Stanford University
Human-Centered
Artificial Intelligence

Issue Brief
HAI Policy & Society
February 2023

Improving Transparency in AI Language Models: A Holistic Evaluation

Rishi Bommasani, Daniel Zhang, Tony Lee, Percy Liang

ARTIFICIAL INTELLIGENCE (AI) LANGUAGE MODELS ARE EVERYWHERE. People can talk to their smartphone through voice assistants like Siri and Cortana. Consumers can play music, turn up thermostats, and check the weather through smart speakers, like Google Nest or Amazon Echo, that likewise use language models to process commands. Online translation tools help people traveling the world or learning a new language. Algorithms flag "offensive" and "obscene" comments on social media platforms. The list goes on.

The rise of language models, like the text generation tool ChatGPT, is just the tip of the iceberg in the larger paradigm shift toward foundation models—machine learning models, including language models, trained on massive datasets to power an unprecedented array of applications. Their meteoric rise is only surpassed by their sweeping impact: They are reconstituting established industries like web search, transforming practices in classroom education, and capturing widespread media attention. Consequently, characterizing these models is a pressing social matter: If an AI-powered content moderation tool that flags toxic online comments cannot distinguish between offensive and satirical uses of the same word, it could censor speech from marginalized communities.

Key Takeaways

Transparency into AI systems is necessary for policymakers, researchers, and the public to understand these systems and their impacts. We introduced Holistic Evaluation of Language Models (HELM) as a framework to benchmark language models as a concrete path to provide this transparency.

Traditional methods for evaluating language models focus on model accuracy in specific scenarios. Since language models are already used for many different purposes (e.g., summarizing documents, answering questions, retrieving information), HELM covers a broader range of use cases, evaluating for the many relevant metrics (e.g., fairness, efficiency, robustness, toxicity).

In the absence of a clear standard, language model evaluation has been uneven: Different model developers evaluate on different benchmarks, meaning their models cannot be easily compared. We establish a clear head-to-head comparison by evaluating 34 prominent language models from 12 different providers (e.g., OpenAI, Google, Microsoft, Meta).

HELM serves as public reporting on AI models—especially for those that are closed-access or widely deployed—empowering decision-makers to understand their function and impact and to ensure their design aligns with human-centered values.

But how to evaluate foundation models is an open question. The public lacks adequate transparency into these models, from the code underpinning the model to the training and testing data used to bring it into the world. Evaluation presents a way forward by concretely measuring the capabilities, risks, and limitations of foundation models.

In our paper from a 50-person team at the Stanford Center for Research on Foundation Models (CRFM), we propose a framework, *Holistic Evaluation of Language Models (HELM)*, to address the lack of transparency for language models. HELM implements these comprehensive assessments—yielding results that researchers, policymakers, the broader public, and other stakeholders can use.

The Importance of Transparency and Evaluation

Transparency is critical for understanding AI systems and designing better policies around them. Black-box decision-making remains a challenge for policymakers, researchers, company executives, and the public seeking to understand why an AI model is generating a particular output. Further, language models developed and used by companies like Google and Microsoft in search engines, content moderation, and translation services may be *closed*—meaning they are not accessible to regulators and external researchers, limiting outsiders' ability to understand the system.

Researchers evaluate AI models to increase transparency into black-box decision-making. We highlight three factors essential for effective evaluations. First, *which* models to evaluate: Evaluating models requires *access* to models. Second, *what* to evaluate against: Models should be evaluated on a unified *standard*. Third, *how* to evaluate: Evaluations should consider every factor—from fairness to robustness to ability to generate disinformation—in a *comprehensive* way.

> *Evaluation presents a way forward by concretely measuring the capabilities, risks, and limitations of foundation models.*

Making these decisions intentionally is vital, because evaluation encodes values and priorities into AI systems. For example, evaluation focusing only on accuracy overlooks the fact that many other criteria (from fairness to efficiency) matter when a language model is deployed in practice. Testing a model in different environments, against different performance metrics, exposes real problems with the potential for serious harm, including models that are toxic, dishonest, used to spread disinformation, and more. A holistic evaluation would provide a more complete picture of model behavior and allows their design to align with human-centered values.

Transparency through evaluation helps researchers, policymakers, and the public. It enables a better understanding of how to correct for mistakes and minimize the likelihood of undesirable outcomes, whether for a developer engineering a language model or a member of a regulatory agency seeking to assess

Evaluation focusing only on accuracy overlooks the fact that many other criteria (from fairness to efficiency) matter when a language model is deployed in practice.

a model's impact on a given community. In addition, pushing for transparency through evaluation reinforces the practice of promoting transparency around AI—a critical step to enabling broader trust in AI systems and empowering individuals to make their own assessments of AI in society.

Holistic Evaluation of Language Models (HELM)

Using HELM, we improved transparency of language models along several fronts. HELM has three core elements: (1) We clearly state the evaluation goal and clearly track where the implementation falls short of that goal; (2) We evaluate multiple metrics for every use case because models should satisfy multiple desiderata (e.g., fairness and accuracy); and (3) We run evaluations on all existing models to standardize the results and directly compare models.

We evaluated 34 different language models across 16 different core scenarios and 7 metrics. Each of these models is an interface that takes text as input and emits text as output: The model can be given a document and asked to summarize it, or posed a question and asked to answer it.

Drawing on some of the most prominent and publicly available language models, the 34 models we tested were built by a dozen organizations around the world, including Meta, Microsoft, OpenAI, Tsinghua University, Yandex, Google, and more. The scenarios we tested ranged from answering questions to retrieving information to detecting toxicity.

In addition, the metrics for HELM include accuracy (average correctness), calibration (know what it doesn't know), robustness (perform well across typos and dialects), fairness (perform well for different demographic groups), bias and stereotypes (represent demographic groups equally), toxicity (likelihood of toxic content produced by models), and efficiency (time and energy use for model training and inference).

HELM differs from traditional evaluations that focus on one specific scenario or metric to better improve transparency. Past evaluations might assess how accurately a model classifies the toxicity of a user's social media comment. While useful, this is inadequate for multipurpose language models that should satisfy many desiderata. In the aforementioned hypothetical, one should also evaluate the model's ability to answer questions and summarize documents. And we should require more than just accuracy: The model should not perform worse for some demographics than others, and it should express uncertainty when it does not know the right answer. Grappling with the broader space of use cases and desiderata enables researchers and policymakers to holistically understand models.

Evaluation Outcomes

First, we observed a gap in language model accuracy that varied based on whether the model was limited-access via an application programming interface (API), closed access, or openly accessible to the public. Specifically, limited-access and closed-access models (such as those from Microsoft/NVIDIA, OpenAI, and Anthropic) are more accurate than open models (such as those from Meta, BLOOM, and Tsinghua University).

It is worth noting that the non-open models we evaluated were more accurate, because the public does not have open access to them beyond an API. Yet, the public has a stake in deployed language models because those models affect them—from facilitating language translation every day to impacting the accessibility of language AI to different communities. And the fact that most researchers do not have open access to limited-access and closed-access models constrains the advancement of language AI by preventing them from accessing the evidently state-of-the-art models. It also impedes public understanding of language AI, because researchers cannot analyze the models and describe them clearly to a general audience.

Second, we found that accuracy, robustness, and fairness were highly correlated, meaning the specific models we evaluated that are most accurate are also more fair and robust (under our given definitions). However, this is only in relative terms: Models show significant drops in accuracy when evaluated on language involving typos (low robustness) or spoken in African American English (low fairness compared to standard American English). There are also other models we did not evaluate, and it is possible to evaluate language models under slightly different definitions of terms like fairness and robustness.

> *...the fact that most researchers do not have open access to limited-access and closed-access models constrains the advancement of language AI by preventing them from accessing the evidently state-of-the-art models.*

Hence, it is important to conduct more work to look at robustness and fairness and what relationships may exist between them in different scenarios.

Additionally, we found that fine-tuning language models with human feedback (from OpenAI and Anthropic) can help with accuracy, robustness, and fairness. In fact, smaller language models could compete with models 10 times the size, in some cases. We also found that human evaluation was sometimes essential. Some language models produced effective summaries, but some of the summaries included in the datasets we examined were less accurate.

Given the size and diversity of the language model space, we acknowledge that our study has its limits. The paper is limited by the models it included. For example, some language models are not publicly disclosed, let alone released, and some highly accurate models were not yet evaluated through HELM. The lack of studies of models in languages other than English points to another gap in the current study. There is a substantial opportunity for future work to build upon our framework for holistically evaluating language models.

Policy Implications

Language models affect the public, and policymakers must pay attention to their impacts across speech analysis, disinformation, the accessibility of language translation, web search, and many other areas. Given their already immense and rapidly accelerating societal impact, policymakers should push for greater transparency into these technologies and enhance evaluation efforts. HELM charts a path forward.

> *Given their already immense and rapidly accelerating societal impact, policymakers should push for greater transparency into these technologies and enhance evaluation efforts. HELM charts a path forward.*

Transparency in AI models, including language models, is the subject of growing policy attention. In Congress, numerous bills seek to tackle these issues in ways that encompass or could encompass language models. The Algorithmic Justice and Online Platform Transparency Act would require internet platforms to prevent algorithmic discrimination on the basis of protected classes and increase transparency around their use of algorithms. The Algorithmic Accountability Act would require covered companies to assess high-risk systems that could contribute to bias, inaccuracy, and other harmful outcomes.

However, our research suggests that some closed-access models perform more accurately than some open-access models. Because closed-access models developed by companies are more likely to be deployed, the lack of transparency into those models hampers policymakers, researchers, and the public from understanding these systems. We cannot adjust these systems to protect against harmful biases and discriminatory patterns, nor can we effectively regulate them, if we do not understand them. Public evaluations reporting on all models—especially those that are closed-access or widely deployed—is one area in which policy can help move the needle.

Importantly, HELM also enables researchers, policymakers, and other stakeholders to evaluate AI systems in a more holistic manner. Given the wide implications of language models, in areas ranging from language translation to content analysis to disinformation, measuring accuracy is not sufficient to understand the scope of a model's behavior and its potential benefits and risks to society. Instead, policymakers should understand factors like fairness, robustness (performance across variations), and toxicity on top of accuracy—as well as the relationships between those factors. If trade-offs exist, HELM enables stakeholders to understand those balances too.

Policymakers should also remember that algorithmic evaluations serve different purposes for different stakeholders—and that policymakers could develop their own ways to holistically evaluate language models. The more ubiquitous language models become, the more important their accuracy, fairness, and efficiency, among other metrics, becomes. If American classrooms adopt more AI in learning activities, for example, language algorithms that perform poorly on certain English speech could exacerbate inequities in student learning. If language models become more adept at producing disinformation-carrying headlines, to give another example, policymakers will have to grapple with a range of domestic political as well as national security risks stemming from those models.

Language models are here to stay. HELM enables decision-makers to understand their function and impact—and to ensure their design aligns with human-centered priorities and values. Policymakers should consider the value of holistic algorithmic benchmarks to evaluate commercial application of AI use cases.

The original article is accessible at Percy Liang, Rishi Bommasani, Tony Lee et al., "Holistic Evaluation of Language Models, https://crfm.stanford.edu/helm

Stanford University's Institute on Human-Centered Artificial Intelligence (HAI), applies rigorous analysis and research to pressing policy questions on artificial intelligence. A pillar of HAI is to inform policymakers, industry leaders, and civil society by disseminating scholarship to a wide audience. HAI is a nonpartisan research institute, representing a range of voices. The views expressed in this policy brief reflect the views of the authors. For further information, please contact **HAI-Policy@stanford.edu.**

Rishi Bommasani is a Ph.D. student in computer science at Stanford University who works at the Center for Research on Foundation Models, part of the Stanford Institute for Human-Centered Artificial Intelligence.

Daniel Zhang is a policy research manager and a social science researcher at the Stanford Institute for Human-Centered Artificial Intelligence.

Tony Lee is a research engineer at the Center for Research on Foundation Models, part of the Stanford Institute for Human-Centered Artificial Intelligence.

Percy Liang is an associate professor of computer science and statistics at Stanford University and the director of the Center for Research on Foundation Models, part of the Stanford Institute for Human-Centered Artificial Intelligence.

HAI **Stanford University**
Human-Centered
Artificial Intelligence

Stanford HAI: Cordura Hall, 210 Panama Street, Stanford, CA 94305-1234
T 650.725.4537 **F** 650.123.4567 **E** HAI-Policy@stanford.edu **hai.stanford.edu**