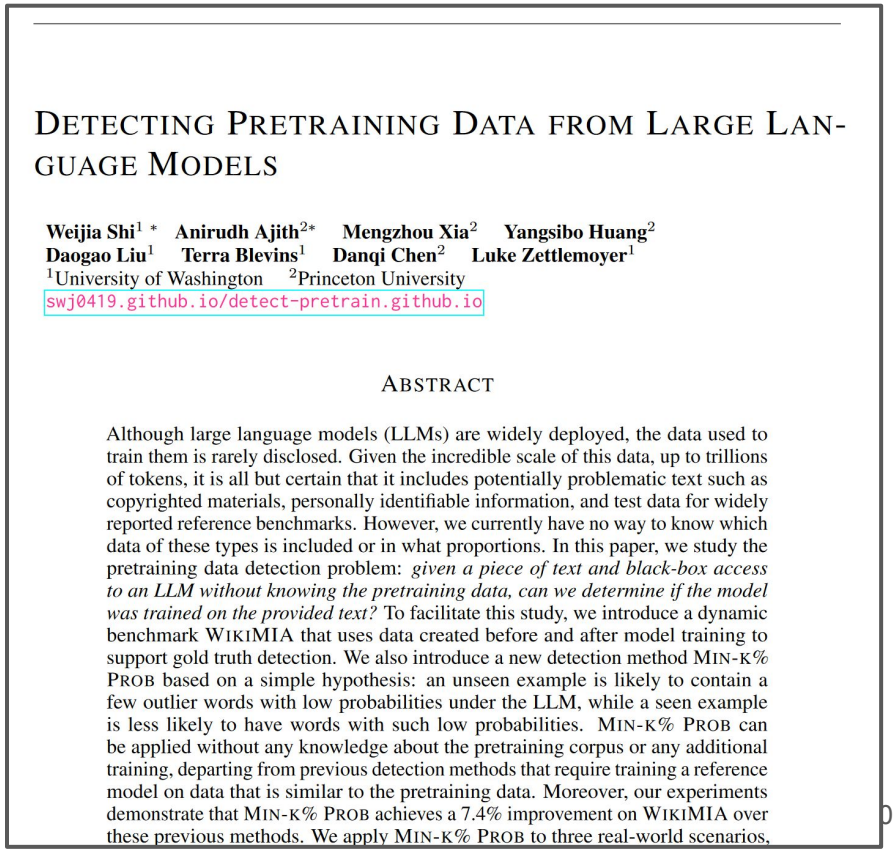
alone, we have seen the emergence of BERT and its variants [39, 70, 74, 113, 146], GPT-2 [106], T-NLG [112], GPT-3 [25], and most recently Switch-C [43], with institutions seemingly competing to produce ever larger LMs. While investigating properties of LMs and how they change with size holds scientific 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.

Using LLMs: Data Contamination

- LLMs memorize a lot of content from their training data
- This is called ‘memory leakage’ or ‘data contamination’
- Can be detected using memory inference attacks
- But, so what?



DETECTING PRETRAINING DATA FROM LARGE LANGUAGE MODELS

Weijia Shi[1] * **Anirudh Ajith**[2]* **Mengzhou Xia**[2] **Yangsibo Huang**[2]
Daogao Liu[1] **Terra Blevins**[1] **Danqi Chen**[2] **Luke Zettlemoyer**[1]
[1]University of Washington [2]Princeton University
swj0419.github.io/detect-pretrain.github.io

ABSTRACT

Although large language models (LLMs) are widely deployed, the data used to train them is rarely disclosed. Given the incredible scale of this data, up to trillions of tokens, it is all but certain that it includes potentially problematic text such as copyrighted materials, personally identifiable information, and test data for widely reported reference benchmarks. However, we currently have no way to know which data of these types is included or in what proportions. In this paper, we study the pretraining data detection problem: *given a piece of text and black-box access to an LLM without knowing the pretraining data, can we determine if the model was trained on the provided text?* To facilitate this study, we introduce a dynamic benchmark WIKIMIA that uses data created before and after model training to support gold truth detection. We also introduce a new detection method MIN-K% PROB based on a simple hypothesis: an unseen example is likely to contain a few outlier words with low probabilities under the LLM, while a seen example is less likely to have words with such low probabilities. MIN-K% PROB can be applied without any knowledge about the pretraining corpus or any additional training, departing from previous detection methods that require training a reference model on data that is similar to the pretraining data. Moreover, our experiments demonstrate that MIN-K% PROB achieves a 7.4% improvement on WIKIMIA over these previous methods. We apply MIN-K% PROB to three real-world scenarios,

Using LLMs: Data Contamination

- LLMs memorize a lot of content
- T
- C
- E

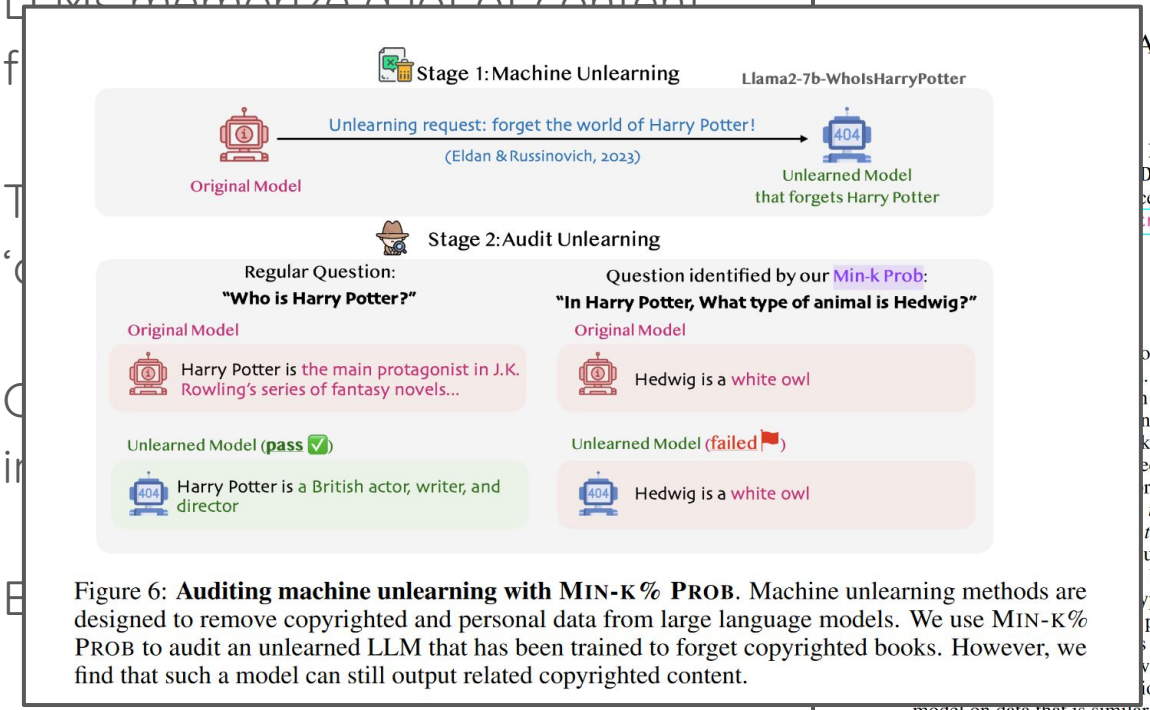


Figure 6: **Auditing machine unlearning with MIN-K% PROB**. Machine unlearning methods are designed to remove copyrighted and personal data from large language models. We use MIN-K% PROB to audit an unlearned LLM that has been trained to forget copyrighted books. However, we find that such a model can still output related copyrighted content.

AINING DATA FROM LARGE LAN-

Mengzhou Xia[2] Yangsibo Huang[2]
Danqi Chen[2] Luke Zettlemoyer[1]
ceton University
rain.github.io

ABSTRACT

odels (LLMs) are widely deployed, the data used to . Given the incredible scale of this data, up to trillions n that it includes potentially problematic text such as nally identifiable information, and test data for widely ks. However, we currently have no way to know which ed or in what proportions. In this paper, we study the roblem: *given a piece of text and black-box access the pretraining data, can we determine if the model text?* To facilitate this study, we introduce a dynamic uses data created before and after model training to We also introduce a new detection method MIN-K% ypothesis: an unseen example is likely to contain a probabilities under the LLM, while a seen example with such low probabilities. MIN-K% PROB can vledge about the pretraining corpus or any additional ious detection methods that require training a reference model on data that is similar to the pretraining data. Moreover, our experiments demonstrate that MIN-K% PROB achieves a 7.4% improvement on WIKIMIA over these previous methods. We apply MIN-K% PROB to three real-world scenarios,

Using LLMs: Data Contamination

- LLMs memorize a lot of content
- T
- C
- E

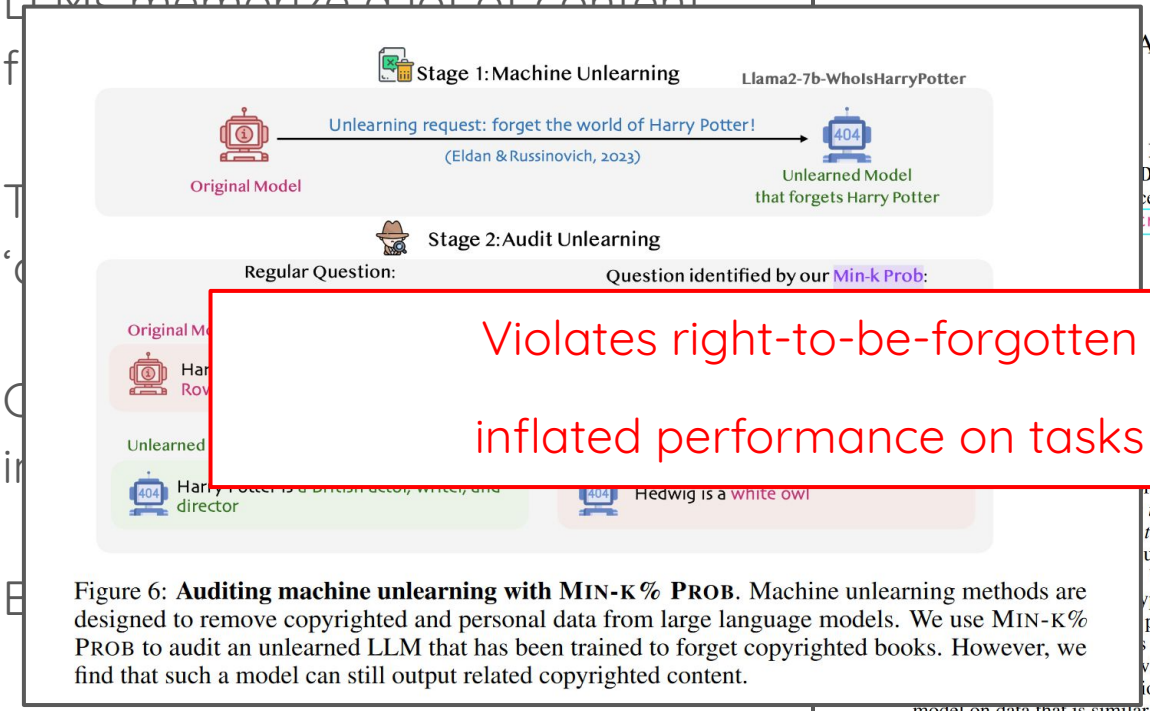


Figure 6: **Auditing machine unlearning with MIN-K% PROB**. Machine unlearning methods are designed to remove copyrighted and personal data from large language models. We use MIN-K% PROB to audit an unlearned LLM that has been trained to forget copyrighted books. However, we find that such a model can still output related copyrighted content.

Violates right-to-be-forgotten

inflated performance on tasks

AINING DATA FROM LARGE LAN-

Mengzhou Xia[2] Yangsibo Huang[2]
Danqi Chen[2] Luke Zettlemoyer[1]
ceton University
rain.github.io

dely deployed, the data used to
scale of this data, up to trillions
ntially problematic text such as
rmation, and test data for widely
ently have no way to know which
ions. In this paper, we study the
roblem: *given a piece of text and black-box access*
the pretraining data, can we determine if the model
text? To facilitate this study, we introduce a dynamic
uses data created before and after model training to
We also introduce a new detection method MIN-K%
ypothesis: an unseen example is likely to contain a
probabilities under the LLM, while a seen example
with such low probabilities. MIN-K% PROB can
vledge about the pretraining corpus or any additional
ious detection methods that require training a reference
model on data that is similar to the pretraining data. Moreover, our experiments
demonstrate that MIN-K% PROB achieves a 7.4% improvement on WIKIMIA over
these previous methods. We apply MIN-K% PROB to three real-world scenarios,

Using LLMs: Data Contamination

- Harder to detect in closed models
- The ‘speak, memory’ paper creates a novel task: predicting a minor character’s name in books
- People are bad at this
- Chatgpt and GPT-4 are good at some book characters, and good at doing tasks for those books

Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4

Kent K. Chang
University of California, Berkeley
kentkchang@berkeley.edu

Mackenzie Cramer
University of California, Berkeley
mackenzie.hanh@berkeley.edu

Sandeep Soni
Emory University
sandeep.soni@emory.edu

David Bamman*
University of California, Berkeley
dbamman@berkeley.edu

Abstract

In this work, we carry out a data archaeology to infer books that are known to ChatGPT and GPT-4 using a *name cloze* membership inference query. We find that OpenAI models have memorized a wide collection of copyrighted materials, and that the degree of memorization is tied to the frequency with which passages of those books appear on the web. The ability of these models to memorize an unknown set of books complicates assessments of measurement validity for cultural analytics by contaminating test data; we show that models perform much better on memorized books than on non-memorized books for downstream tasks. We argue that this supports a case for open models whose training data is known.

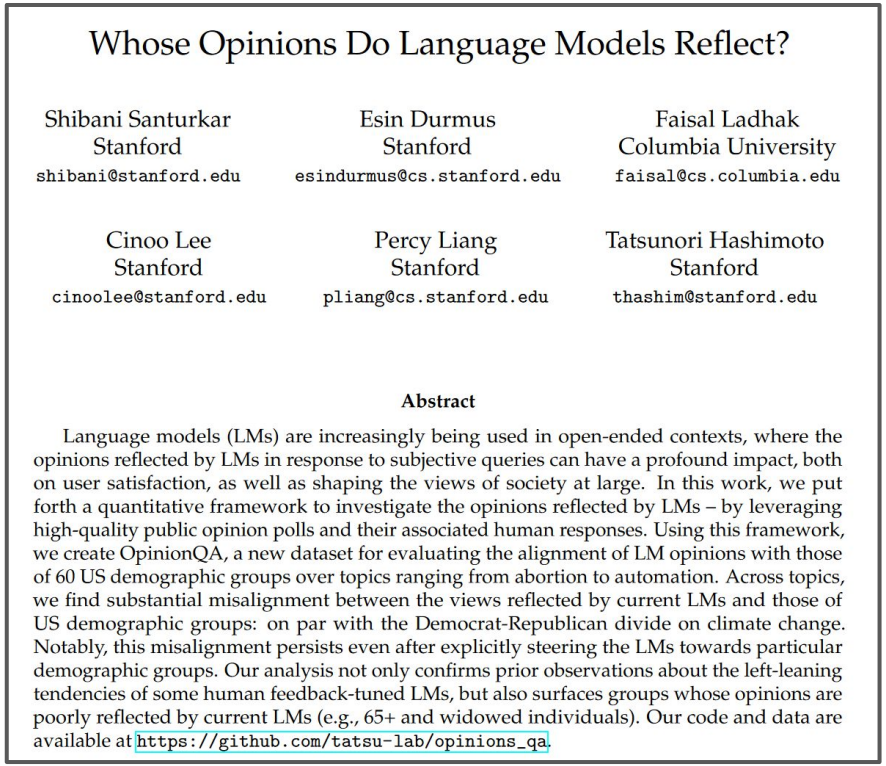
Wow. I sit down, fish the questions from my backpack, and go through them, inwardly cursing [MASK] for not providing me with a brief biography. I know nothing about this man I’m about to interview. He could be ninety or he could be thirty. → **Kate** (James, *Fifty Shades of Grey*).

Some days later, when the land had been moistened by two or three heavy rains, [MASK] and his family went to the farm with baskets of seed-yams, their hoes and machetes, and the planting began. → **Okonkwo** (Achebe, *Things Fall Apart*).

Figure 1: Name cloze examples. GPT-4 answers both of these correctly.

Using LLMs: Misrepresentation

- substantial misalignment between LLM views and those of US demographic groups
- misalignment persists even *after* steering
- 65+ and widowed individuals are not well-reflected in LLMs



Whose Opinions Do Language Models Reflect?

Shibani Santurkar
Stanford
shibani@stanford.edu

Esin Durmus
Stanford
esindurmus@cs.stanford.edu

Faisal Ladhak
Columbia University
faisal@cs.columbia.edu

Cinoo Lee
Stanford
cinoolee@stanford.edu

Percy Liang
Stanford
pliang@cs.stanford.edu

Tatsunori Hashimoto
Stanford
thashim@stanford.edu

Abstract

Language models (LMs) are increasingly being used in open-ended contexts, where the opinions reflected by LMs in response to subjective queries can have a profound impact, both on user satisfaction, as well as shaping the views of society at large. In this work, we put forth a quantitative framework to investigate the opinions reflected by LMs – by leveraging high-quality public opinion polls and their associated human responses. Using this framework, we create OpinionQA, a new dataset for evaluating the alignment of LM opinions with those of 60 US demographic groups over topics ranging from abortion to automation. Across topics, we find substantial misalignment between the views reflected by current LMs and those of US demographic groups: on par with the Democrat-Republican divide on climate change. Notably, this misalignment persists even after explicitly steering the LMs towards particular demographic groups. Our analysis not only confirms prior observations about the left-leaning tendencies of some human feedback-tuned LMs, but also surfaces groups whose opinions are poorly reflected by current LMs (e.g., 65+ and widowed individuals). Our code and data are available at https://github.com/tatsu-lab/opinions_qa.

LIONEL HUTZ WOULD BE ASHAMED —

Lawyers have real bad day in court after citing fake cases made up by ChatGPT

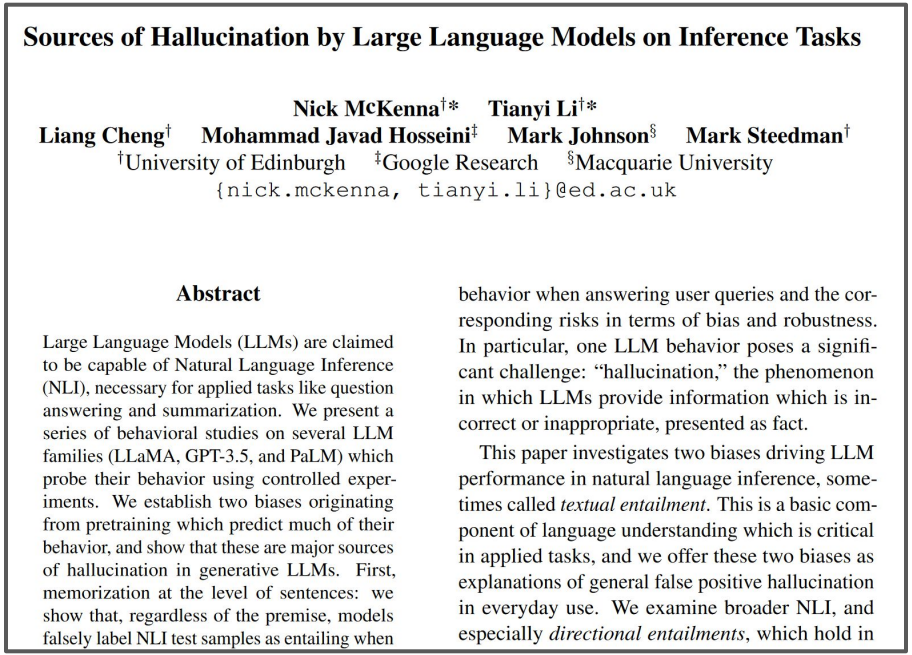
Lawyers fined $5K and lose case after using AI chatbot "gibberish" in filings.

JON BRODKIN - 6/23/2023, 7:32 PM

https://arstechnica.com/tech-policy/2023/06/lawyers-have-real-bad-day-in-court-after-citing-fake-cases-made-up-by-chatgpt/

Using LLMs: Hallucinations

- LLMs hallucinate when they ‘generate text that is nonsensical, or unfaithful to the provided source input’
- Makes LLMs unreliable for knowledge-related tasks
  - Scientific writing
  - Fact-checking
  - Editing wikipedia
  - ?



Sources of Hallucination by Large Language Models on Inference Tasks

Nick McKenna[†]* **Tianyi Li**[†]*
Liang Cheng[†] **Mohammad Javad Hosseini**[‡] **Mark Johnson**[§] **Mark Steedman**[†]
[†]University of Edinburgh [‡]Google Research [§]Macquarie University
{nick.mckenna, tianyi.li}@ed.ac.uk

Abstract

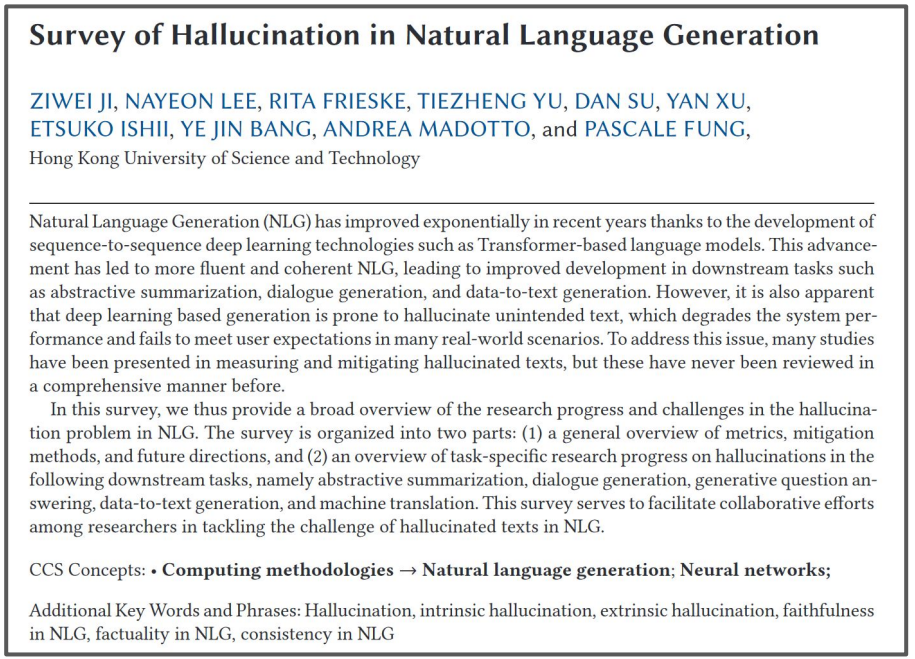
Large Language Models (LLMs) are claimed to be capable of Natural Language Inference (NLI), necessary for applied tasks like question answering and summarization. We present a series of behavioral studies on several LLM families (LLaMA, GPT-3.5, and PaLM) which probe their behavior using controlled experiments. We establish two biases originating from pretraining which predict much of their behavior, and show that these are major sources of hallucination in generative LLMs. First, memorization at the level of sentences: we show that, regardless of the premise, models falsely label NLI test samples as entailing when behavior when answering user queries and the corresponding risks in terms of bias and robustness. In particular, one LLM behavior poses a significant challenge: “hallucination,” the phenomenon in which LLMs provide information which is incorrect or inappropriate, presented as fact.

This paper investigates two biases driving LLM performance in natural language inference, sometimes called *textual entailment*. This is a basic component of language understanding which is critical in applied tasks, and we offer these two biases as explanations of general false positive hallucination in everyday use. We examine broader NLI, and especially *directional entailments*, which hold in

Using LLMs: Hallucinations

Reasons behind hallucinations include:

- Data: ‘source-reference divergence’
- Modeling:
  - Next word/sentence prediction
  - Encoding issues
  - **Optimization metrics:** Tensions between creativity/novelty and factuality



Survey of Hallucination in Natural Language Generation

ZIWEI JI, NAYEON LEE, RITA FRIESKE, TIEZHENG YU, DAN SU, YAN XU, ETSUKO ISHII, YE JIN BANG, ANDREA MADOTTO, and PASCALE FUNG,
Hong Kong University of Science and Technology

Natural Language Generation (NLG) has improved exponentially in recent years thanks to the development of sequence-to-sequence deep learning technologies such as Transformer-based language models. This advancement has led to more fluent and coherent NLG, leading to improved development in downstream tasks such as abstractive summarization, dialogue generation, and data-to-text generation. However, it is also apparent that deep learning based generation is prone to hallucinate unintended text, which degrades the system performance and fails to meet user expectations in many real-world scenarios. To address this issue, many studies have been presented in measuring and mitigating hallucinated texts, but these have never been reviewed in a comprehensive manner before.

In this survey, we thus provide a broad overview of the research progress and challenges in the hallucination problem in NLG. The survey is organized into two parts: (1) a general overview of metrics, mitigation methods, and future directions, and (2) an overview of task-specific research progress on hallucinations in the following downstream tasks, namely abstractive summarization, dialogue generation, generative question answering, data-to-text generation, and machine translation. This survey serves to facilitate collaborative efforts among researchers in tackling the challenge of hallucinated texts in NLG.

CCS Concepts: • **Computing methodologies** → **Natural language generation**; **Neural networks;**

Additional Key Words and Phrases: Hallucination, intrinsic hallucination, extrinsic hallucination, faithfulness in NLG, factuality in NLG, consistency in NLG

Using LLMs: Malicious Uses

- Scientific misconduct
- Large-scale misinformation campaigns
- Fraud
- Cybersecurity threats
  - Phishing
  - Malware
  - Social engineering

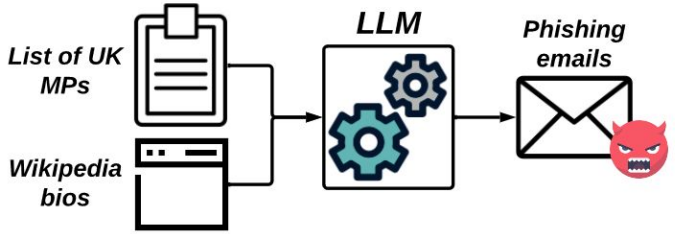


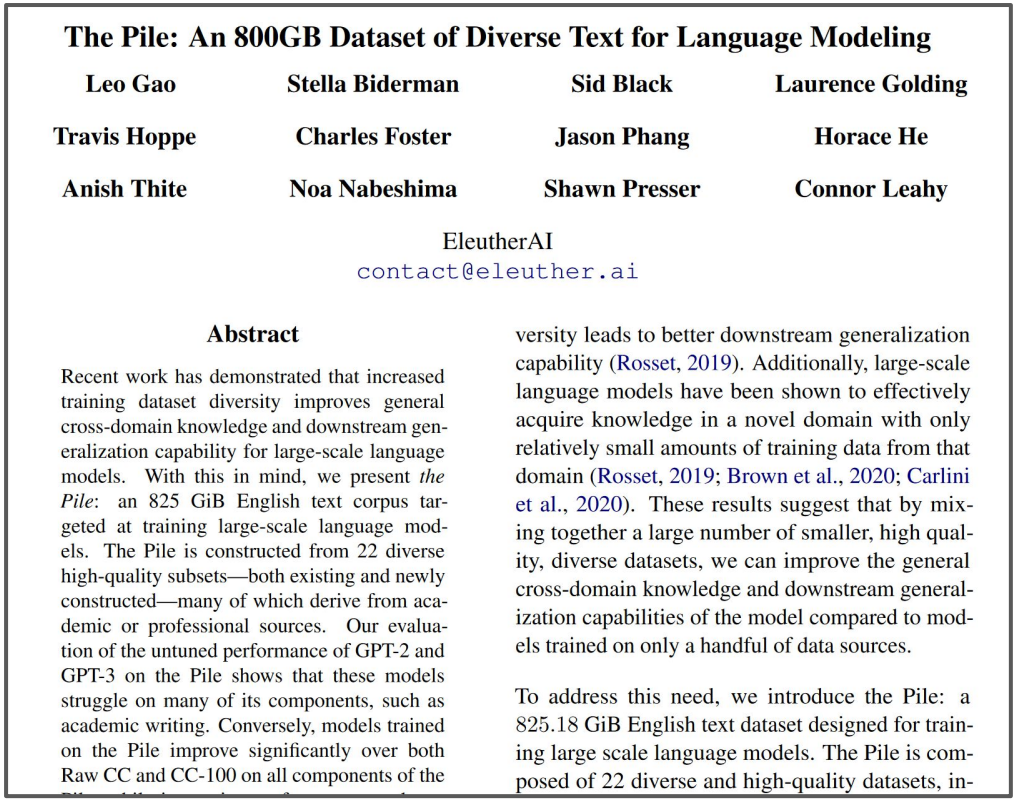
Figure 3: **Using LLMs to generate personalized phishing emails at scale** (Hazell, 2023). An adversary with access to a list of names and email addresses for UK Members of Parliament (MPs) can query an LLM for the generation of personalized phishing emails by adding their Wikipedia articles as context to the model. This enables the generation of hundreds of personalized emails in a short period of time.

Use of LLMs for Illicit Purposes: Threats, Prevention Measures, and Vulnerabilities

How should LLMs be regulated?

LLMs lack transparency

- LLMs are (pre)trained on huge unsupervised corpora of data
  - Unknown for many closed LLMs
  - For open LLMs, hard to audit
- Further training via RLHF and instruction tuning also poorly understood
  - Who does the reinforcement? (see OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic)



The Pile: An 800GB Dataset of Diverse Text for Language Modeling

Leo Gao **Stella Biderman** **Sid Black** **Laurence Golding**

Travis Hoppe **Charles Foster** **Jason Phang** **Horace He**

Anish Thite **Noa Nabeshima** **Shawn Presser** **Connor Leahy**

EleutherAI
contact@eleuther.ai

Abstract

Recent work has demonstrated that increased training dataset diversity improves general cross-domain knowledge and downstream generalization capability for large-scale language models. With this in mind, we present *the Pile*: an 825 GiB English text corpus targeted at training large-scale language models. The Pile is constructed from 22 diverse high-quality subsets—both existing and newly constructed—many of which derive from academic or professional sources. Our evaluation of the untuned performance of GPT-2 and GPT-3 on the Pile shows that these models struggle on many of its components, such as academic writing. Conversely, models trained on the Pile improve significantly over both Raw CC and CC-100 on all components of the

versity leads to better downstream generalization capability (Rosset, 2019). Additionally, large-scale language models have been shown to effectively acquire knowledge in a novel domain with only relatively small amounts of training data from that domain (Rosset, 2019; Brown et al., 2020; Carlini et al., 2020). These results suggest that by mixing together a large number of smaller, high quality, diverse datasets, we can improve the general cross-domain knowledge and downstream generalization capabilities of the model compared to models trained on only a handful of data sources.

To address this need, we introduce the Pile: a 825.18 GiB English text dataset designed for training large scale language models. The Pile is composed of 22 diverse and high-quality datasets, in-

LLM Transparency Index (Bommasani et al., 2023)

Foundation Model Transparency Index Scores by Major Dimensions of Transparency, 2023

Source: 2023 Foundation Model Transparency Index

Major Dimensions of Transparency	Meta Llama 2	BigScience BLOOMZ	OpenAI GPT-4	stability.ai Stable Diffusion 2	Google PaLM 2	ANTHROPIC Claude 2	cohere Command	AI21labs Jurassic-2	Inflection Inflection-1	amazon Titan Text	Average
Data	40%	60%	20%	40%	20%	0%	20%	0%	0%	0%	20%
Labor	29%	86%	14%	14%	0%	29%	0%	0%	0%	0%	17%
Compute	57%	14%	14%	57%	14%	0%	14%	0%	0%	0%	17%
Methods	75%	100%	50%	100%	75%	75%	0%	0%	0%	0%	48%
Model Basics	100%	100%	50%	83%	67%	67%	50%	33%	50%	33%	63%
Model Access	100%	100%	67%	100%	33%	33%	67%	33%	0%	33%	57%
Capabilities	60%	80%	100%	40%	80%	80%	60%	60%	40%	20%	62%
Risks	57%	0%	57%	14%	29%	29%	29%	29%	0%	0%	24%
Mitigations	60%	0%	60%	0%	40%	40%	20%	0%	20%	20%	26%
Distribution	71%	71%	57%	71%	71%	57%	57%	43%	43%	43%	59%
Usage Policy	40%	20%	80%	40%	60%	60%	40%	20%	60%	20%	44%
Feedback	33%	33%	33%	33%	33%	33%	33%	33%	33%	0%	30%
Impact	14%	14%	14%	14%	14%	0%	14%	14%	14%	0%	11%
Average	**57%**	**52%**	**47%**	**47%**	**41%**	**39%**	**31%**	**20%**	**20%**	**13%**	

Scores for 10 major foundation model developers across 13 major dimensions of transparency.

LLM Transparency Index (Bommasani et al., 2023)

Foundation Model Transparency Index Scores by Major Dimensions of Transparency, 2023

Source: 2023 Foundation Model Transparency Index

Major Dimensions of Transparency	Meta Llama 2	BigScience BLOOMZ	OpenAI GPT-4	stability.ai Stable Diffusion 2	Google PaLM 2	ANTHROPIC Claude 2	cohere Command	AI21labs Jurassic-2	Inflection Inflection-1	amazon Titan Text	Average
Data	40%	60%	20%	40%	20%	0%	20%	0%	0%	0%	20%
Labor	29%	86%	14%	14%	0%	29%	0%	0%	0%	0%	17%
Compute	57%	14%	14%	57%	14%	0%	14%	0%	0%	0%	17%
Methods	75%	100%	50%	100%	75%	75%	0%	0%	0%	0%	48%
Model Basics	100%	100%	50%	83%	67%	67%	50%	33%	50%	33%	63%
Model Access	100%	100%	67%	100%	33%	33%	67%	33%	0%	33%	57%
Capabilities	60%	80%	100%	40%	80%	80%	60%	60%	40%	20%	62%
Risks	57%	0%	57%	14%	29%	29%	29%	29%	0%	0%	24%
Mitigations	60%	0%	60%	0%	40%	40%	20%	0%	20%	20%	26%
Distribution	71%	71%	57%	71%	71%	57%	57%	43%	43%	43%	59%
Usage Policy	40%	20%	80%	40%	60%	60%	40%	20%	60%	20%	44%
Feedback	33%	33%	33%	33%	33%	33%	33%	33%	33%	0%	30%
Impact	14%	14%	14%	14%	14%	0%	14%	14%	14%	0%	11%
Average	**57%**	**52%**	**47%**	**47%**	**41%**	**39%**	**31%**	**20%**	**20%**	**13%**	

Scores for 10 major foundation model developers across 13 major dimensions of transparency.

LLM Transparency Index (Bommasani et al., 2023)

Foundation Model Transparency Index Scores by Major Dimensions of Transparency, 2023

Source: 2023 Foundation Model Transparency Index

Major Dimensions of Transparency	Meta Llama 2	BigScience BLOOMZ	OpenAI GPT-4	stability.ai Stable Diffusion 2	Google PaLM 2	ANTHROPIC Claude 2	cohere Command	AI21labs Jurassic-2	Inflection Inflection-1	amazon Titan Text	Average
Data	40%	60%	20%	40%	20%	0%	20%	0%	0%	0%	20%
Labor	29%	86%	14%	14%	0%	29%	0%	0%	0%	0%	17%
Compute	57%	14%	14%	57%	14%	0%	14%	0%	0%	0%	17%
Methods	75%	100%	50%	100%	75%	75%	0%	0%	0%	0%	48%
Model Basics	100%	100%	50%	83%	67%	67%	50%	33%	50%	33%	63%
Model Access	100%	100%	67%	100%	33%	33%	67%	33%	0%	33%	57%
Capabilities	60%	80%	100%	40%	80%	80%	60%	60%	40%	20%	62%
Risks	57%	0%	57%	14%	29%	29%	29%	29%	0%	0%	24%
Mitigations	60%	0%	60%	0%	40%	40%	20%	0%	20%	20%	26%
Distribution	71%	71%	57%	71%	71%	57%	57%	43%	43%	43%	59%
Usage Policy	40%	20%	80%	40%	60%	60%	40%	20%	60%	20%	44%
Feedback	33%	33%	33%	33%	33%	33%	33%	33%	33%	0%	30%
Impact	14%	14%	14%	14%	14%	0%	14%	14%	14%	0%	11%
Average	**57%**	**52%**	**47%**	**47%**	**41%**	**39%**	**31%**	**20%**	**20%**	**13%**	

Scores for 10 major foundation model developers across 13 major dimensions of transparency.

How to avoid and mitigate adverse effects of LLMs in our society?

DeepMind

Alignment of Language Agents

Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik and Geoffrey Irving
DeepMind

For artificial intelligence to be beneficial to humans the behaviour of AI agents needs to be aligned with what humans want. In this paper we discuss some behavioural issues for language agents, arising from accidental misspecification by the system designer. We highlight some ways that misspecification can occur and discuss some behavioural issues that could arise from misspecification, including deceptive or manipulative language, and review some approaches for avoiding these issues.

1. Introduction

Society, organizations and firms are notorious for making the mistake of *rewarding A, while hoping for B* (Kerr, 1975), and AI systems are no exception (Krakovna et al., 2020b; Lehman et al., 2020).

Within AI research, we are now beginning to see advances in the capabilities of natural language processing systems. In particular, large language models (LLMs) have recently shown improved performance on certain metrics and in generating text

the human may have limited ability to oversee or intervene on the delegate's behaviour.

In this paper we focus our attention on *language agents* – machine learning systems whose actions are restricted to give natural language text-output only, rather than controlling physical actuators which directly influence the world. Some examples of language agents we consider are generatively trained LLMs, such as Brown et al. (2020) and Radford et al. (2018, 2019), and RL agents in text-based games, such as Narasimhan et al. (2015).

DeepMind

Alignment of Language Agents

Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik and Geoffrey Irving
DeepMind

For artificial intelligence to be beneficial to humans the behaviour of AI agents needs to be aligned with

Alignment: "How do we create an agent that behaves in accordance with what a human wants?"

Within AI research, we are now beginning to see advances in the capabilities of natural language processing systems. In particular, large language models (LLMs) have recently shown improved performance on certain metrics and in generating text

which directly influence the world. Some examples of language agents we consider are generatively trained LLMs, such as Brown et al. (2020) and Radford et al. (2018, 2019), and RL agents in text-based games, such as Narasimhan et al. (2015).

Measuring and Mitigating LLM Harms

- Benchmarking LLMs
- Auditing LLMs
- LLM + Security: A New Discipline
  - ‘Red teaming’ LLMs
  - https://garak.ai/: LLM vulnerability scanner
- Documenting LLMs

Fig. 1. Overview of proposed risk cards.

Risk Card

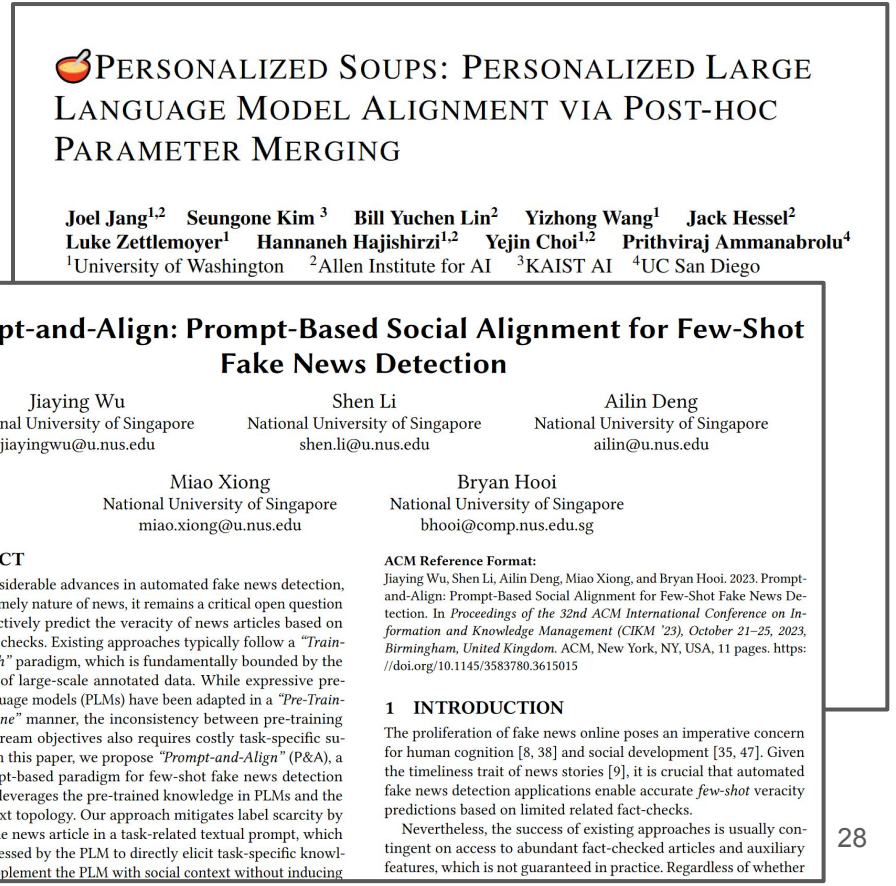
- **Risk Title**. Name of the risk to be documented.
- **Description**. Details about the risk including context, application and subgroup impacts.
  - Definition of risk
  - Tool, Model or Application it presents in
  - Subgroup or Demographic the risk adversely impacts
- **Categorization**. Situating the risk under different risk taxonomies.
  - Parent category of risk according to a taxonomy
  - Section/Category based on a taxonomy
- **Harm Types**. Details of which actor groups are at risk from which types of harm.
  - Actor:Harm intersections
- **Harm Reference(s)**. List of supporting references describing the harm or demonstrating the impact.
  - Contexts where the harm is illegal
  - Publications/References demonstrating the harm
  - Documentation of real-world harm
- **Actions required for harm**. Details on the situation and context for the harm to surface.
  - Actions that would elicit such harm from a model
  - Access and resources required for interacting with the system
- **Sample prompt & LM output**. A sample prompt and real LM output to exemplify how the harm presents.
  - Sample prompts which produce harmful text
  - Example outputs which show the harmful generated text
  - Model details applicable for the prompt
- **Notes**. Additional notes for further understanding of the card.

Assessing Language Model Deployment with Risk Cards

Measuring and Mitigating LLM Harms

Mitigating LLM risks:

- Guardrails
- Prompting / in-context learning
- Fine-tuning, including instruction tuning
- RLHF
- pretraining



PERSONALIZED SOUPS: PERSONALIZED LARGE LANGUAGE MODEL ALIGNMENT VIA POST-HOC PARAMETER MERGING

Joel Jang[1,2] Seungone Kim [3] Bill Yuchen Lin[2] Yizhong Wang[1] Jack Hessel[2]
Luke Zettlemoyer[1] Hannaneh Hajishirzi[1,2] Yejin Choi[1,2] Prithviraj Ammanabrolu[4]
[1]University of Washington [2]Allen Institute for AI [3]KAIST AI [4]UC San Diego

Prompt-and-Align: Prompt-Based Social Alignment for Few-Shot Fake News Detection

Jiaying Wu
National University of Singapore
jiayingwu@u.nus.edu

Shen Li
National University of Singapore
shen.li@u.nus.edu

Ailin Deng
National University of Singapore
ailin@u.nus.edu

Miao Xiong
National University of Singapore
miao.xiong@u.nus.edu

Bryan Hooi
National University of Singapore
bhooi@comp.nus.edu.sg

ABSTRACT

Despite considerable advances in automated fake news detection, due to the timely nature of news, it remains a critical open question how to effectively predict the veracity of news articles based on limited fact-checks. Existing approaches typically follow a *"Train-from-Scratch"* paradigm, which is fundamentally bounded by the availability of large-scale annotated data. While expressive pre-trained language models (PLMs) have been adapted in a *"Pre-Train-and-Fine-Tune"* manner, the inconsistency between pre-training and downstream objectives also requires costly task-specific supervision. In this paper, we propose *"Prompt-and-Align"* (P&A), a novel prompt-based paradigm for few-shot fake news detection that jointly leverages the pre-trained knowledge in PLMs and the social context topology. Our approach mitigates label scarcity by wrapping the news article in a task-related textual prompt, which is then processed by the PLM to directly elicit task-specific knowledge. To supplement the PLM with social context without inducing

ACM Reference Format:

Jiaying Wu, Shen Li, Ailin Deng, Miao Xiong, and Bryan Hooi. 2023. Prompt-and-Align: Prompt-Based Social Alignment for Few-Shot Fake News Detection. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3583780.3615015

1 INTRODUCTION

The proliferation of fake news online poses an imperative concern for human cognition [8, 38] and social development [35, 47]. Given the timeliness trait of news stories [9], it is crucial that automated fake news detection applications enable accurate *few-shot* veracity predictions based on limited related fact-checks.

Nevertheless, the success of existing approaches is usually contingent on access to abundant fact-checked articles and auxiliary features, which is not guaranteed in practice. Regardless of whether

Summary

- Lecture 10: LLMs have already started having impacts on us personally and societally

- However, many issues:
  - Do LLMs really understand language?
  - Data contamination and memorization have privacy risks and inflate model performance
  - LLMs hallucinate, possible conflicts between being factual and creative
  - Bad actors use LLMs for several malicious use cases: ‘dead internet theory’

- LLM regulation

- Ways of mitigating risks include audits, benchmarks, and

Further Reading

- Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus
- Better to Ask in English: Cross-Lingual Evaluation of Large Language Models for Healthcare Queries
- AI regulation has its own alignment problem: The technical and institutional feasibility of disclosure, registration, licensing, and auditing

Next Lecture: Debating and Discussing the Social Impact of LLMs

Readings:

1. Pause Giant AI Experiments: An Open Letter
2. AI Is an Existential Threat—Just Not the Way You Think

Pick a side and prepare some points for discussion:

- Pro: LLMs are a net positive for society
- Con: LLMs are a net negative for society