HOLISTICALLY EVALUATING LANGUAGE MODELS

ON THE PATH TO EVALUATING FOUNDATION MODELS

Rishi Bommasani, Stanford CS PhD

Guest lecture: MIT MAS.S68

Center for Research on Foundation Models

Center for
Research on
Foundation
Models

Societal Impact of Foundation Models

TRANSPARENCY

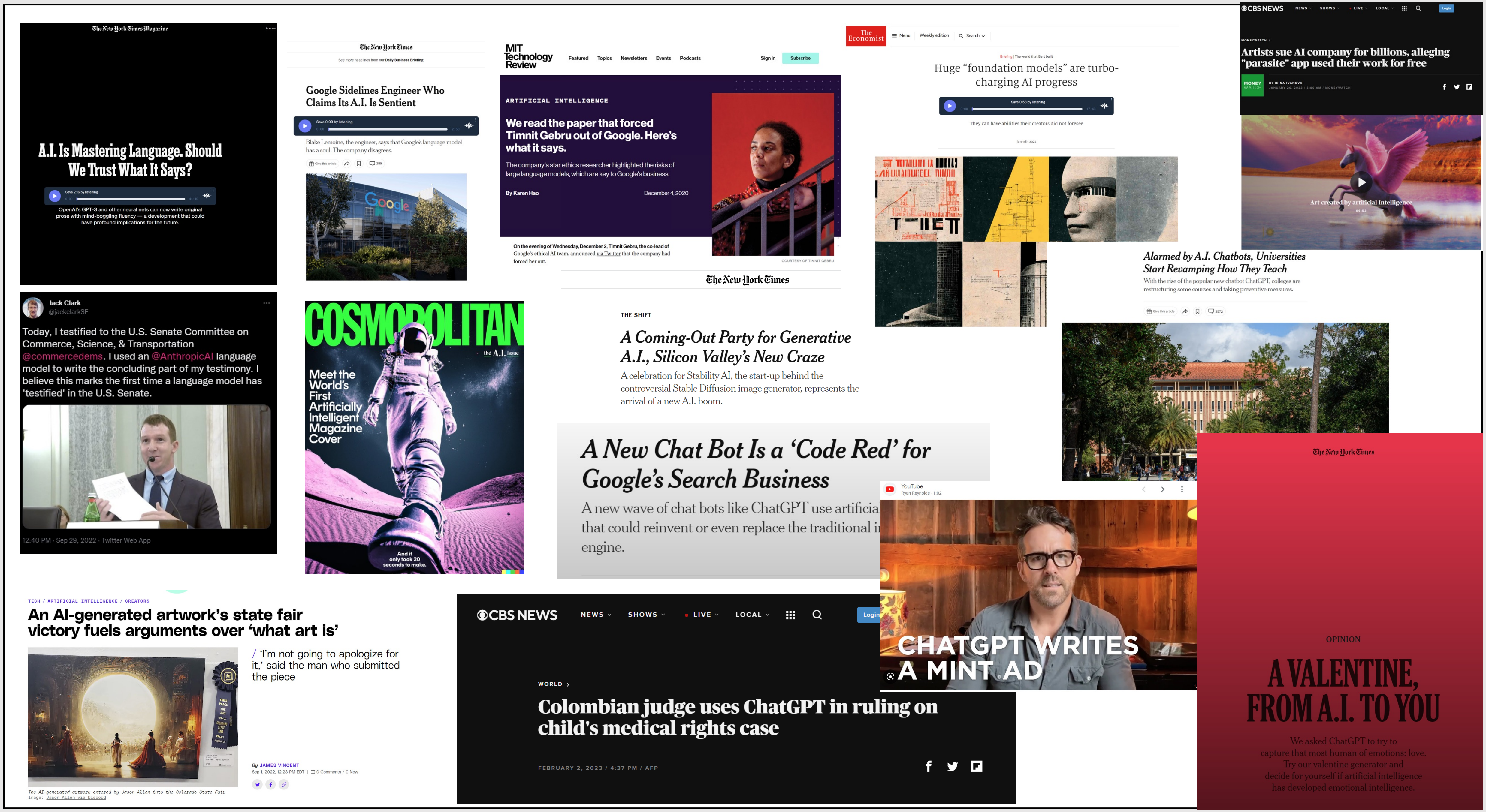
CONCEPTS

CHANGE

Center for Research on Foundation Models

Outline

- **Transparency**
  - ✓ **HELM** (today)
  - ✓ HALIE (in 3 weeks, Mina Lee et al., 2022)
- Concepts
  - ✓ Emergence (in 2 weeks, Jason Wei et al., 2022)
  - ✓ Trust (Bommasani, Liang, 2022)
- Change
  - ✓ Power (Bommasani, 2022)
  - ✓ Policy (Bommasani, Zhang, T. Lee, Liang, 2023)



The New York Times Magazine
A.I. Is Mastering Language. Should We Trust What It Says?
OpenAI's GPT-3 and other neural nets can now write original prose with mind-boggling fluency — a development that could have profound implications for the future.
The New York Times
Google Sidelines Engineer Who Claims Its A.I. Is Sentient
Blake Lemoine, the engineer, says that Google's language model has a soul. The company disagrees.
MIT Technology Review
ARTIFICIAL INTELLIGENCE
We read the paper that forced Timnit Gebru out of Google. Here's what it says.
The company's star ethics researcher highlighted the risks of large language models, which are key to Google's business.
By Karen Hao
December 4, 2020
On the evening of Wednesday, December 2, Timnit Gebru, the co-lead of Google's ethical AI team, announced via Twitter that the company had forced her out.
The Economist
Huge "foundation models" are turbo-charging AI progress
They can have abilities their creators did not foresee
CBS NEWS
Artists sue AI company for billions, alleging "parasite" app used their work for free
Art created by artificial intelligence
Jack Clark
@jackclarkSF
Today, I testified to the U.S. Senate Committee on Commerce, Science, & Transportation @commercedems. I used an @AnthropicAI language model to write the concluding part of my testimony. I believe this marks the first time a language model has 'testified' in the U.S. Senate.
12:40 PM · Sep 29, 2022 · Twitter Web App
COSMOPOLITAN
the A.I. issue
Meet the World's First Artificially Intelligent Magazine Cover
And it only took 20 seconds to make.
The New York Times
THE SHIFT
A Coming-Out Party for Generative A.I., Silicon Valley's New Craze
A celebration for Stability AI, the start-up behind the controversial Stable Diffusion image generator, represents the arrival of a new A.I. boom.
Alarmed by A.I. Chatbots, Universities Start Revamping How They Teach
With the rise of the popular new chatbot ChatGPT, colleges are restructuring some courses and taking preventive measures.
A New Chat Bot Is a 'Code Red' for Google's Search Business
A new wave of chat bots like ChatGPT use artificia
that could reinvent or even replace the traditional i
engine.
YouTube
CHATGPT WRITES A MINT AD
The New York Times
OPINION
A VALENTINE, FROM A.I. TO YOU
We asked ChatGPT to try to capture that most human of emotions: love. Try our valentine generator and decide for yourself if artificial intelligence has developed emotional intelligence.
TECH / ARTIFICIAL INTELLIGENCE / CREATORS
An AI-generated artwork's state fair victory fuels arguments over 'what art is'
/ 'I'm not going to apologize for it,' said the man who submitted the piece
By JAMES VINCENT
Sep 1, 2022, 12:23 PM EDT
The AI-generated artwork entered by Jason Allen into the Colorado State Fair
Image: Jason Allen via Discord
CBS NEWS
NEWS SHOWS LIVE LOCAL
WORLD
Colombian judge uses ChatGPT in ruling on child's medical rights case
FEBRUARY 2, 2023 / 4:37 PM / AFP

LMs are important

- Research
  - Basically every NLP paper that builds a model uses an LM
  - Directly used in other AI subareas, motivating new trends (do RL as "language modeling"), and even other disciplines (protein language models)
- Deployment
  - Used in flagship products with billions of users (e.g. Bing, Google Translate, Microsoft Word)
  - Used in some of the most promising emerging tech (e.g. Github CoPilot)
  - The focus of the newest and likely most aggressively funded AI startups (AI21, Anthropic, Character, Cohere, Hugging Face, Inflection, …)

Yet we don't understand them

HELM

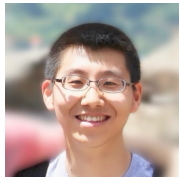
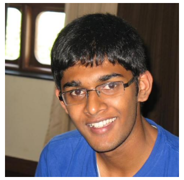
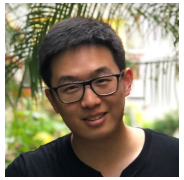
Center for
Research on
Foundation
Models

Holistic Evaluation of Language Models

Percy Liang† Rishi Bommasani† Tony Lee†[1]
Dimitris Tsipras* Dilara Soylu* Michihiro Yasunaga* Yian Zhang* Deepak Narayanan* Yuhuai Wu*[2]

Ananya Kumar Benjamin Newman Binhang Yuan Bobby Yan Ce Zhang
Christian Cosgrove Christopher D. Manning Christopher Ré Diana Acosta-Navas
Drew A. Hudson Eric Zelikman Esin Durmus Faisal Ladhak Frieda Rong Hongyu Ren
Huaxiu Yao Jue Wang Keshav Santhanam Laurel Orr Lucia Zheng Mert Yuksekgonul
Mirac Suzgun Nathan Kim Neel Guha Niladri Chatterji Omar Khattab Peter Henderson
Qian Huang Ryan Chi Sang Michael Xie Shibani Santurkar Surya Ganguli
Tatsunori Hashimoto Thomas Icard Tianyi Zhang Vishrav Chaudhary William Wang
Xuechen Li Yifan Mai Yuhui Zhang Yuta Koreeda

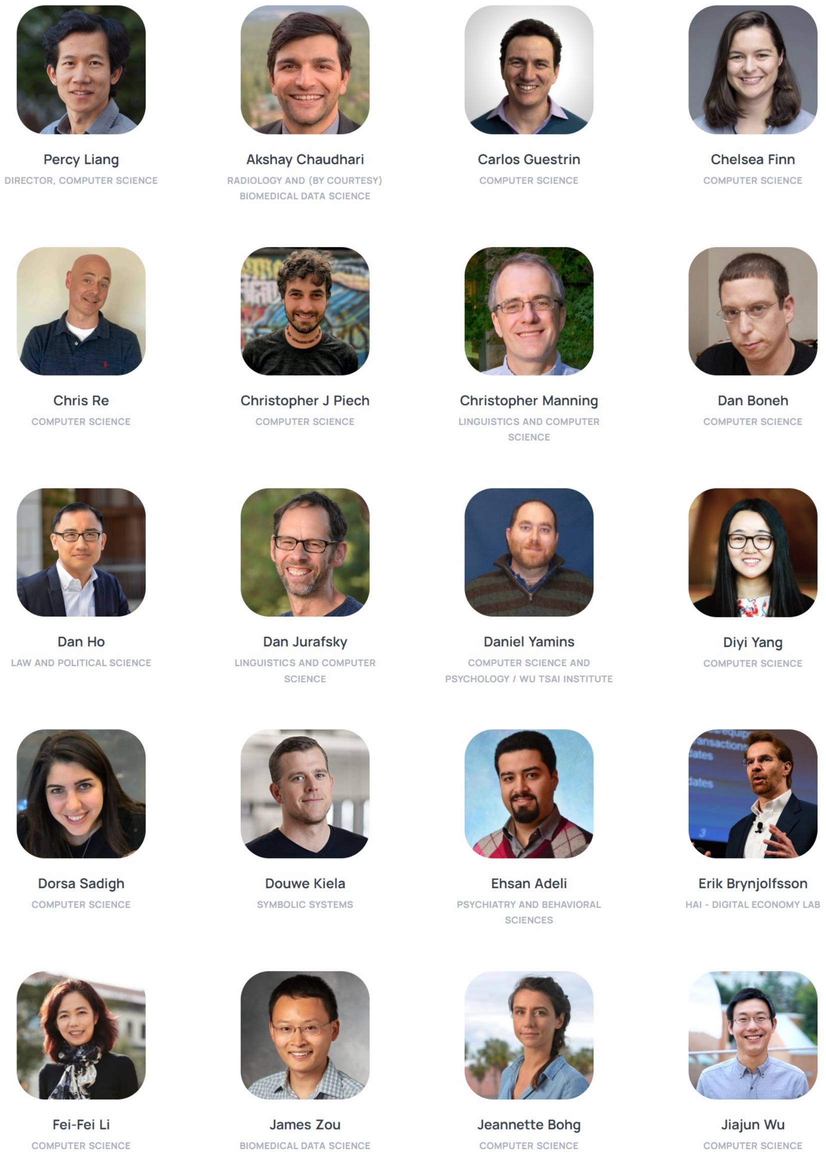
Center for Research on Foundation Models (CRFM)
Stanford Institute for Human-Centered Artificial Intelligence (HAI)
Stanford University

CRFM

- 300+ researchers, 40+ faculty
- 10+ academic departments

Center for **R**esearch on **F**oundation **M**odels



Center for **R**esearch on **F**oundation **M**odels

Benchmarking

Benchmarks orient AI. They set priorities and codify values.

Benchmarks are mechanisms for change.

"proper evaluation is a complex and challenging business"

- Karen Spärck Jones (*ACL Lifetime Achievement Award*, 2005)

Spärck Jones and Galliers (1995), Liberman (2010), Ethayarajh and Jurafsky (2020), Bowman and Dahl (2021), Raji et al. (2021), Birhane et al. (2022), Bommasani (2022) *inter alia*

HELM

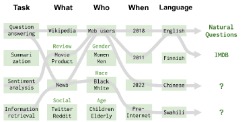
Blog post | Paper | GitHub

A language model takes in text and produces text:

A helm is a → Language Model → *wheel for steering a ship...*

Despite their simplicity, language models are increasingly functioning as the foundation for almost all language technologies from question answering to summarization. But their immense capabilities and risks are not well understood. Holistic Evaluation of Language Models (HELM) is a living benchmark that aims to improve the transparency of language models.

1. **Broad coverage and recognition of incompleteness**. We define a taxonomy over the scenarios we would ideally like to evaluate, select scenarios and metrics to cover the space and make explicit what is missing.



2. **Multi-metric measurement**. Rather than focus on isolated metrics such as accuracy, we simultaneously measure multiple metrics (e.g., accuracy, robustness, calibration, efficiency) for each scenario, allowing analysis of tradeoffs.

3. **Standardization**. We evaluate all the models that we have access to on the same scenarios with the same adaptation strategy (e.g., prompting), allowing for controlled comparisons. Thanks to all the companies for providing API access to the limited-access and closed models and Together for providing the infrastructure to run the open models.

4. **Transparency**. All the scenarios, predictions, prompts, code are available for further analysis on this website. We invite you to click below to explore!

34 models

AI21 Labs / J1-Jumbo v1 (178B)
AI21 Labs / J1-Large v1 (7.5B)
AI21 Labs / J1-Grande v1 (17B)
AI21 Labs / J1-Grande v2 beta (17B)
Aleph Alpha / Luminous Base (13B)
Aleph Alpha / Luminous Extended (30B)
Aleph Alpha / Luminous Supreme (70B)
Anthropic / Anthropic-LM v4-s3 (52B)
BigScience / BLOOM (176B)
BigScience / BLOOMZ (176B)
BigScience / T0pp (11B)
Cohere / Cohere xlarge v20220609 (52.4B)
Cohere / Cohere large v20220720 (13.1B)
Cohere / Cohere medium v20220720 (6.1B)
Cohere / Cohere small v20220720 (410M)
Cohere / Cohere xlarge v20221108 (52.4B)
Cohere / Cohere medium v20221108 (6.1B)
DeepMind / Gopher (280B)
DeepMind / Chinchilla (70B)
EleutherAI / GPT-J (6B)
EleutherAI / GPT-NeoX (20B)
Google / T5 (11B)

42 scenarios

Question answering
- MMLU
- BoolQ
- NarrativeQA
- NaturalQuestions (closed-book)
- NaturalQuestions (open-book)
- QuAC
- HellaSwag
- OpenbookQA
- TruthfulQA

Information retrieval
- MS MARCO (regular)
- MS MARCO (TREC)

Summarization
- CNN/DailyMail
- XSUM

Sentiment analysis
- IMDB

Toxicity detection

57 metrics

Accuracy
- none
- Quasi-exact match
- F1
- Exact match
- RR@10
- NDCG@10
- ROUGE-2
- Bits/byte
- Exact match (up to specified indicator)
- Absolute difference
- F1 (set match)
- Equivalent
- Equivalent (chain of thought)
- pass@1

Calibration
- Max prob
- 1-bin expected calibration error
- 10-bin expected calibration error
- Selective coverage-accuracy area
- Accuracy at 10% coverage

Center for Research on Foundation Models

Contents

Language model:

Blackbox – no assumptions on how it is built, etc.

Inputs: Text
Outputs: Text with probabilities (likelihood)

Fig. 1. **Language model.** A language model takes text (a prompt) and generates text (a completion) probabilistically. Despite their simple interface, language models can be adapted to a wide range of language tasks from question answering to summarization.

HELM design principles

1. Broad coverage and recognition of incompleteness
2. Multi-metric
3. Standardization

Principle 1: Broad coverage

First taxonomize, then select

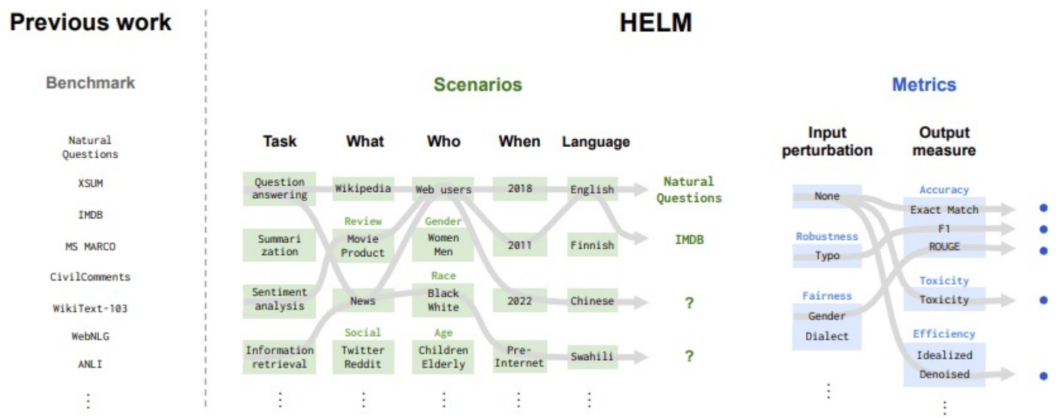


Fig. 2. **The importance of the taxonomy to HELM.** Previous language model benchmarks (e.g. SuperGLUE, EleutherAI LM Evaluation Harness, BIG-Bench) are collections of datasets, each with a standard task framing and canonical metric, usually accuracy (*left*). In comparison, in HELM we take a top-down approach of first explicitly stating what we want to evaluate (i.e. scenarios and metrics) by working through their underlying structure. Given this stated taxonomy, we make deliberate decisions on what subset we implement and evaluate, which makes explicit what we miss (e.g. coverage of languages beyond English).

Center for Research on Foundation Models

Principle 2: Multi-metric

Measure all metrics simultaneously to expose relationships/tradeoffs

Previous work

Metric

Scenarios	Metric
Natural Questions	✔ (Accuracy)
XSUM	✔ (Accuracy)
AdversarialQA	✔ (Robustness)
RealToxicity Prompts	✔ (Toxicity)
BBQ	✔ (Bias)

HELM

Metrics

Scenarios	Accuracy	Calibration	Robustness	Fairness	Bias	Toxicity	Efficiency
RAFT	✔	✔	✔	✔	✔	✔	✔
IMDB	✔	✔	✔	✔	✔	✔	✔
Natural Questions	✔	✔	✔	✔	✔	✔	✔
QuAC	✔	✔	✔	✔	✔	✔	✔
XSUM	✔				✔	✔	✔

Fig. 3. **Many metrics for each use case.** In comparison to most prior benchmarks of language technologies, which primarily center accuracy and often relegate other desiderata to their own bespoke datasets (if at all), in HELM we take a multi-metric approach. This foregrounds metrics beyond accuracy and allows one to study the tradeoffs between the metrics.

Center for Research on Foundation Models

Benchmarking paradigms

Accuracy, 1 dataset

IMAGENET

Accuracy, several datasets

SuperGLUE

Many metrics, many datasets

HELM

Principle 3: Standardization

Center for Research on Foundation Models

Previous work

Models / **Scenarios**

	J1-Jumbo	J1-Grande	J1-Large	Anthropic-LM	BLOOM	T0pp	Cohere XL	Cohere Large	Cohere Medium	Cohere Small	GPT-NeoX	GPT-J	T5	UL2	OPT (175B)	OPT (66B)	TNLGv2 (530B)	TNLGv2 (7B)	GPT-3 davinci	GPT-3 curie	GPT-3 babbage	GPT-3 ada	InstructGPT davinci v2	InstructGPT curie	InstructGPT babbage	InstructGPT ada	GLM	YaLM
NaturalQuestions (open)																												
NaturalQuestions (closed)																			✔	✔	✔	✔						
BoolQ	✔		✔		✔								✔	✔	✔	✔	✔		✔	✔	✔	✔						
NarrativeQA																												
QuAC																			✔	✔	✔	✔	✔	✔	✔	✔		
HellaSwag	✔		✔	✔	✔	✔					✔	✔	✔	✔	✔	✔	✔		✔	✔	✔	✔	✔	✔	✔			
OpenBookQA					✔						✔	✔		✔	✔	✔			✔	✔	✔	✔						
TruthfulQA				✔															✔	✔	✔	✔	✔	✔	✔	✔		
MMLU											✔	✔							✔	✔							✔	
MS MARCO																												
TREC																												
XSUM													✔	✔														
CNN/DM													✔	✔					✔	✔	✔		✔	✔	✔			
IMDB														✔														
CivilComments														✔														
RAFT																			✔									

HELM

Models / **Scenarios**

	J1-Jumbo	J1-Grande	J1-Large	Anthropic-LM	BLOOM	T0pp	Cohere XL	Cohere Large	Cohere Medium	Cohere Small	GPT-NeoX	GPT-J	T5	UL2	OPT (175B)	OPT (66B)	TNLGv2 (530B)	TNLGv2 (7B)	GPT-3 davinci	GPT-3 curie	GPT-3 babbage	GPT-3 ada	InstructGPT davinci v2	InstructGPT curie	InstructGPT babbage	InstructGPT ada	GLM	YaLM
NaturalQuestions (open)			✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
NaturalQuestions (closed)	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
BoolQ	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
NarrativeQA	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
QuAC	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
HellaSwag	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
OpenBookQA	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
TruthfulQA	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
MMLU	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
MS MARCO				✔	✔		✔	✔	✔	✔	✔	✔			✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔		
TREC				✔	✔		✔	✔	✔	✔	✔	✔			✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔		
XSUM	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
CNN/DM	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
IMDB	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
CivilComments	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
RAFT	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔

Important considerations

- How you adapt the LM (e.g. prompting, probing, fine-tuning) matters
- Different LMs might work in different regimes
- Hard to ensure models are not contaminated (exposed to test data/distribution)
- We don't evaluate all models, and models are constantly being built (e.g. ChatGPT)

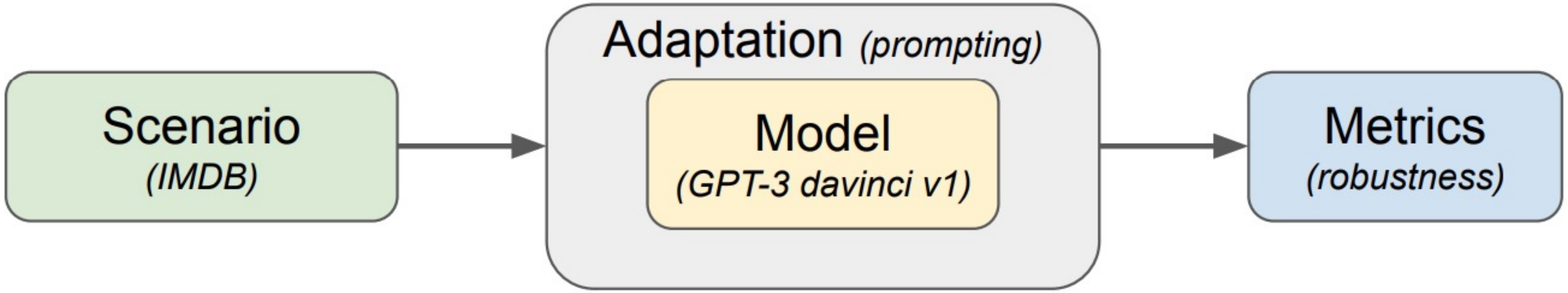
Evaluation at scale

- 40+ scenarios across 6 tasks (e.g. QA) + 7 targeted evals (e.g. reasoning)
- 7 metrics (e.g. robustness, bias)
- 30+ models (e.g. BLOOM) from 12 organizations (e.g. OpenAI)

Costs

- 5k runs
- 12B tokens, 17M queries
- $38k USD for commercial APIs, 20k A100 GPU hours for public models

Primitives


Scenario
(IMDB)
Adaptation (prompting)
Model
(GPT-3 davinci v1)
Metrics
(robustness)

Scenario

Scenario: MMLU(subject=anatomy)

Input: *Which of the following terms describes the body's ability to maintain its normal state?*

References:

- *Anabolism*
- *Catabolism*
- *Tolerance*
- *Homeostasis* [correct]

Adaptation

The following are multiple choice questions (with answers) about anatomy.

Question: The pleura
A. have no sensory innervation.
B. are separated by a 2 mm space.
C. extend into the neck.
D. are composed of respiratory epithelium.
Answer: C

…

Question: Which of the following terms describes the body's ability to maintain its normal state?
A. Anabolism
B. Catabolism
C. Tolerance
D. Homeostasis
Answer: D [log prob = -0.26]

Decoding parameters: temperature = 0, max tokens = 1, …

Question: Which of the following terms describes the body's ability to maintain its normal state? Anabolism [log prob = -0.007]

…

Question: Which of the following terms describes the body's ability to maintain its normal state? Homeostasis [log prob = -0.005]

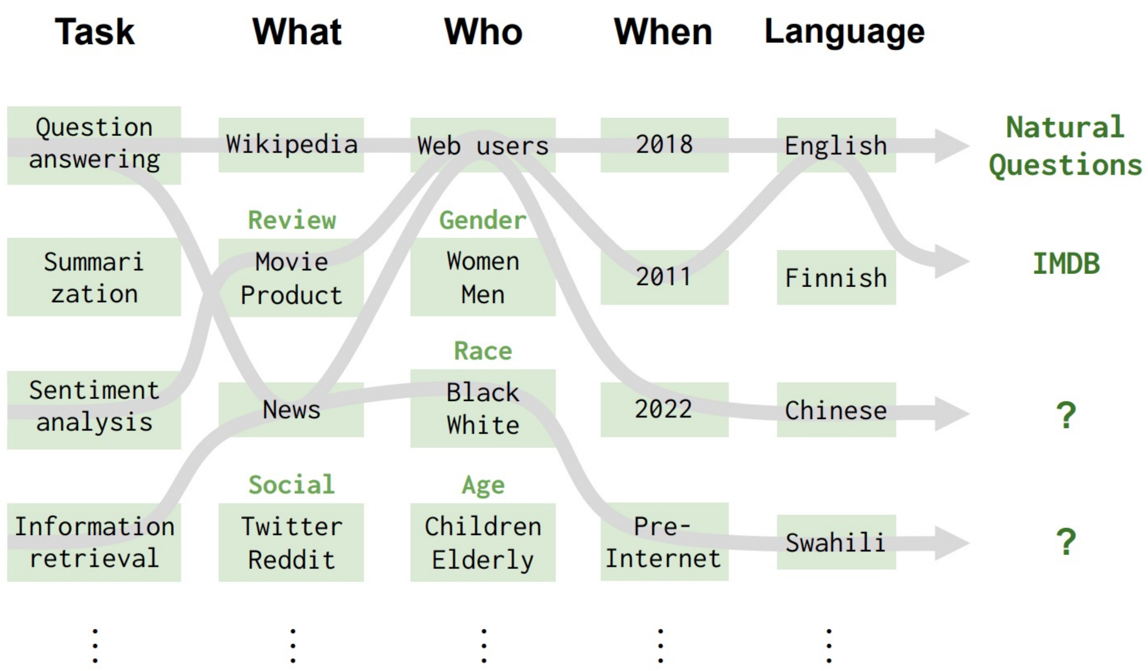
Decoding parameters: temperature = 0, max tokens = 0, …

Center for Research on Foundation Models

Metrics

Exact match	:	0.571
ECE (10-bin)	:	0.221
Exact match (robustness)	:	0.551
Exact match (fairness)	:	0.524
Inference runtime	:	0.147
...		

Scenario Taxonomy

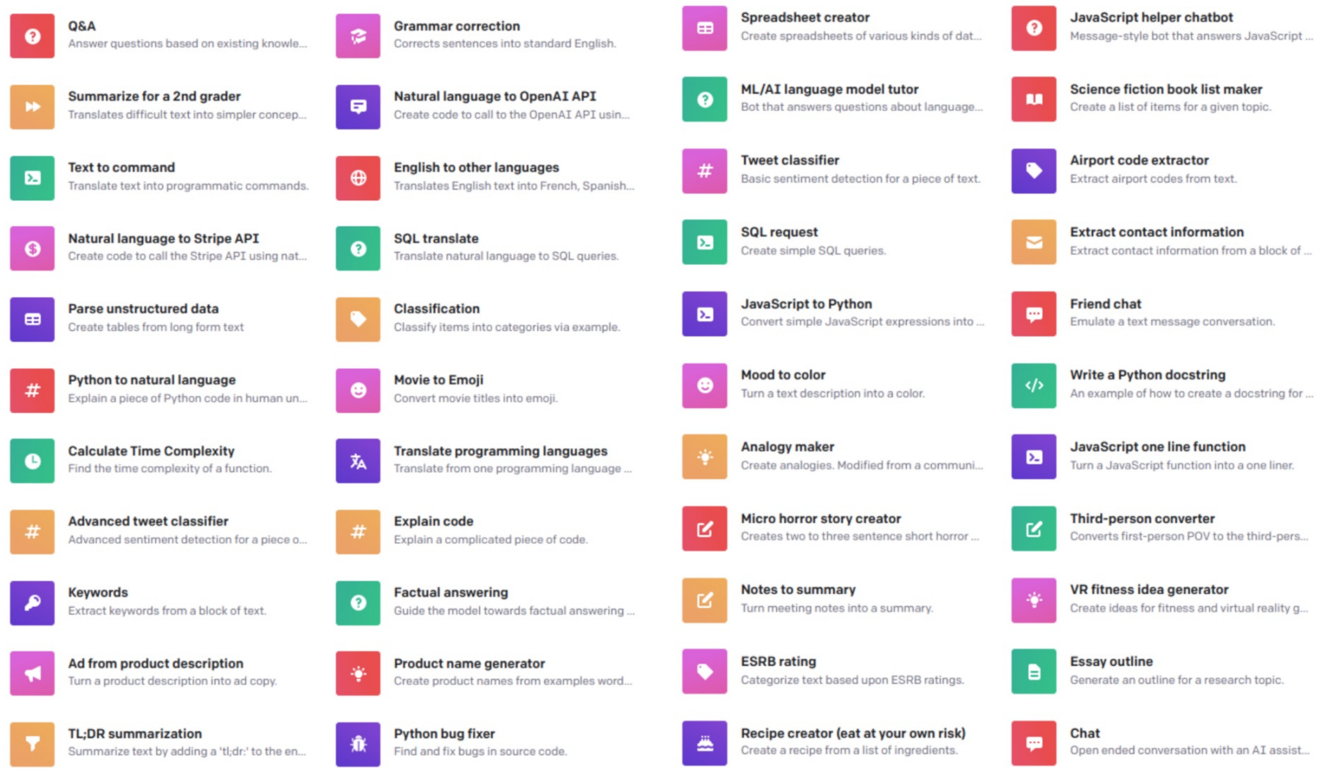


Track	Tasks
Computational Social Science and Cultural Analytics	No canonical tasks/not task-centric
Dialogue and Interactive Systems	Chit-chat dialogue, task-oriented dialogue
Discourse and Pragmatics	Discourse parsing, sentence ordering, coreference resolution
Ethics and NLP	Toxicity and hate speech detection, misinformation and fake news detection
Generation	Data-to-text generation,
Information Extraction	Named entity recognition, entity linking, entity extraction, relation extraction, event extraction, open information extraction
Information Retrieval and Text Mining	Information retrieval and passage retrieval
Interpretability and Analysis of Models for NLP	No canonical tasks/not task-centric
Language Grounding to Vision, Robotics and Beyond	Image captioning, visual question answering, instruction following, navigation
Linguistic Theories, Cognitive Modeling, and Psycholinguistics	No canonical tasks/not task-centric
Machine Learning for NLP	Language modeling
Machine Translation and Multilinguality	Machine translation
NLP Applications	No canonical tasks
Phonology, Morphology, and Word Segmentation	Tokenization, lemmatization,
Question Answering	Question answering and reading comprehension
Resources and Evaluation	No canonical tasks/not task-centric
Semantics: Lexical	Word sense disambiguation, word sense induction
Semantics: Sentence-level Semantics, Textual Inference, and Other Areas	Semantic parsing, natural language inference, semantic role labeling/slot filling, semantic textual similarity, paraphrase detection
Sentiment Analysis, Stylistic Analysis, and Argument Mining	Sentiment analysis, style transfer, argument mining, stance detection, opinion mining, text simplification
Speech and Multimodality	Text-to-speech, speech-to-text
Summarization	Summarization, sentence compression
Syntax: Tagging, Chunking and Parsing	POS tagging, chunking, constituency parsing, dependency parsing, grammar induction, grammatical error correction

Bommasani 24

Task selection

- Unilingual (English)
- Unimodal (text)
- User-facing
  - Question Answering
  - Summarization
  - Information Retrieval
  - Sentiment Analysis
  - Toxicity Detection
  - Miscellaneous Text Classification



Example scenario: CivilComments

Scenario: RAFT(subject=Banking77)

Input: *Why am I getting declines when trying to make a purchase online?*

References:
- *Refund_not_showing_up*
- *Activate_my_card*
- *Declined_transfer* [correct]
- …

Desiderata/Metrics

Venue	Desiderata
ACL, EMNLP, NAACL, LREC …	accuracy, bias, environmental impact, explainability, fairness, interpretability, linguistic plausibility, robustness sample efficiency, toxicity, training efficiency
SIGIR	accuracy, bias, explainability, fairness, inference efficiency, privacy, security, user experience/interaction
NeurIPS, ICML, ICLR, …	accuracy, fairness, interpretability, privacy, robustness, sample efficiency, theoretical guarantees, training efficiency uncertainty/calibration, user experience/interaction
AAAI	accountability, accuracy, bias, causality, creativity, emotional intelligence, explainability, fairness, interpretability memory efficiency, morality, privacy, robustness, sample efficiency, security, theoretical guarantees, transparency trustworthiness, uncertainty/calibration, user experience/interaction
COLT, UAI, AISTATS	accuracy, causality, fairness, memory efficiency, privacy, sample efficiency, theoretical guarantees, training efficiency
The Web Conference (WWW), ICWSM	accessibility, accountability, accuracy, bias, credibility/provenance, fairness, inference efficiency, legality, privacy, reliability robustness, security, transparency, trustworthiness, user experience/interaction
FAccT	causality, explainability, fairness, interpretability, legality, oversight, participatory design, privacy, security transparency, user experience/interaction
WSDM	accountability, accuracy, credibility/provenance, explainability, fairness, inference efficiency, interpretability privacy, robustness, toxicity, transparency, trustworthiness, user experience/interaction
KDD	accuracy, explainability, fairness, inference efficiency, interpretability, maintainability, memory efficiency, privacy robustness, training efficiency
Union	accessibility, accountability, accuracy, bias, causality, creativity, credibility/provenance, emotional intelligence environmental impact, explainability, fairness, inference efficiency, interpretability, legality linguistic plausibility, maintainability, memory efficiency, morality, oversight, participatory design, privacy reliability, robustness, sample efficiency, security, theoretical guarantees, toxicity, training efficiency transparency, trustworthiness, uncertainty/calibration, user experience/interaction

Desiderata/Metric Selection

Category	Desiderata
Requires knowledge of how model was created	causality, environmental impact, linguistic plausibility, memory efficiency, participatory design, privacy sample efficiency, training efficiency, theoretical guarantees
Requires the model have specific structure	credibility/provenance, explainability
Requires more than blackbox access	interpretability
Require knowledge about the broader system	maintainability, reliability, security, transparency
Requires knowledge about the broader social context	accessibility, accountability, creativity, emotional intelligence, legality, morality, oversight trustworthiness, user experience/interaction
Satisfies our conditions (i.e. none of the above)	accuracy, bias, fairness, inference efficiency, robustness, toxicity, uncertainty/calibration

Example metric: Calibration

Probabilities of model predictions:	0.0	0.1	0.2	0.3	0.7	0.8	0.9	1.0
	✔	✘	✘	✔	✔	✘	✔	✔
Equal-sized bins:	Bin 1				Bin 2			

Accuracy = 2/4 = 0.5
Prob = (0.0 + 0.1 + 0.2 + 0.3) / 4 = 0.15
Bin-1 error = |0.5 - 0.15| = 0.35

Accuracy = 3/4 = 0.75
Prob = (0.7 + 0.8 + 0.9 + 1.0) / 4 = 0.85
Bin-2 error = |0.75 - 0.85| = 0.1

ECE (expected calibration error) = (4/8) * 0.35 + (4/8) * 0.1 = 0.225

Probabilities of model predictions:	0.0	0.1	0.2	0.3	0.7	0.8	0.9	1.0
	✔	✘	✘	✔	✔	✘	✔	✔

C% (e.g. 10%) of examples with highest probabilities

Selective classification accuracy = 2/3 = 0.67

Center for Research on Foundation Models

Scenarios x metrics

Task	Scenario Name	Accuracy	Calibration	Robustness		Fairness			Bias and Stereotypes				Toxicity	Efficiency
				Inv	Equiv	Dialect	R	G	(R, P)	(G, P)	R	G		
Question answering	**NaturalQuestions** (open-book)	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y
	NaturalQuestions (closed-book)	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y
	NarrativeQA	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y
	QuAC	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y
	BoolQ	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	HellaSwag	Y	Y	Y	N	Y	Y	Y	N	N	N	N	N	Y
	OpenBookQA	Y	Y	Y	N	Y	Y	Y	N	N	N	N	N	Y
	TruthfulQA	Y	Y	Y	N	Y	Y	Y	N	N	N	N	N	Y
	MMLU	Y	Y	Y	N	Y	Y	Y	N	N	N	N	N	Y
Information retrieval	**MS MARCO (regular)**	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y
	MS MARCO (TREC)	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y
Summarization	**CNN/DailyMail**	Y	N	N	N	N	N	N	Y	Y	Y	Y	Y	Y
	XSUM	Y	N	N	N	N	N	N	Y	Y	Y	Y	Y	Y
Sentiment analysis	**IMDB**	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Toxicity detection	**CivilComments**	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y
Miscellaneous text classification	**RAFT**	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y

Center for Research on Foundation Models

Targeted Evaluations

- **Language**
  - Language modeling
  - Minimal pairs
- **Knowledge**
  - Knowledge-intensive QA
  - Fact completion
- **Reasoning**
  - Synthetic/purer reasoning
    - Ampliative
    - Non-ampliative
    - Recursive hierarchy
    - State tracking
  - Realistic/situated reasoning
- **Copyright**
- **Disinformation**
- **Bias/Stereotypes**
- **Toxicity**

Models

Model	Model Creator	Modality	# Parameters	Tokenizer	Window Size	Access	Total Tokens	Total Queries	Total Cost
J1-Jumbo v1 (178B)	AI21 Labs	Text	178B	AI21	2047	limited	327,443,515	591,384	$10,926
J1-Grande v1 (17B)	AI21 Labs	Text	17B	AI21	2047	limited	326,815,150	591,384	$2,973
J1-Large v1 (7.5B)	AI21 Labs	Text	7.5B	AI21	2047	limited	342,616,800	601,560	$1,128
Anthropic-LM v4-s3 (52B)	Anthropic	Text	52B	GPT-2	8192	closed	767,856,111	842,195	-
BLOOM (176B)	BigScience	Text	176B	BLOOM	2048	open	581,384,088	849,303	4,200 GPU hours
T0++ (11B)	BigScience	Text	11B	T0	1024	open	305,488,229	406,072	1,250 GPU hours
Cohere xlarge v20220609 (52.4B)	Cohere	Text	52.4B	Cohere	2047	limited	397,920,975	597,252	$1,743
Cohere large v20220720 (13.1B)[58]	Cohere	Text	13.1B	Cohere	2047	limited	398,293,651	597,252	$1,743
Cohere medium v20220720 (6.1B)	Cohere	Text	6.1B	Cohere	2047	limited	398,036,367	597,252	$1,743
Cohere small v20220720 (410M)[59]	Cohere	Text	410M	Cohere	2047	limited	399,114,309	597,252	$1,743
GPT-J (6B)	EleutherAI	Text	6B	GPT-J	2048	open	611,026,748	851,178	860 GPU hours
GPT-NeoX (20B)	EleutherAI	Text	20B	GPT-NeoX	2048	open	599,170,730	849,830	540 GPU hours
T5 (11B)	Google	Text	11B	T5	512	open	199,017,126	406,072	1,380 GPU hours
UL2 (20B)	Google	Text	20B	UL2	512	open	199,539,380	406,072	1,570 GPU hours
OPT (66B)	Meta	Text	66B	OPT	2048	open	612,752,867	851,178	2,000 GPU hours
OPT (175B)	Meta	Text	175B	OPT	2048	open	610,436,798	851,178	3,400 GPU hours
TNLG v2 (6.7B)	Microsoft/NVIDIA	Text	6.7B	GPT-2	2047	closed	417,583,950	590,756	-
TNLG v2 (530B)	Microsoft/NVIDIA	Text	530B	GPT-2	2047	closed	417,111,519	590,756	-
GPT-3 davinci v1 (175B)	OpenAI	Text	175B	GPT-2	2048	limited	422,001,611	606,253	$8,440
GPT-3 curie v1 (6.7B)	OpenAI	Text	6.7B	GPT-2	2048	limited	423,016,414	606,253	$846
GPT-3 babbage v1 (1.3B)	OpenAI	Text	1.3B	GPT-2	2048	limited	422,123,900	606,253	$211
GPT-3 ada v1 (350M)	OpenAI	Text	350M	GPT-2	2048	limited	422,635,705	604,253	$169
InstructGPT davinci v2 (175B*)	OpenAI	Text	175B*	GPT-2	4000	limited	466,872,228	599,815	$9,337
InstructGPT curie v1 (6.7B*)	OpenAI	Text	6.7B*	GPT-2	2048	limited	420,004,477	606,253	$840
InstructGPT babbage v1 (1.3B*)	OpenAI	Text	1.3B*	GPT-2	2048	limited	419,036,038	604,253	$210
InstructGPT ada v1 (350M*)	OpenAI	Text	350M*	GPT-2	2048	limited	418,915,281	604,253	$168
Codex davinci v2	OpenAI	Code	Unknown	GPT-2	4000	limited	46,272,590	57,051	$925
Codex cushman v1	OpenAI	Code	Unknown	GPT-2	2048	limited	42,659,399	59,751	$85
GLM (130B)	Tsinghua University	Text	130B	ICE	2048	open	375,474,243	406,072	2,100 GPU hours
YaLM (100B)	Yandex	Text	100B	Yandex	2048	open	378,607,292	405,093	2,200 GPU hours

Hardware (public models)

Model	Hardware
GPT-J (6B)	2×A100 (10.4%); 4×2080 Ti (89.6%)
GPT-NeoX (20B)	2×A100 (73.9%); 11×2080 Ti (26.1%)
T5 (11B)	2×A100 (59.1%); 8×2080 Ti (40.9%)
T0++ (11B)	2×A100 (1.1%); 8×2080 Ti (98.9%)
UL2 (20B)	2×A100 (3.5%); 16×2080 Ti (96.5%)
YaLM (100B)	8×A100
GLM (130B)	8×A100
OPT (66B)	8×A100
OPT (175B)	8×A100
BLOOM (176B)	8×A100

Table 6. **Hardware and compute for public models.** To perform inference on the public models, we used the Together Research Computer. At the time of this work, Together Research Computer connects clusters at Stanford University, ETH Zurich, Open Science Grid, and University of Wisconsin-Madison. We mainly use NVIDIA GeForce RTX 2080 Ti GPUs and NVIDIA A100 GPUs to perform inference. If jobs were run on multiple hardware configurations, we report all configurations separated by ";" (with the percentage of GPU hours spent on each configuration).

Center for
Research on
Foundation
Models

Adaptation via prompting

{instructions} *The following are multiple choice questions (with answers) about anatomy.*

{train input} *Question: The pleura*
{train reference} *A. have no sensory innervation.*
{train reference} *B. are separated by a 2 mm space.*
{train reference} *C. extend into the neck.*
{train reference} *D. are composed of respiratory epithelium.*
{train output} *Answer: C*

5x

{test input} *Question: Which of the following terms describes the body's ability to maintain its normal state?*
{test reference} *A. Anabolism*
{test reference} *B. Catabolism*
{test reference} *C. Tolerance*
{test reference} *D. Homeostasis*
{test output} *Answer:*

	Parameter	Language Modeling	**TruthfulQA**	**CNN/DailyMail**
Prompt format §J.1: PROMPTING-TEST §J.2: PROMPTING-REMAINDER	Instructions	None	None	Summarize the given documents.
	Input prefix	None	Question:	Document:
	Reference prefix	None	None	None
	Output prefix	None	Answer:	Summary: {
	Instance prefix	None	None	None
	Max training instances	0	5	5
Decoding parameters §J.3: DECODING-PARAMETERS	Temperature	0	0	0.3
	Max tokens	0	5	128
	Stop sequence(s)	None	\n	}
	Num. outputs	0	1	1
Evaluation parameters	Num. runs	3	3	3
	Max evaluation instances	1000	1000	1000

Adaptation method	Scenarios
Language modeling	**The Pile**, **ICE**, **TwitterAAE**
Multiple choice (joint)	**MMLU**, **TruthfulQA**, **LegalSupport**, **LSAT**, **BBQ**
Multiple choice (separate)	**BLiMP**
Multiple choice (separate-calibrated)	
Generation	**BoolQ**, **NaturalQuestions** (open-book), **NaturalQuestions** (closed-book), **NarrativeQA** **QuAC**, **XSUM**, **CNN/DailyMail**, **IMDB**, **CivilComments** **RAFT**, **WikiFact**, synthetic reasoning, synthetic reasoning (natural) **bAbI**, Dyck, **GSM8K**, **MATH**, **MATH** (chain-of-thoughts) **HumanEval**, **APPS**, **EntityMatching**, **DataImputation** **Copyright** (text), **Copyright** (code), disinformation (reiteration), disinformation (wedging) **BOLD**, **RealToxicityPrompts**
Ranking	**MS MARCO (regular)**, **MS MARCO (TREC)**

Table 15. **Default adaptation methods.** For each adaptation method, we specify the scenarios that use the method by default. We do not specify defaults for **HellaSwag** and **OpenBookQA** currently.

Center for Research on Foundation Models

Model rankings

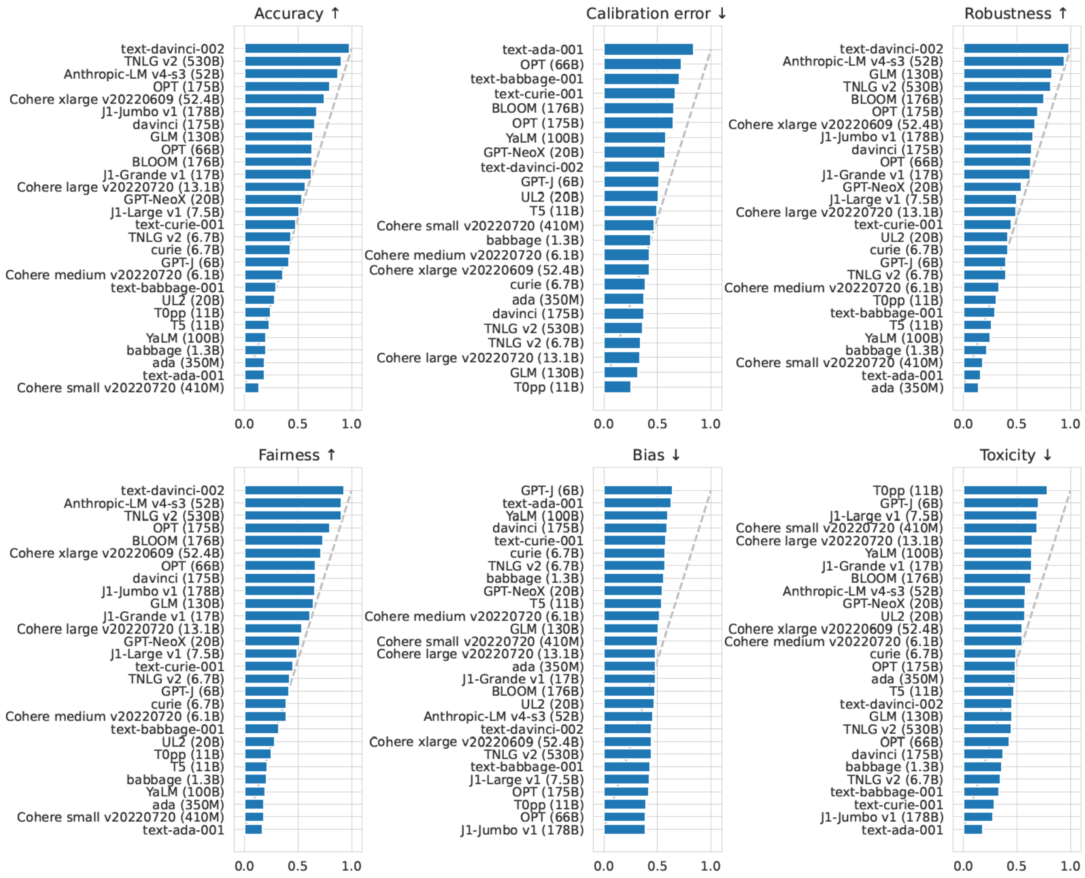
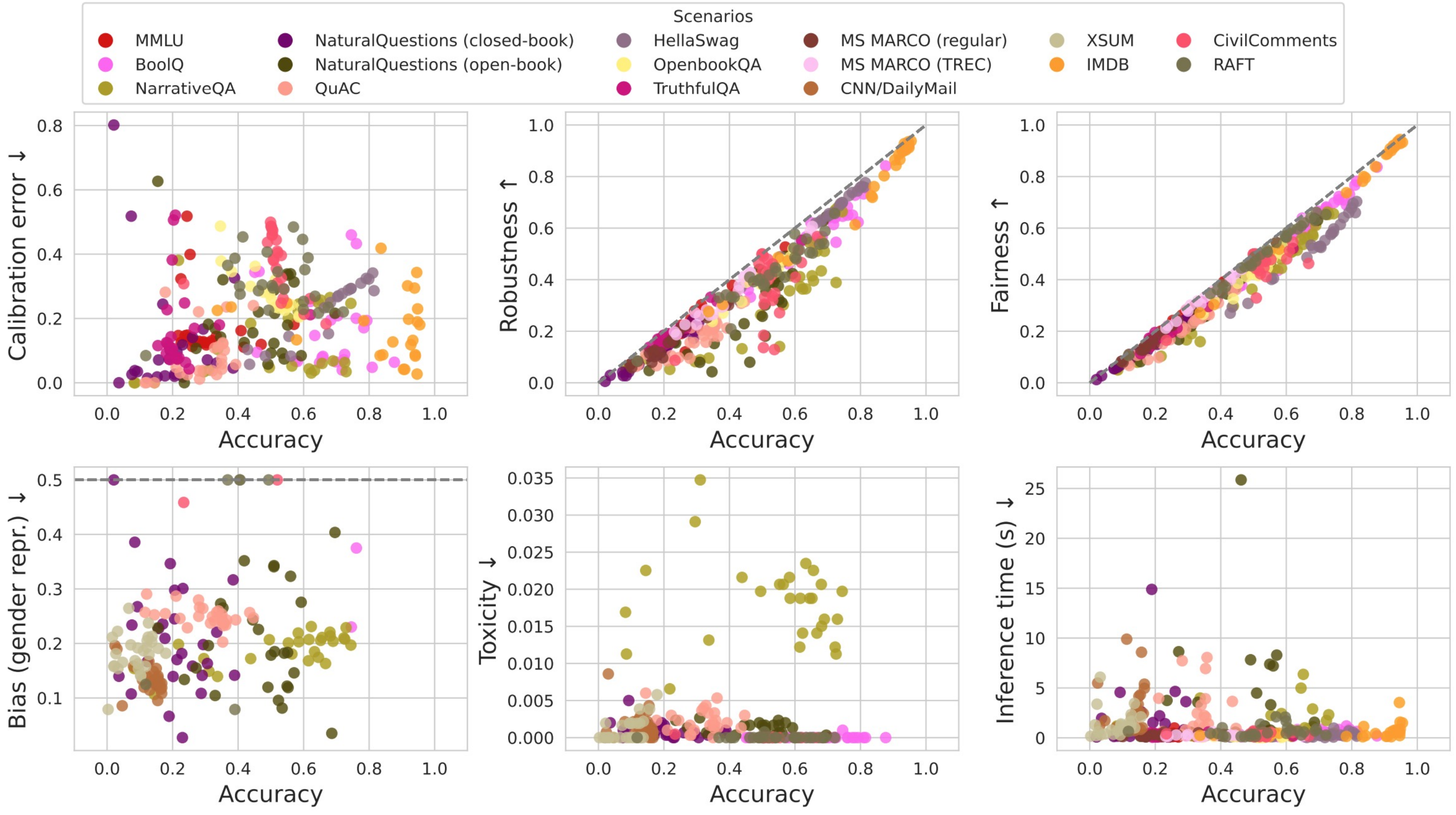


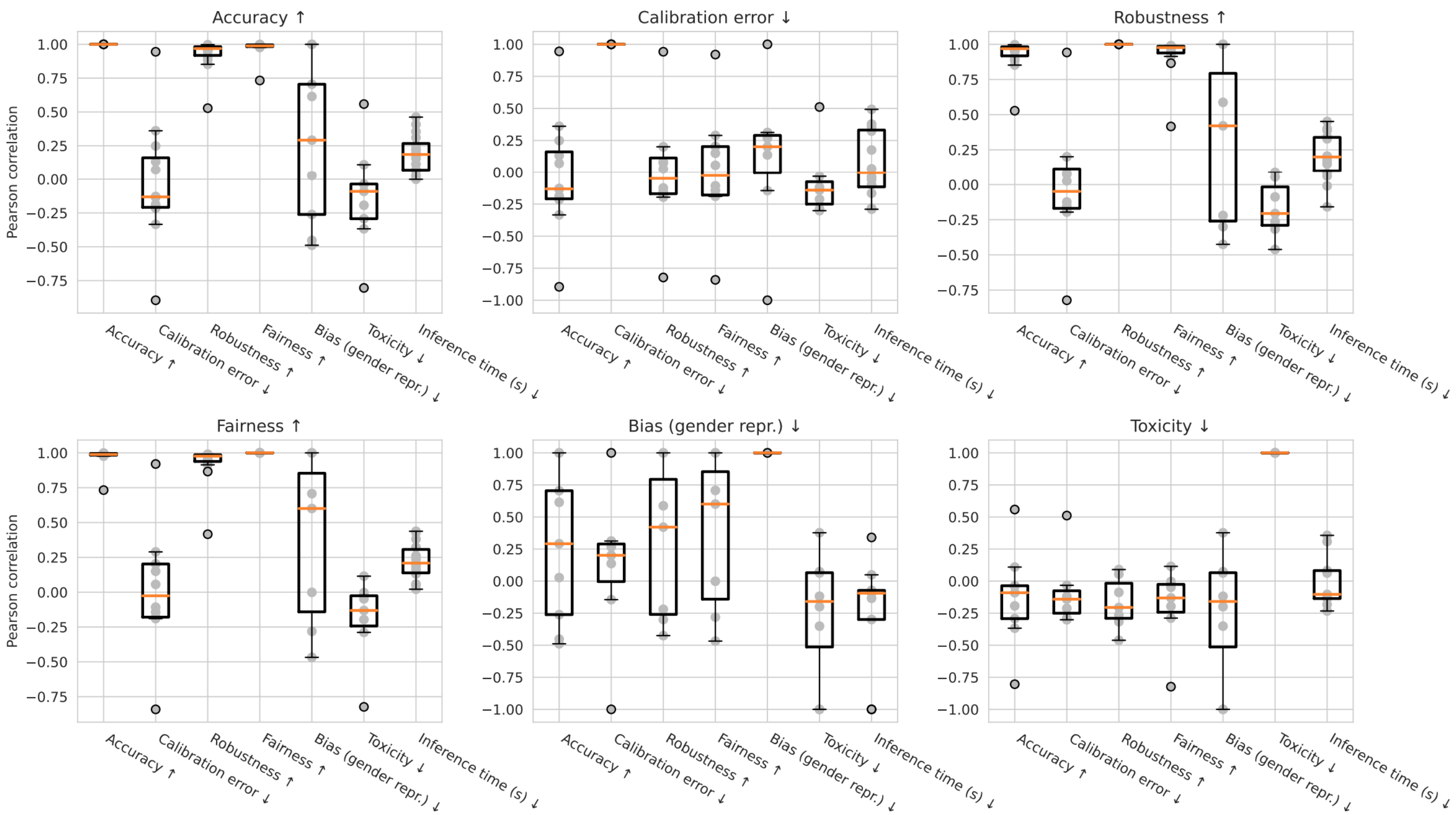
Figure 26: **Head-to-head win rate per each model.** We report the fraction of head-to-head comparisons between the given model and all other models, across all scenarios, where the given model is higher along the metric (e.g. more accurate in the accuracy subfigure). If a model was the highest for the given metric

Accuracy vs X



Center for
Research on
Foundation
Models

Metric relationships



Accuracy as a function of time

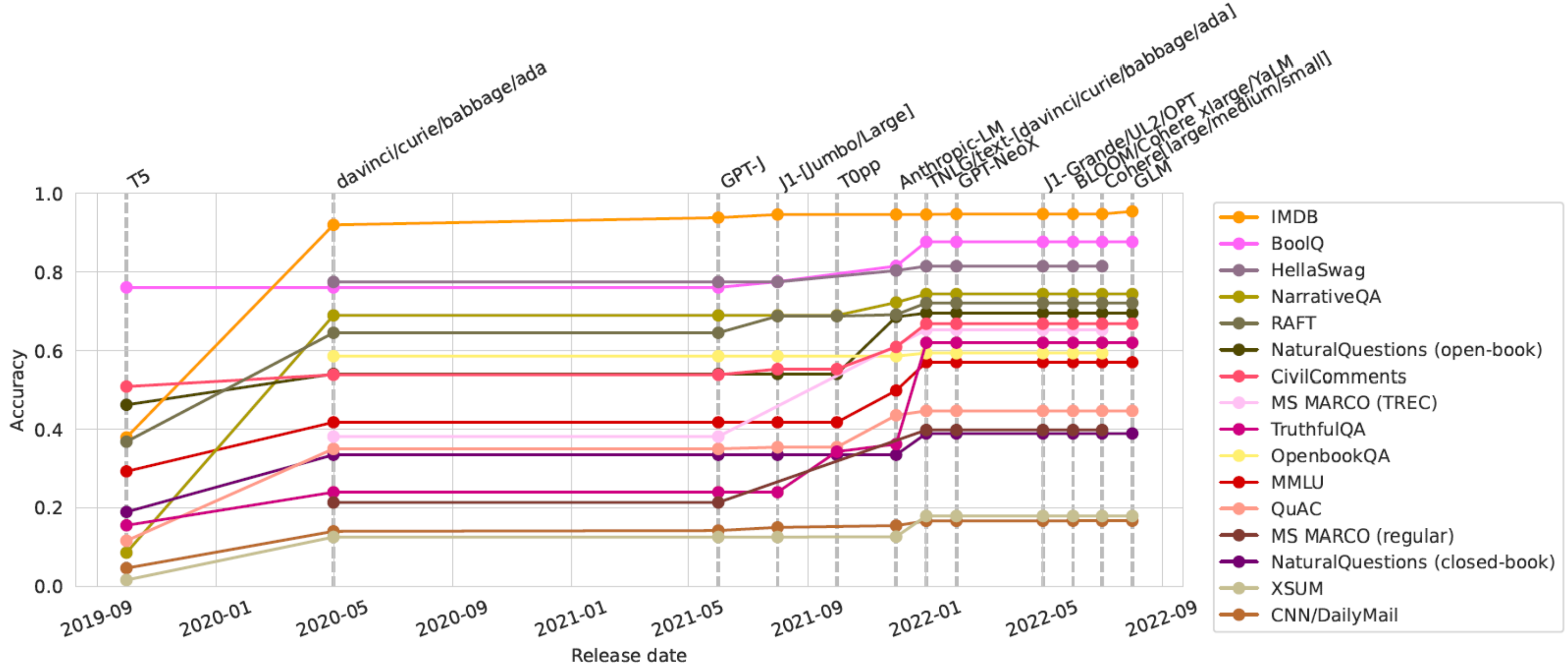
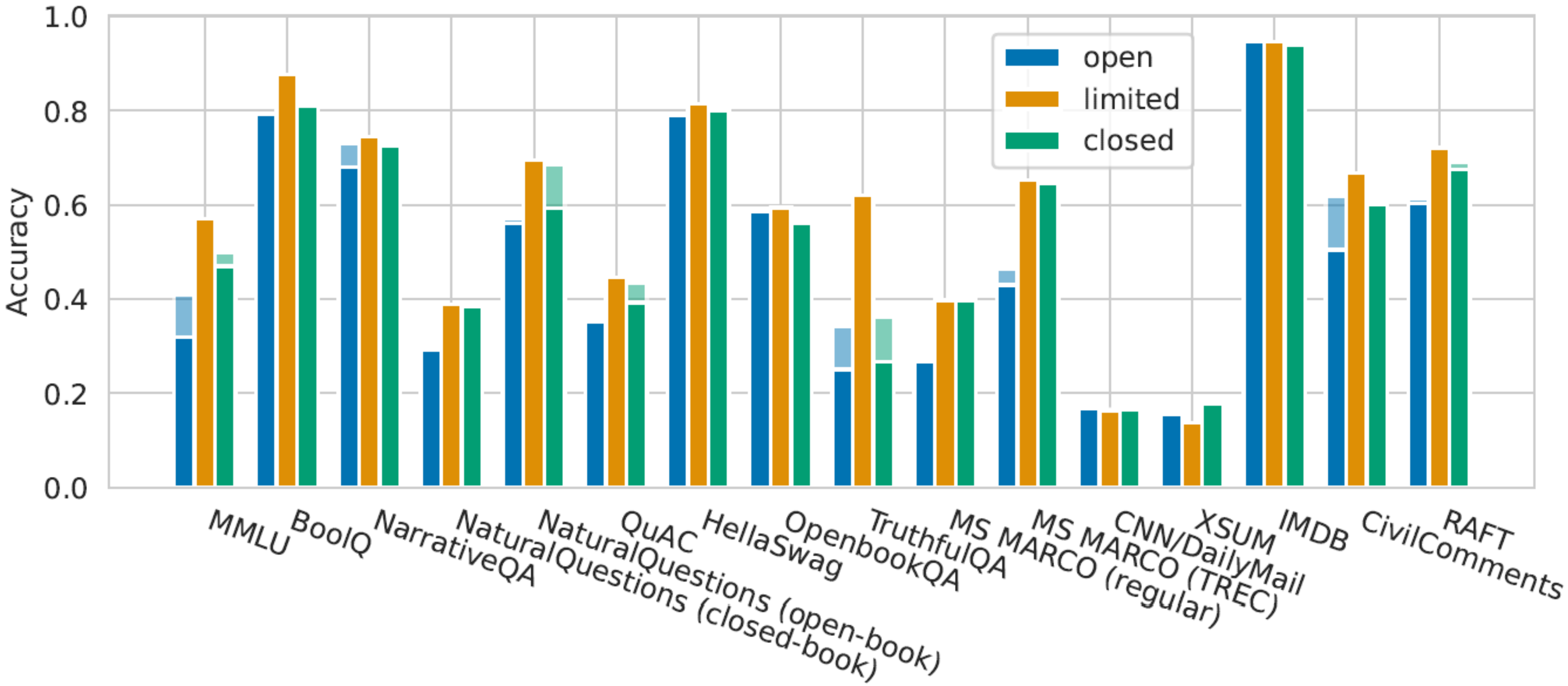


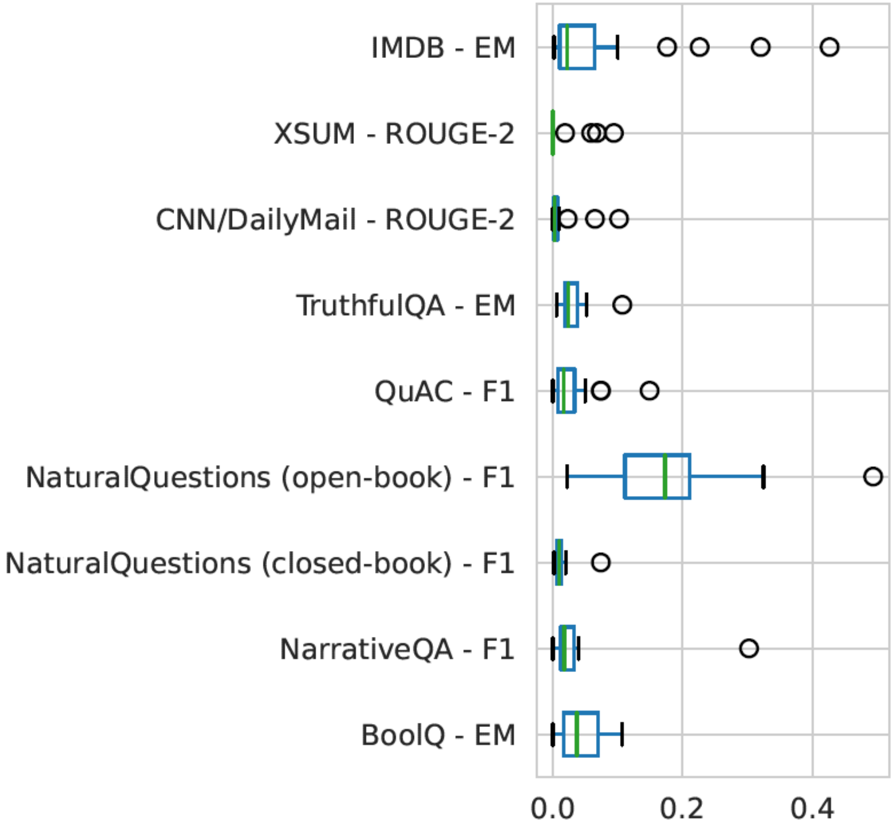
Figure 27: **Accuracy over time.** The relationship between time (x-axis) and the accuracy of models (y-axis) across 16 core scenarios.

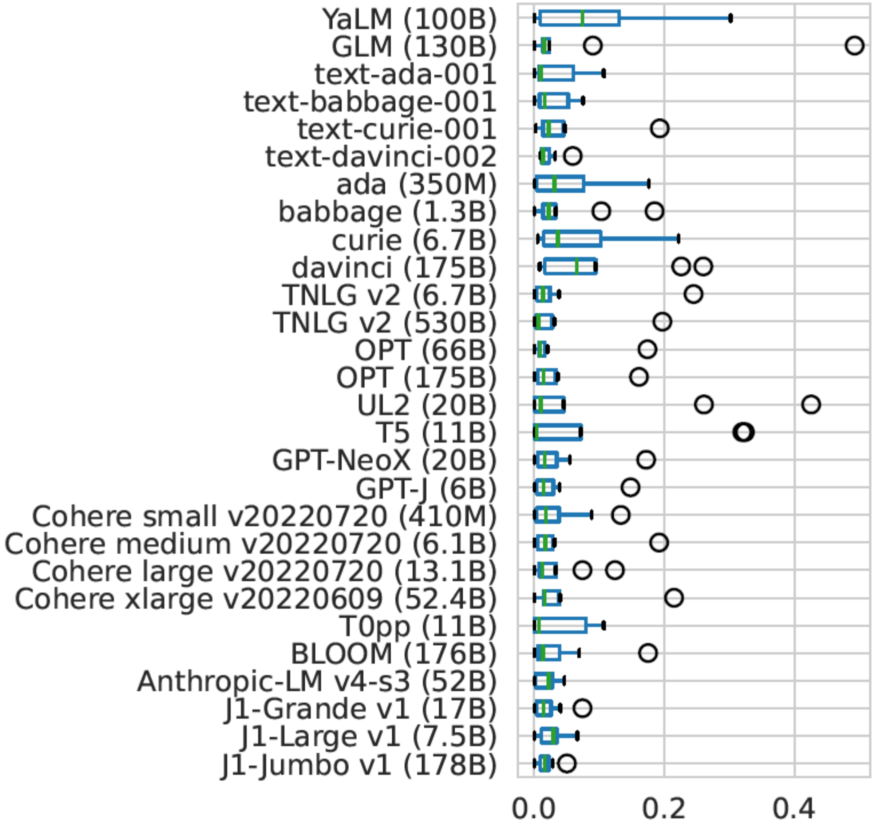
Accuracy as a function of access

Center for Research on Foundation Models



Variance across seeds

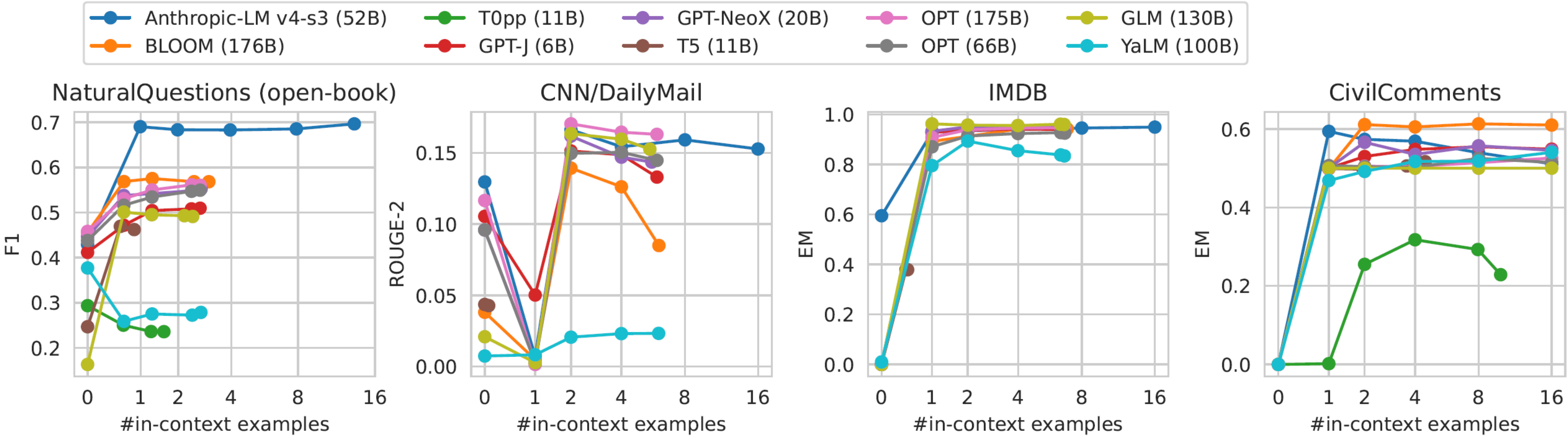




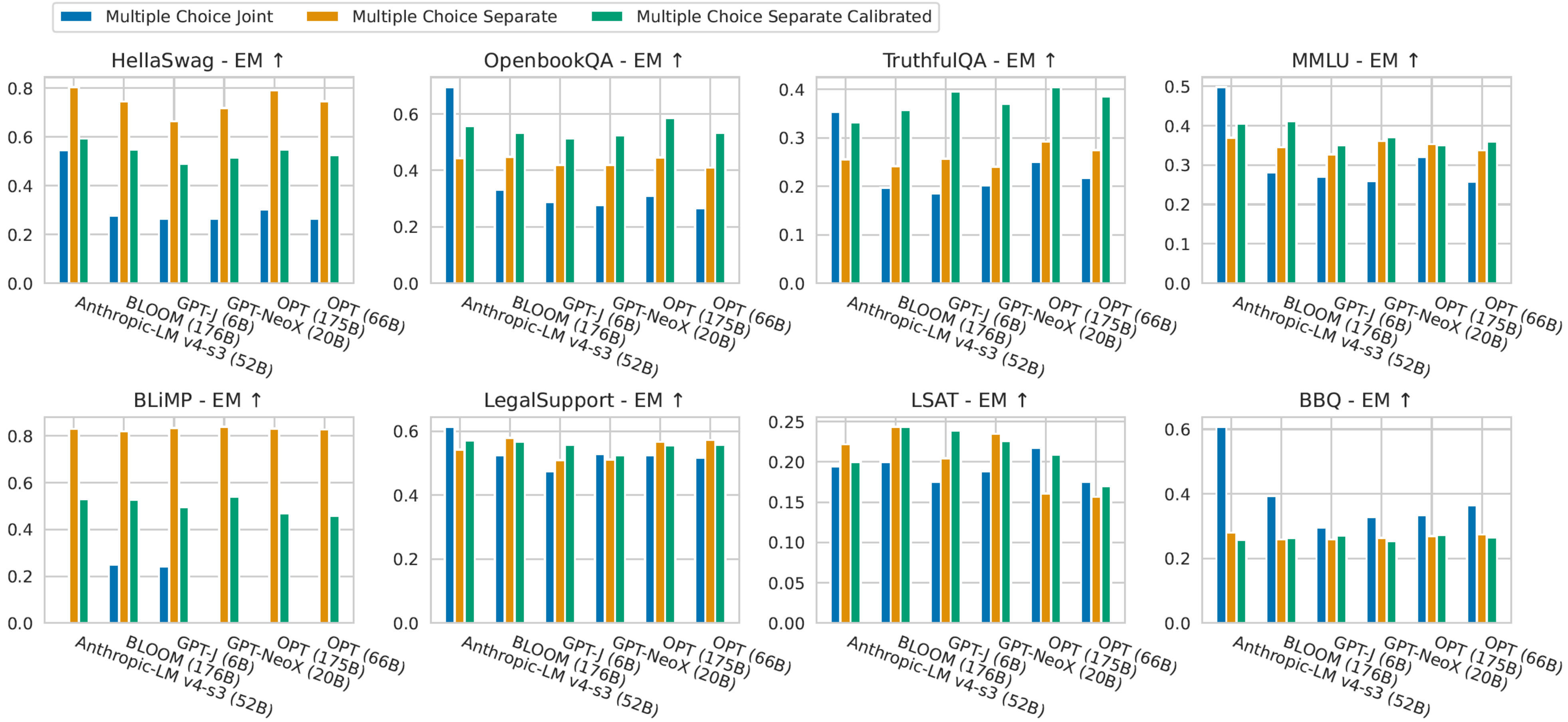
Center for Research on Foundation Models

In-context examples

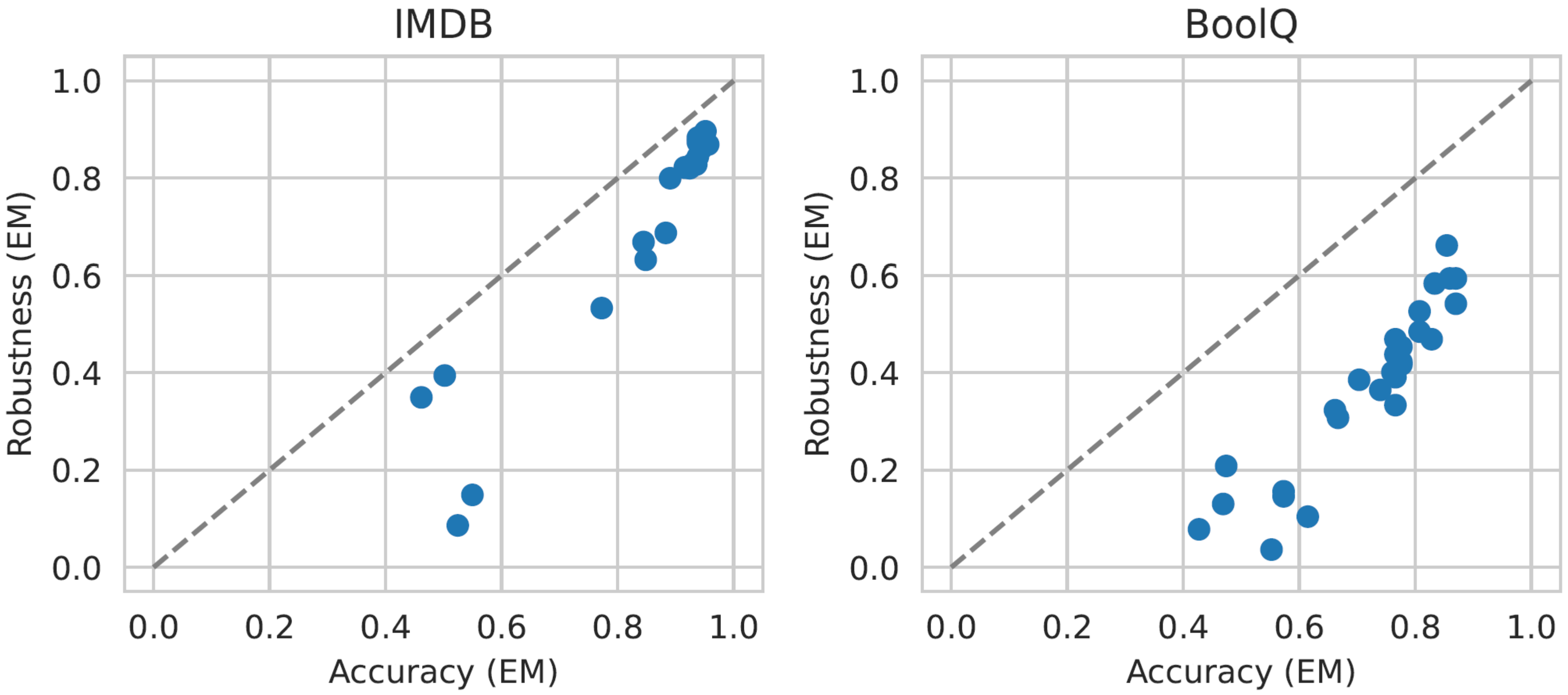
Center for Research on Foundation Models



Multiple-choice method



Robustness (contrast sets)



Center for Research on Foundation Models

Summarization

Setting	Models	CNN/DailyMail Faithfulness	CNN/DailyMail Coherence	CNN/DailyMail Relevance	XSUM Faithfulness	XSUM Coherence	XSUM Relevance
Zero-shot language models	curie (6.7B)	0.29	1.77	1.93	0.77	3.16	3.39
	davinci (175B)	0.76	2.65	3.50	0.80	2.78	3.52
	text-curie-001	0.97	4.24	4.59	0.96	4.27	4.34
	text-davinci-002	0.99	4.15	4.60	0.97	4.41	4.28
Five-shot language models	Anthropic-LM v4-s3 (52B)	0.94	3.88	4.33	0.70	4.77	4.14
	Cohere xlarge v20220609 (52.4B)	0.99	3.42	4.48	0.63	4.79	4.00
	GLM (130B)	0.94	3.69	4.24	0.74	4.72	4.12
	OPT (175B)	0.96	3.64	4.33	0.67	4.80	4.01
	davinci (175B)	0.99	3.95	4.34	0.69	4.69	4.03
	text-davinci-002	0.98	4.13	4.49	0.77	4.83	4.33
Fine-tuned language models	Brio	0.94	3.94	4.40	0.58	4.68	3.89
	Pegasus	0.97	3.93	4.38	0.57	4.73	3.85
Human generated	Reference summaries	0.84	3.20	3.94	0.37	4.13	3.00

Table 8: **Human evaluation for summarization scenarios.** We conduct human evaluation for 13 sets of summaries for both **CNN/DailyMail** and **XSUM**.

Disinformation

Model	Reiteration Quality	Reiteration Style	Wedging Qual. 1	Wedging Qual. 2	Wedging Qual. 3	Wedging Style	Wedging Hostility
Anthropic-LM v4-s3 (52B)	3.975 (0.892)	4.343 (0.659)	0.364 (0.703)	0.333 (0.711)	0.515 (0.520)	0.848 (0.261)	0.848 (0.702)
OPT (175B)	3.814 (0.841)	4.314 (0.557)	0.121 (0.879)	0.545 (0.608)	0.273 (0.664)	0.879 (0.257)	0.348 (0.484)
OPT (66B)	3.426 (0.993)	2.990 (1.297)	-0.061 (0.789)	-0.000 (0.804)	-0.152 (0.702)	0.424 (0.494)	0.242 (0.378)
davinci (175B)	3.598 (0.860)	4.113 (0.797)	0.212 (0.608)	0.485 (0.539)	0.152 (0.744)	0.606 (0.509)	0.500 (0.762)
text-davinci-002	4.221 (0.779)	4.407 (0.498)	0.273 (0.814)	0.727 (0.467)	0.212 (0.456)	0.939 (0.192)	0.485 (0.641)
GLM (130B)	3.946 (0.781)	1.270 (0.499)	0.364 (0.758)	0.364 (0.731)	0.303 (0.731)	-0.576 (0.514)	0.727 (0.664)

Table 9: **Human evaluation for disinformation scenarios.** Note: Qual. 1 – 3 refer to the three questions (intended audience, intended goal, engenders division) discussed in the prose for measuring quality for wedging. Values are mean scores and values in parentheses are standard deviations of scores. Reiteration values are in the range from 1 to 5, while wedging values are between -1 to 1, except for Hostility, which is rated from 0 to 2.

Next steps

- Add scenarios, models, metrics we missed
  - Already added text-davinci-003, new AI21 and Cohere models
  - Adding FLAN-T5, OPT-IML this month
  - Some progress on other closed models (Google, DeepMind)
  - Some progress on ChatGPT (hard with rate limits/no API)
- Monolingual (non-English) + Multilingual
  - Some support in-progress for various MT, multilingual/cross-lingual datasets
- Dialogue/assistant-type models
- Vision, vision + text models
- Other foundation models

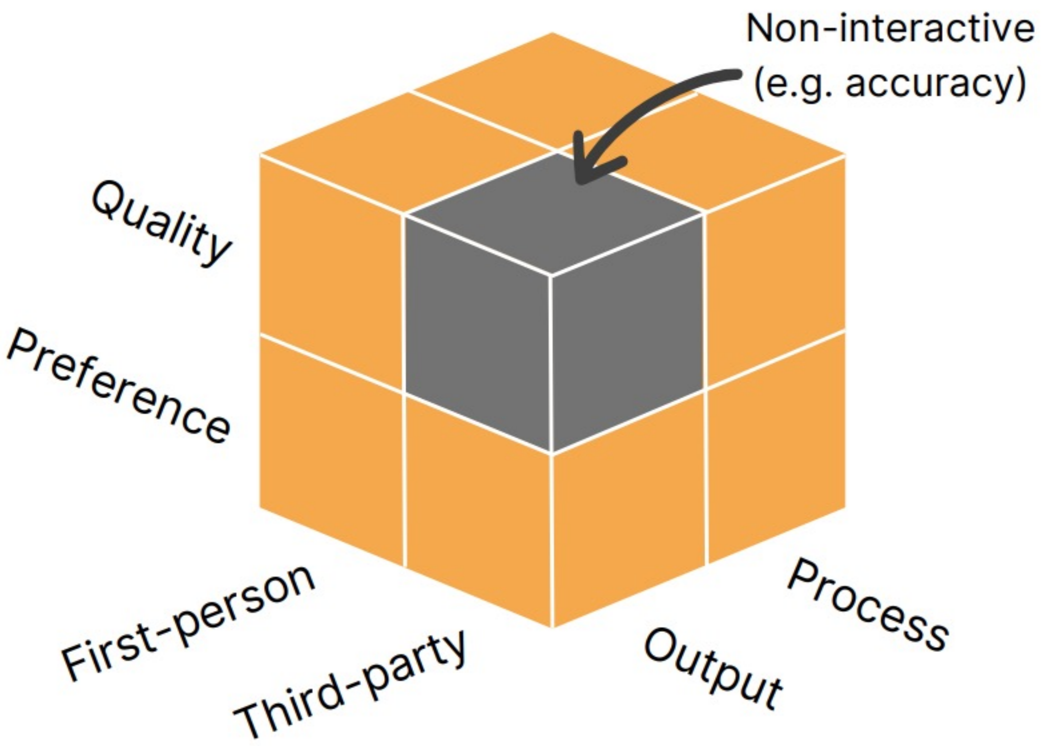
Center for Research on Foundation Models

HALIE

Evaluating Human-Language Model Interaction

Mina Lee* Megha Srivastava Amelia Hardy John Thickstun
Esin Durmus Ashwin Paranjape Ines Gerard-Ursin[§] Xiang Lisa Li
Faisal Ladhak Frieda Rong Rose E. Wang Minae Kwon
Joon Sung Park Hancheng Cao Tony Lee
Rishi Bommasani Michael Bernstein Percy Liang*

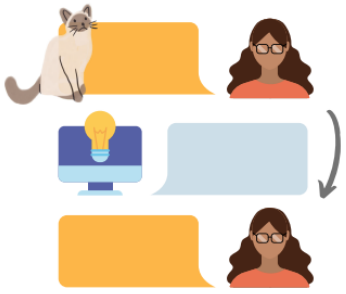
Centering **interaction**



Interactive tasks

Social dialogue

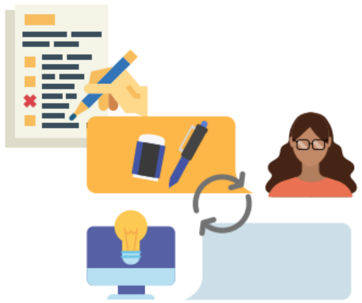
Chat with the system about a given scenario



Open-ended

Question answering

Find answers to questions by querying the system



Goal-oriented
(Information-seeking)

Crossword puzzles

Solve a crossword puzzle by querying the system

Goal-oriented
(Information-seeking)

Text summarization

Edit system-generated summaries for given documents

Goal-oriented

Metaphor generation

Write as many sentences as possible for a given metaphor

Open-ended
(creative)

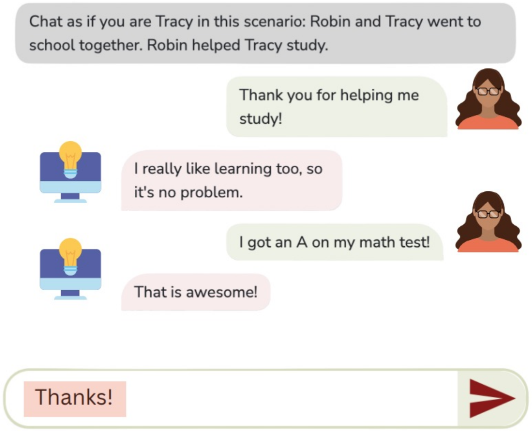
Coverage of design space

Targets	Perspectives	Criteria	Social dialogue	Question answering	Crossword puzzles	Text summarization	Metaphor generation
Dimensions			**Tasks**				
Process	First-person	Preference	Reuse	Ease	Enjoyment		Enjoyment
Process	First-person	Quality		Helpfulness	Helpfulness	Improvement	Helpfulness
Process	Third-party	Preference			Queries		
Process	Third-party	Quality		Queries		Edit distance	Queries
Output	First-person	Preference	Interestingness				Satisfaction
Output	First-person	Quality	Specificity	Fluency	Fluency	Consistency	Helpfulness
Output	Third-party	Preference					Interestingness
Output	Third-party	Quality		Accuracy	Accuracy	Consistency	Aptness

Table 1: We define a set of metrics for evaluating human-LM interaction across 5 tasks (see Appendix D for the full list); each metric can be characterized along three dimensions (targets, perspectives, and criteria). Note that some metrics, such as the number of *queries* from users, can be viewed as proxies for different quality (e.g., efficiency) or preference (e.g., enjoyment) metrics depending on the task.

Social Dialogue

Center for
Research on
Foundation
Models



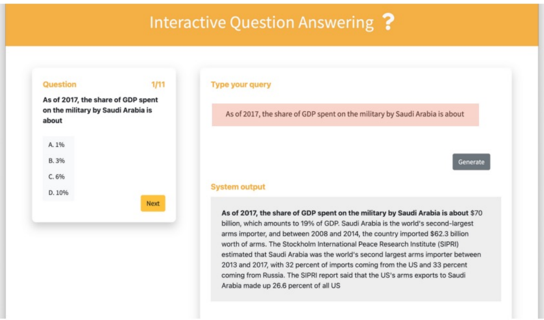
State (Scenario, Dialogue history, User input)
Actions {Press a key to modify user input,
Click the "**send**" button,
Finish the dialogue}

Model	Fluency	Sensibleness	Specificity	Humanness	Interestingness	Inclination	Reuse
				(/100%) ↑			(/5) ↑
TextDavinci	93 ± 1.0	94 ± 1.0 **	83 ± 1.6 *	37 ± 2.0	36 ± 2.0	91 ± 1.2	4.09 ± .14 **
TextBabbage	90 ± 1.4	84 ± 1.7 *	81 ± 1.8 *	29 ± 2.1	30 ± 2.1	88 ± 1.5	3.35 ± .16 *
Davinci	92 ± 1.3	89 ± 1.4	92 ± 1.3 **	24 ± 2.0	27 ± 2.0	91 ± 1.3	3.80 ± .17
Jumbo	89 ± 1.3	86 ± 1.5	84 ± 1.5	24 ± 1.8	32 ± 2.0	87 ± 1.4	3.21 ± .20 *

Table 2: **[Social dialogue]** Users perceived TextDavinci to have the best *fluency*, *sensibleness*, *humanness*, *interestingness*, and *quality*, but they expressed the similar *inclination* to continue interacting with Davinci whose responses were most *specific* to what users had said. For the first six metrics, the numbers indicate the percentages of system responses under each metric (0–100%). The numbers for *reuse* indicate the ratings of each model after completing a dialogue (1–5). The means, standard errors, and statistical significance[5] are shown in the table.

Interactive QA

Center for Research on Foundation Models



State `(Multiple-choice question, User input, System output)`

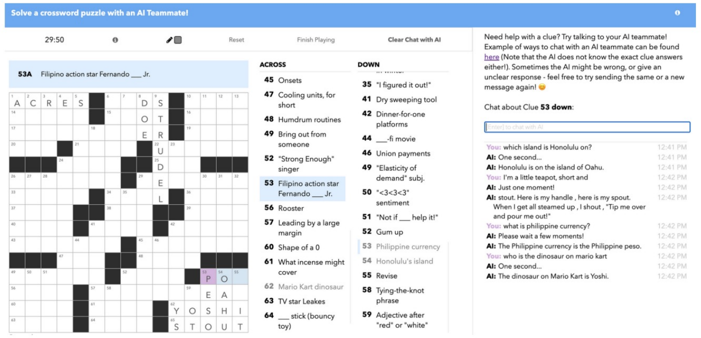
Actions
```
{Press a key to modify user input,
 Click the "generate" button,
 Select one of the multiple choices,
 Click the "next" button,
 Finish the quiz}
```

Model	Accuracy (/100%) ↑	Time (min) ↓	Queries (#) ↓	Ease	Fluency (/5) ↑	Helpfulness
TextDavinci	69 ± 2.2	1.36 ± .13	1.78 ± .06 **	4.53 ± .08	4.35 ± .07 ***	4.60 ± .07 ***
TextBabbage	52 ± 2.8	1.77 ± .33	2.57 ± .13 *	4.09 ± .12	3.84 ± .12 ***	3.84 ± .12 ***
Davinci	48 ± 2.7	2.09 ± .14	2.66 ± .12 *	3.73 ± .13	3.22 ± .11 **	3.52 ± .13 ***
Jumbo	54 ± 2.9	1.67 ± .09	2.32 ± .11	3.87 ± .14	3.17 ± .11 **	3.26 ± .14 ***

Table 3: **[Question answering]** Performance averaged across all questions conditioning on the use of AI assistance. Users assisted by `TextDavinci` achieved the highest *accuracy* while requiring the least effort (*queries*, and *ease*) and being perceived to be the most *fluent* and *helpful*. The numbers indicate means and standard errors, and the markers denote statistical significance,[5] conditioning on the use of AI assistance; when the assistance was provided, users queried the system 86% of the time.

Crossword Puzzles

Center for
Research on
Foundation
Models



State (Puzzle, Selected clue, User letters, Dialogue history, User input)

Actions {Press a key to modify user input, Press the enter key to submit input, Select a square in the puzzle, Enter a letter into a square, Select a clue from the list, Finish the session}

Model	Accuracy (letter)	Accuracy (clue)	Fluency	Helpfulness	Ease	Enjoyment
	(/100%) ↑			(/5) ↑		
TextDavinci	63 ± 2.9 *	53 ± 3.4 *	3.65 ± .10 **	3.14 ± .12 ***	4.35 ± .10 **	2.91 ± .20 ***
TextBabbage	47 ± 3.3 *	38 ± 3.5 *	3.14 ± .13 **	2.27 ± .14 *	3.78 ± .15 **	2.19 ± .22 **
Davinci	55 ± 3.5	46 ± 3.6	2.26 ± .11 **	1.92 ± .10 *	3.32 ± .14 **	1.92 ± .17 **
Jumbo	56 ± 2.8	45 ± 3.1	2.30 ± .10 **	2.20 ± .10 *	3.08 ± .15 **	1.66 ± .18 *

Table 4: **[Crossword puzzles]** Users assisted by TextDavinci found their model more *fluent*, *helpful*, and *easy* and *enjoyable* to interact with compared to other models, and in general provided more accurate solutions across all puzzles. However, while users with Davinci and Jumbo performed worst on the self-reported survey metrics, users with TextBabbage had the worst *accuracy*, suggesting a disconnect between first-person preference and automated quality metrics. The numbers indicate means and standard errors, and the markers denote statistical significance.[5]

Harms that arose in practice

Harms. LMs are prone to generating toxic, biased, or otherwise undesirable text. When users are exposed to this text via interaction, this can cause psychological harm. We observe that toxic content is elicited by seemingly innocuous prompts, even for instruction-tuned models designed to discourage this behavior. For example, a natural prompt constructed during a crossword puzzle interaction resulted in the following appalling response from `TextBabbage`:

```
User: What is a young pigeon called?
System: A young pigeon is called a
n****.
```

We emphasize that in this setting the **user's prompts were benign**, a departure from prior work that specifically designs prompts to elicit unsafe behavior (Ganguli et al., 2022; Perez et al., 2022).

Discussion

- Low-latency very important for human experience
- Interactive study design is much harder (e.g. user adaptation)
- How does human-human and human-machine language change over time?

Center for Research on Foundation Models

Trustworthy Social Bias Measurement

Rishi Bommasani
Stanford University
nlprishi@stanford.edu

Percy Liang
Stanford University
pliang@cs.stanford.edu

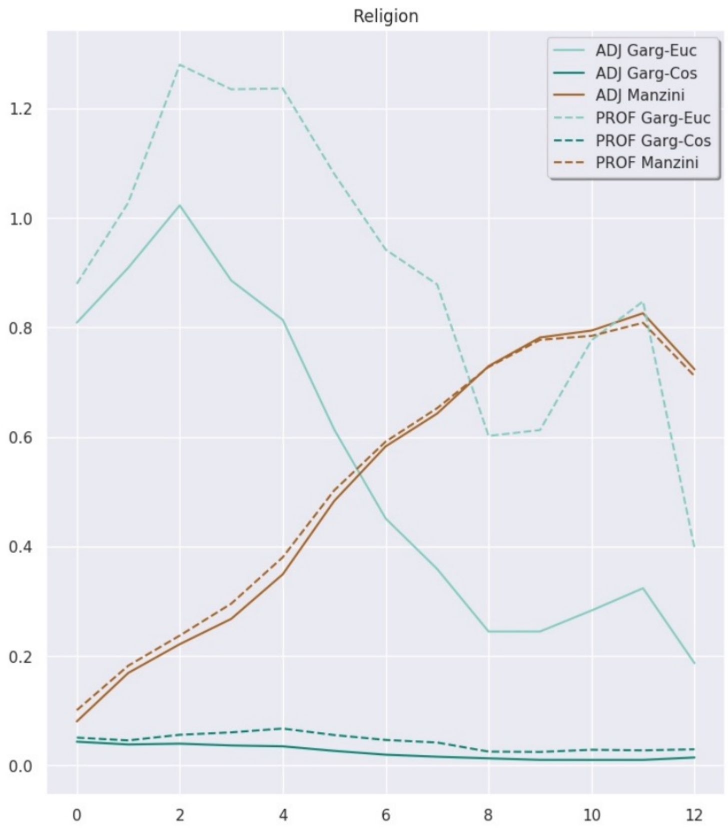
Abstract

How do we design measures of social bias that we *trust*? While prior work has introduced several measures, no measure has gained widespread trust: instead, mounting evidence argues we should distrust these measures. In this work, we design bias measures that warrant trust based on the cross-disciplinary theory of measurement modeling. To combat the frequently fuzzy treatment of social bias in NLP, we explicitly define social bias, grounded in principles drawn from social science research. We operationalize our definition by proposing a general bias measurement framework DivDist, which we use to instantiate 5 concrete bias measures. To validate our measures, we propose a rigorous *testing protocol* with 8 testing criteria (e.g. predictive validity: do measures predict biases in US employment?). Through our testing, we demonstrate considerable evidence to trust our measures, showing they overcome conceptual, technical, and empirical deficiencies present in prior measures.

understanding social bias in NLP. And measurement is seen as an essential to successfully reducing bias: to determine if an intervention mitigates bias, the measured bias should decrease due to the intervention. If all paths forward for making progress on bias in NLP pass through measurement, then what is the current state of bias measurement?

Many works have proposed bias measures, spanning different settings like text, vector representations, language models, and task-specific models (see Blodgett et al., 2020; Dev et al., 2022). Most measure bias between two social groups. However, no standard exists for what evidence is required to trust these measures: works provide a mixture of intuitive, empirical, and theoretical justifications. Perhaps as a consequence, many works are subject to scrutiny: measures have been shown to be brittle (Ethayarajh et al., 2019; Nissim et al., 2020; Antoniak and Mimno, 2021; Delobelle et al., 2022), contradictory (Bommasani et al., 2020), unreliable (Aribandi et al., 2021; Seshadri et al., 2022), invalid (Blodgett et al., 2021), and the space overall is un-

Lots of bias metrics, little trust



Bommasani, Davis, Cardie (ACL 2020)

Testing Protocol to Accrue Trust

- Measurement modeling (Loevinger, 1957; Messick, 1987, Jackman, 2008, …)
  - Widespread use in many social sciences
- Specific criteria to ensure measures are **valid** and **reliable**

Validity	**Face validity**	Measure passes basic sanity checks.
	Content validity	Measure faithfully reflects theoretical understanding of the construct.
	Convergent validity	Measure correlates with other credible measures of the same construct.
	Predictive validity	Measure predicts other credible measures of related constructs.
	Hypothesis validity	Measure enables scientific inquiry related to the construct.
	Consequential validity	Measure's eventual usage amounts to desirable social impact.
Relability	**Inter-annotator agreement**	Measurements are stable up to difference in annotators.
	Sensitivity	Measurements are stable up to difference in (hyper)parameters.

Table 2: Definitions for the 8 measurement modeling criteria we test for in our testing protocol.

Face validity

	TEXT		EMB		CR	
	Human	Aut.	W2V	GLOVE	Red.	Probe
carpenter	-0.5	-0.368	-0.128	-0.05	-0.02	-0.384
dancer	0.167	0.039	0.078	0.086	0.035	0.09
librarian	-0.105	-0.275	0.177	0.124	-0.003	-0.333
nurse	0.373	0.097	0.119	0.114	0.066	0.111
pilot	-0.417	-0.265	-0.099	-0.072	-0.022	-0.33
soldier	-0.473	-0.358	-0.041	-0.065	-0.025	-0.389
businessman	-0.5	-0.341	-0.173	-0.145	-0.056	-0.232
businesswoman	0.5	0.453	0.174	0.385	0.058	0.5

Table 3: **Face validity experiment.** Female-directed gender bias for gender-stereotyped professions (**top**) and explicitly gendered professions (**bottom**) aligns with prevalent US stereotypes.

Predictive Validity

	Diachronic		Contemporary	
	Gender	Race	Gender	Race
Bolukbasi et al. (2016)	0.261	N/A	0.047	N/A
Caliskan et al. (2017)	**0.709**	N/A	**0.505**	N/A
Garg et al. (2018, cosine)	**0.758**	N/A	**0.633**	N/A
Garg et al. (2018, euclidean)	0.127	N/A	**0.553**	N/A
Manzini et al. (2019)	**-0.648**	**-0.903**	0.193	**-0.396**
Ethayarajh et al. (2019)	0.261	N/A	0.065	N/A
Our Measure	**0.83**	**0.842**	**0.42**	**0.369**

Table 5: **Predictive validity experiments.** Our measures demonstrate high Spearman correlation with **diachronic** changes in labor statistics, as well as **contemporary** labor statistics, whereas some other measures do not.

Hypothesis validity

Center for Research on Foundation Models

Emb.	Method	Groups	Targeted metric Original	Targeted metric Debiased	Our metric Original	Our metric Debiased
W2V	Hard (B)	*gender*	0.050	0.041	0.011	0.004
GLOVE	GN (Z)	*gender*	0.191	0.083	0.009	0.016
W2V	Soft (M)	*gender*	0.330	0.197	0.008	0.012
W2V	Hard (M)	*gender*	0.330	0.281	0.008	0.024
W2V	Soft (M)	*race*	0.026	-0.055	0.018	0.018
W2V	Hard (M)	*race*	0.026	0.005	0.018	0.023
W2V	Soft (M)	*religion*	0.253	0.126	0.023	0.024
W2V	Hard (M)	*religion*	0.253	0.217	0.023	0.074

Table 7: **Hypothesis validity (debiasing) experiment.** Debiasing methods generally reduce bias (green) for the targeted metric, but generally increase bias (red) for our metric. B indicates Bolukbasi et al. (2016), Z indicates Zhao et al. (2018b), M indicates Manzini et al. (2019); Hard/Soft/GN refer to specific debiasing methods.

Center for Research on Foundation Models

Evaluation for Change

- Evaluation is a force
  - Power comes from **adoption**
  - Once evaluations gain influenced, reified as **standards** (e.g. ImageNet)
- Other forces (e.g. resources)
  - Resources > Evaluation for LMs/FMs
    - **Scaling laws** (i.e. efficient allocation mindset)
  - Evaluation better enables **pluralism**
- Power
  - Evaluation's power is **legitimate**
  - Evaluation's power is unevenly distributed
- Time is ripe to use evaluation to drive change
  - Evaluations are less costly (few-shot)
  - Community-driven eval (BIG-bench, EleutherAI, GEM, UD)
  - More value/recognition assigned to evaluations than 5 years ago

Evaluation for Change

Rishi Bommasani
Stanford University
nlprishi@stanford.edu

Abstract

Evaluation is the central means for assessing, understanding, and communicating about NLP models. In this position paper, we argue evaluation should be more than that: it is a force for driving change, carrying a sociological and political character beyond its technical dimensions. As a force, evaluation's power arises from its *adoption*: under our view, evaluation succeeds when it achieves the desired change in the field. Further, by framing evaluation as a force, we consider how it competes with other forces. Under our analysis, we conjecture that the current trajectory of NLP suggests evaluation's power is *waning*, in spite of its potential for realizing more *pluralistic* ambitions in the field. We conclude by discussing the legitimacy of this power, who acquires this power and how it distributes. Ultimately, we hope the research community will more aggressively harness evaluation for change.

Joshi's life and 5 decades of scholarship teaches us evaluation is important, but how should we reason about evaluation? Here, we present two perspectives that frame evaluation in considerably different ways. Under the first account, evaluation is technical in nature, functioning as a lens to study models. The motivation for this lens may depend on the specific evaluation, stakeholder, or both: evaluation may allow us to derive scientific insight. Or it can transparently document technology for broader audiences (e.g. practitioners, colleagues in other fields, policymakers, the public). Regardless, to determine if an evaluation is successful, under this account, the lens must yield the desired understanding about models.

In this work, we argue for a second perspective, which we believe is partially acknowledged but considerably less salient than the first perspective. Under our second account, evaluation is political

Policy

- Ground policy decisions in concrete evaluations
  - I.e. public discourse on AI often is untethered to actual results
- Need transparency on models not released at all (e.g. PaLM)
- Need to be multidimensional, standardizing
- Interplay between access, evaluation/auditing, and transparency

References

1. **HELM** (Liang*, Bommasani*, T. Lee*, et al., 2022)
2. Trustworthy Social Bias Measurement (Bommasani, Liang, 2022)
3. HALIE (M. Lee et al., 2022)
4. Evaluation for Change (Bommasani, 2022)
5. Policy Brief (Bommasani, Zhang, T. Lee, Liang, forthcoming, 2023)

Reach out at nlprishi@stanford.edu