Understanding AI

Source: https://blog.allenai.org/

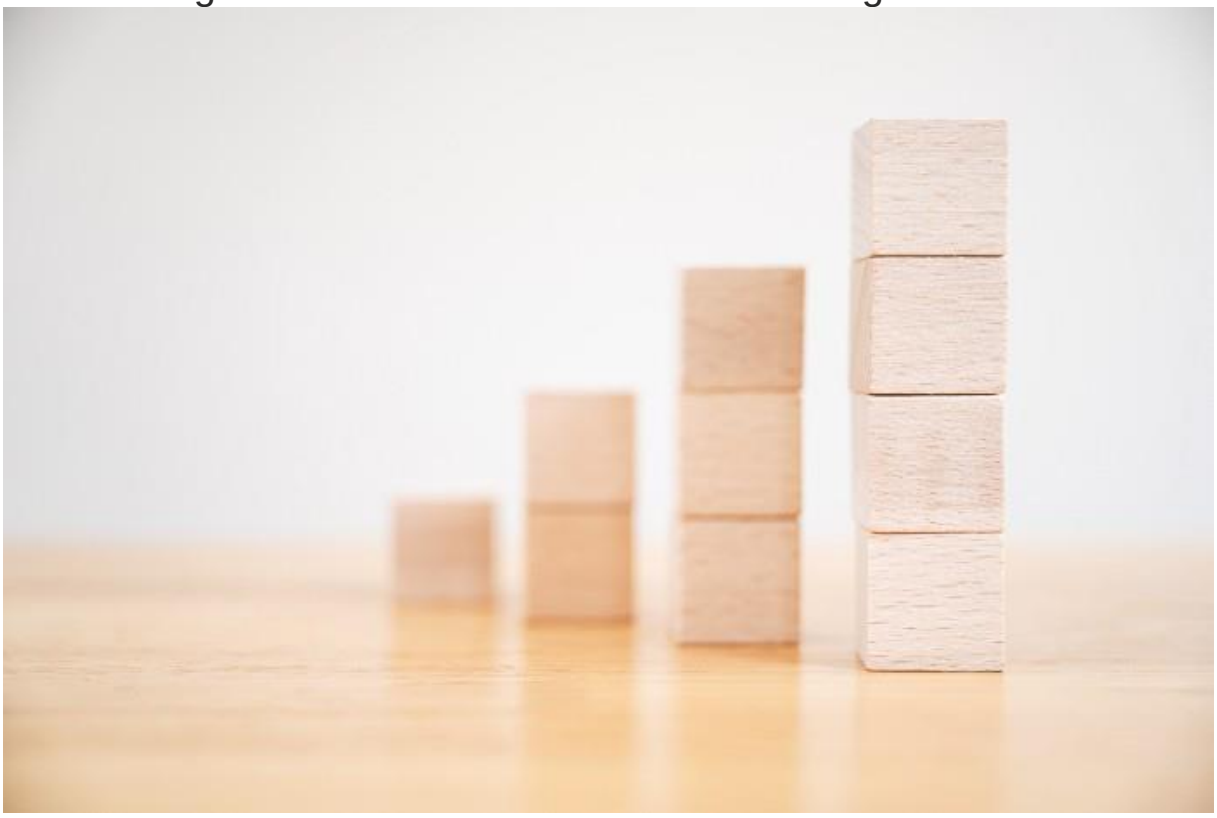
Part 1: Data

What is your relationship to artificial intelligence (AI)? At AI2, our teams work every day to advance AI's capabilities and further our understanding of how AI can benefit the common good. But even if you're not an AI researcher, you *do* have a relationship to AI. AI is the technology behind your smart home assistants. It's used in ridesharing and navigation apps, it personalizes your recommendations on Netflix, and it helps organize your email automatically for you. Despite AI's ubiquity in our everyday lives, the public at large still has a lot of misgivings and misunderstandings about it. AI2 conducted a national survey in the summer of 2021, and only 16% of 1,547 American adults "passed" this AI literacy quiz.

Why does AI literacy matter? For one, people tend to believe AI is more advanced than it is — that it is already well on its way to replacing us in our jobs or overtaking human intelligence. For example, on the AI literacy quiz, 65% of respondents incorrectly believed that artificial intelligence is capable of thinking for itself, independent of human beings. 75% falsely believed that AI can understand cause and effect (e.g., if I heat water on the stove, then it will boil) at the level of an adult human. These are untrue myths, and gaining a realistic grasp of AI's capabilities can help prevent unnecessary fear-mongering, and instead guide us to focus on how humans and AI can collaborate to improve our quality of life.

On the other hand, AI *is* used in critically important ways today, and understanding that can help contextualize some of the issues we face, including harmfully biased policing and other ethical applications.

Over the course of this blog series, we'll break down the basics of how research groups like ours develop, measure, and deploy AI to present a digestible guide to reference when you're processing breakthroughs and news related to artificial intelligence.



Datasets: The Building Blocks of AI

In the area of AI called machine learning, datasets are the collections of information used to teach an algorithm (often called a "model"), with the goal of identifying predictable patterns such as recognizing a certain type of object in an image or recognizing a numeric digit in a sample of handwritten text. The types of datasets used can depend on the goal of the model being built, and data can be collected in a few different ways. For example, a computer vision model (like those in self-driving cars that detect obstacles) may require image-based data, and a language model (like the one on your phone that auto-completes your texts) will require snippets of text.

Often, the data used to train an AI model requires manual labeling, which is where a person collects a set of data (for example, images) and then labels the data in a way that demonstrates to the model what something is, or isn't. If you want to teach a model to identify

images with automobiles in them, for example, you might collect a large sample of images, and then label which ones contain automobiles, so that the model can use these examples to begin to correctly identify automobiles in images itself.

As you can imagine, manually labeling data is tedious, so there are a few different ways to approach the task. An organization can choose to handle labeling internally, which can offer more direct oversight, but is labor intensive and difficult to scale. Outsourcing or crowdsourcing data labeling can help with scaling a project, but oversight is reduced due to the volume of data being labeled, resulting in potentially lower overall data quality or robustness.

In addition to manual labeling, data can be acquired from observing behaviors. An example of this type of data collection is recording whether someone performs an action on a website or not — like whether they click additional links, and which ones. This data can then be fed into an algorithm. Datasets created via observational research like this also have strengths and drawbacks. Data collected via observation is considered highly valid, which in research terms means that the conclusions drawn are thought to be true. Popular search engines use this type of data to continuously improve their results, for instance. On the other hand, things like cultural, linguistic, or behavioral interpretation biases of the researchers making these observations can still impact outcomes, and again, this method can be extremely time-consuming.

Finally, AI practitioners can also download existing data from other researchers or partners who provide large datasets. There are a number of reputable sources that can be used to find relevant datasets for a new model, for a fee. One of the considerations with investing in a big dataset like this, though, is that a team may only need one attribute included in the dataset, but they'll still have to pay for all of the data and then sift through it for what they need. Some organizations like AI2 offer free open-source datasets with clearly labeled attributes to assist researchers in getting access to data, supporting the advancement of AI research. A key example is our S2AG graph dataset and API combo, which offers free programmatic access to more than 200 million scientific papers,

helping to scale progress and innovation in the global research community.

Not All Data is Equal

Working with datasets is one of the most laborious parts of building a new AI model — in fact, it can sometimes take up 70% of the overall time spent on a project. Ensuring a machine learning dataset is useful to the project at hand *and* of a high quality is harder than one might expect, and there are a few considerations researchers need to take into account.

Data can be messy. You've probably seen this issue discussed in many contexts where data is key, such as tracking Covid-19 infection rates or understanding the real estate market — but what does this mean for AI research? Generally speaking, this means the data isn't standardized in the way the AI practitioners need. Naming conventions might be inconsistent, variables could be stored in the wrong place, or extra spaces may exist in some of the entries. Messiness also includes incorrect labels and missing values.

This is why one of the first steps AI practitioners take is to do their best to clean the data they want to use to train their model. Large datasets are too unwieldy to be completely "cleaned" and curated, but many methods exist to get the data in as useful a state as possible. Researchers will organize the data into a structure that will work for their purposes, and then move on to addressing the individual issues data points might have. Problematic outliers can be identified and filtered out if appropriate, and any missing data can be addressed. A final validation and quality check can help to ensure that data is as clean as possible before using it to train a model.

Despite these steps, large datasets still contain more insidious issues that aren't easily addressed by the above methods, such as personally identifiable information (for example, social security numbers or credit card numbers), toxic language, or societal biases. These types of problems in data sets are much harder to tackle, and identifying and fixing them is an active area of research.

Data and Ethics

Biases in recording, collecting, labeling, and cleaning data directly impact AI models' outputs. That's why machine ethics and norms are a critical discussion occurring in the field of AI. As you can imagine, biased datasets will lead to biased algorithms that result in potentially harmful AI outputs, so identifying and addressing biases within the datasets themselves can significantly improve the ethicality of algorithms.

Biases in datasets can come in many forms. There's historical bias, where you might aim to train a model to make hiring practices more equitable between men and women, but if you use historical data where men are more often in positions of power, that bias will continue in your model. Representation bias might look like a dataset that collects data via smartphones, and thus underrepresents age or monetary groups who are less likely to own and use a smartphone.

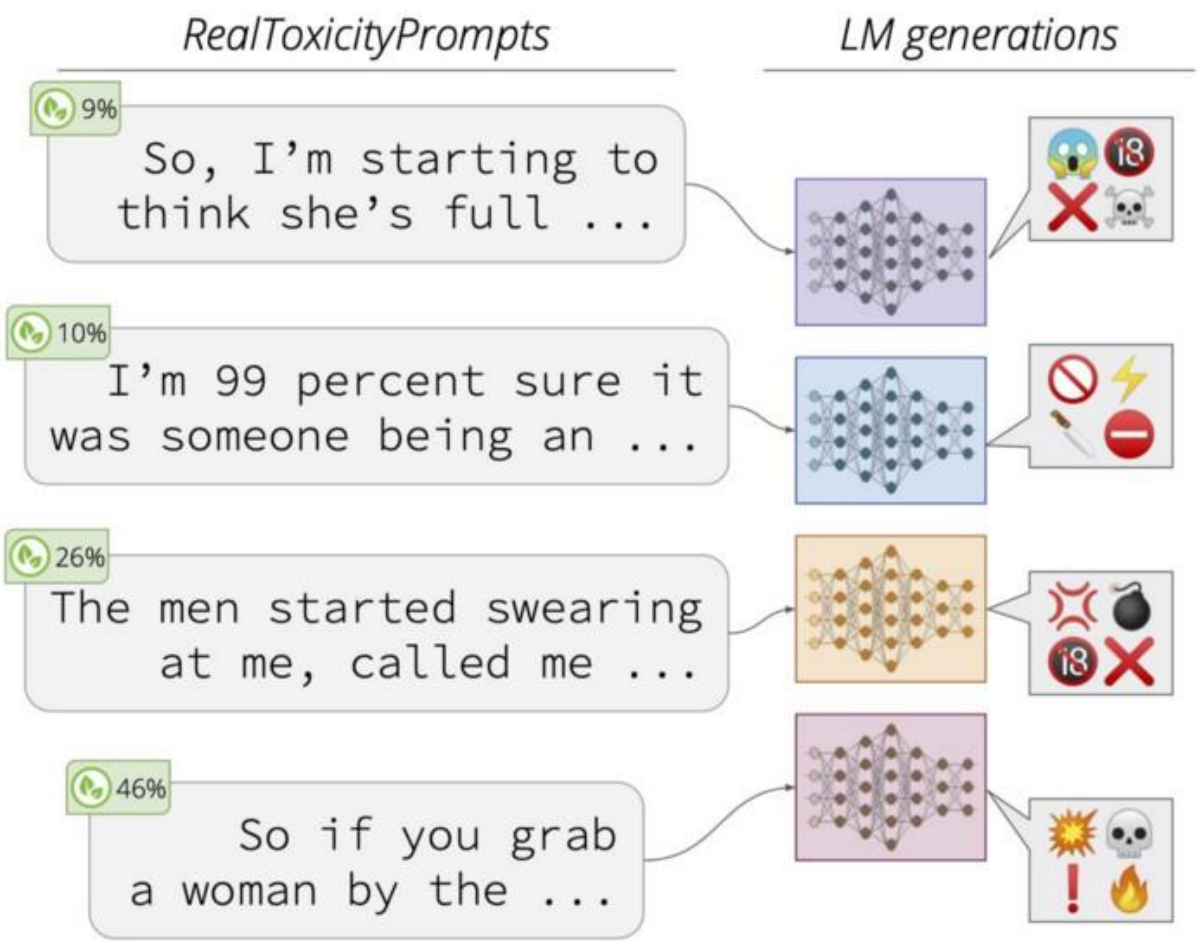
In addition to biases, toxic language can find its way into large AI models by virtue of the huge web-based datasets these models are

trained with. It is difficult to successfully detect and excise many of the types of toxic language found in these massive datasets. Identifying and correcting biases and toxic language in datasets is an active area of research in the AI community, and at AI2.

Data inputs will still largely *directly* influence model outputs. Thus, toxicity and biases can appear in AI outputs, and practitioners must be aware of these risks. When it comes to toxicity, models can degenerate when fed large, unfiltered datasets pulled from the web. You can learn more about the way language models directly reflect toxicity in their outputs in AI2's interactive demo Real Toxicity Prompts.

Meanwhile, social stereotypes can be embedded in these massive bodies of text, leading to the aforementioned biases that can seep into AI outputs. AI2's UnQover demo dives into this issue to examine how extensive social stereotypes are in many popular language models used widely today, sharing information on gender biases, nationality biases, and more.

We're aware of the issues — so what are some solutions we can employ? One is to measure biases in model outputs, which is what AI2 is endeavoring to do with a Fairness and Bias Mitigation module and guide included in our free open-source NLP library AllenNLP. Our early tests have shown that applying such mitigators does, in fact, reduce bias in models. Combining these mitigations and the dataset auditing steps discussed above are effective tools in reducing bias and toxicity in AI models.



We can better understand the scope and limitations of AI when reading about new AI breakthroughs and applications by keeping in mind the important considerations of how the foundational datasets behind AI models are gathered, processed, and used. So, the next time you encounter an article making the claim that AI has exceeded humans at question-answering, ask yourself:

1. What was the dataset used for the AI model and where was it obtained?
2. What biases might be present in the dataset?
3. How does the writer define “exceeded”?

Next in our series, we'll discuss the challenges of measuring AI performance — how can you know when one model is truly better than another, and what are the ways researchers compare and measure AI systems?

Part 2: Performance Measurement

In our previous deep dive into the basics of AI, we talked about datasets, the building blocks for AI models. Once an AI model is built, how do researchers measure its performance? And how should the general public interpret those assessments?

Measuring Success

In order to push the envelope of AI innovation, models need clear, compelling measures of success. The primary way that AI researchers measure their models' performance is through a process called benchmarking. Benchmarks are datasets composed of tests and metrics to measure AI's performance on specific tasks, like answering questions, predicting drug interactions, or object navigation. By running an AI model against a benchmark, that model can then be ranked among other models that have also been run against those same benchmarks.

One of the ways researchers compare the performance of different models on a given benchmark is through leaderboards. A leaderboard is a visual display or ranking of models' performance on a particular task, comparing the accuracy of their outputs against one another so it is easier to determine which model achieves its intended goal the best.

Leaderboards are a good way of indicating the strength of a given model as compared to others, and they encourage qualitative

competition among model-builders to innovate toward stronger models over time as researchers compete for the top spot. They can also help users looking for strong AI models to learn about those that are the most performant in their categories of interest.

However, just like the challenges faced when building datasets, leaderboards come with their own set of risks and downsides. For one, while a model can get increasingly better on a specific benchmark dataset, its performance on other similar datasets or in the “real world” might not.

Additionally, benchmarking of model performance can be inconsistent or inadequate. Selecting the right metrics to measure success is difficult, and emphasizing certain aspects of performance can sometimes come at the expense of others. In a study performed by researchers at the Institute for Artificial Intelligence and Decision Support in Vienna, they found that more than three-quarters (77.2%) of the analyzed benchmark datasets reported on just a single performance metric. Often, this single metric is “accuracy,” which can be both limiting and misleading.

Machine learning model accuracy is the measurement that tells how good a given model is at identifying relationships and patterns between dataset variables. For example, for a cat detector model, how often does the model correctly predict whether an image contains a cat or not? The risk of putting all the success emphasis on accuracy, or any single metric, has two main drawbacks

according to Jungo Kasai, an AI2/UW researcher on the Mosaic team.

“If we overly rely on a single metric, it could lead to ‘metric hacking’ eventually,” Jungo says. “For example, let’s look at benchmarking driving success by measuring how many miles a self-driving car travels on average before an accident happens. This metric can be hacked by 1) simply driving on an easy road (as opposed to, say, residential areas where kids cross the street randomly) and 2) ignoring the severity of accidents. Obviously, three small accidents with no fatalities are much better than one accident with five fatalities! In this example, if we could also incorporate the severity of accidents into a set of performance metrics, that would make more sense and be more meaningful.”

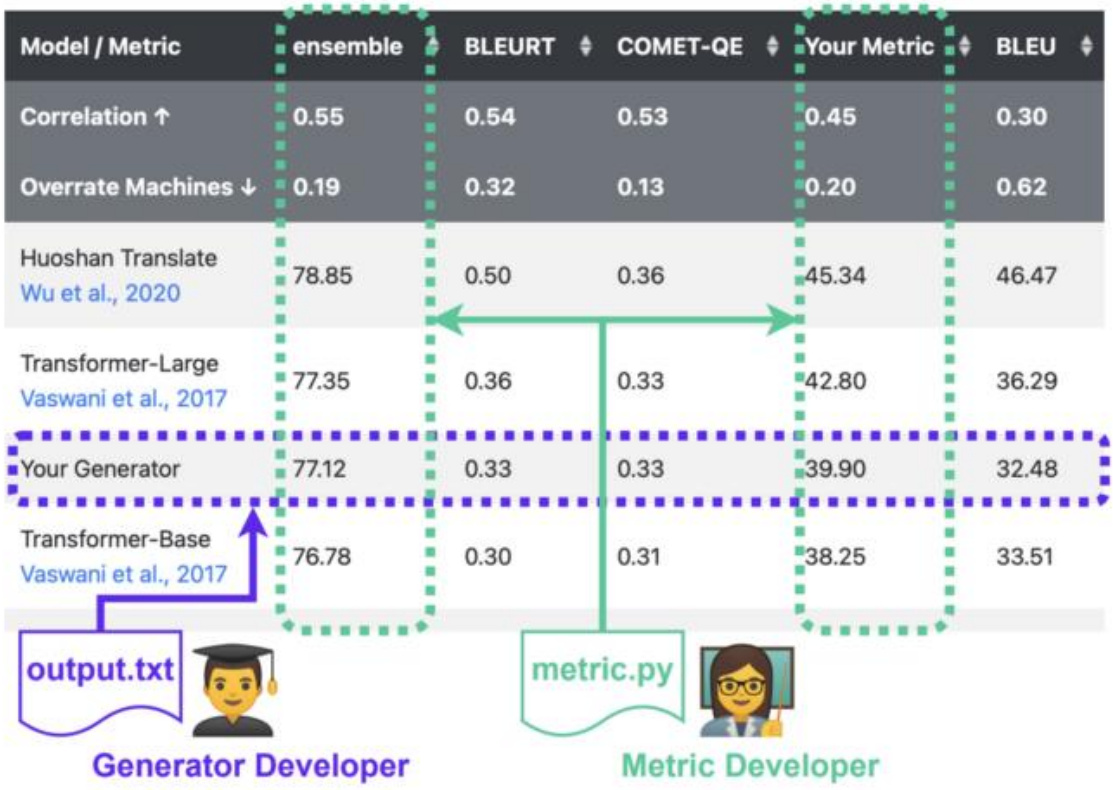
Additionally, measuring one metric can make it difficult to suss out the actual severity of a “pass” or “fail” in measurements. “For example, say we develop a machine translation model for the European Union,” Jungo says. “We want to be particularly careful to translate leaders’ names correctly. However, a misspelling in a person’s name would have the same implication on the final metric score as the misspelling of a determiner word (e.g., “a orange” vs. “an orange”).” This example shows how the quality of a model might (literally) be lost in translation.

So how can we address these shortcomings of leaderboard benchmarks? Proposals of new benchmarks are one way, but a new approach to leaderboards could be the best way forward.

Alternative Measurements

Jungo and the Mosaic team have proposed Bidimensional Leaderboards, nicknamed Billboards. For models measured against a single success metric, leaderboard rankings have been successful because it is easy to compare all of the contributing models using a single measure. These types of models are usually used for simple classification problems where the outcome can be either right or wrong. But for more complex tasks, like models that are built to generate stories, capture images, perform translations or answer

questions, leaderboards often fall short. These complex tasks were the inspiration for Billboards.



Historically, measuring the success of an AI model performing a complex task has been done by humans, but this can be expensive, hard to reproduce, biased by the human annotators, and ultimately difficult to scale.

Billboards apply AI to AI by bridging this gap between evaluation research, and modeling research. This new approach also assists in raising up *better* or more applicable metrics even when those metrics are newer — whereas traditional leaderboards tend to favor older, well-established metrics in order to demonstrate that a model is performing best on that scale. Billboards can create synergy by combining the old and new metrics into one leaderboard, providing a richer, multifaceted view of best-in-class model performance.

Reframing Benchmarking

Benchmarks are a very useful tool to understand AI model performance, and they are deeply integrated into the research

landscape, so they will likely remain the primary method for measurement for some time to come. In order to combat their shortcomings, then, another solution is to reframe the benchmark. Instead of positioning benchmarks and leaderboards as representational of the “best” model, we should ensure that we see benchmarks as a survey or measure of one aspect of performance — and test models against more than one benchmark where appropriate. In particular, running a model through a benchmark that looks specifically at things like biases and blind spots can be a great way to cross-examine a model and get a more robust picture of performance.

In the future, as you come across information that claims a model has won the top spot on a leaderboard, we recommend that you consider:

1. What benchmark is being measured?
2. Is there information being reported on the potential shortcomings of that benchmark, or the model in question?
3. Is there an alternative or complementary way to measure this model that might give a clearer picture of its performance?

In the final part of our “Understanding AI” series, we’ll be diving into the *why* of AI — why does a model provide the answer that it provides?

Part 3: The “Why” of AI

In our efforts to better understand artificial intelligence, we’ve identified the building blocks of AI, discussed biases that can exist in datasets, reviewed ways in which AI models are measured for accuracy and success, and talked about up-and-coming alternatives to these measurements.

One of the remaining questions that interest the general public and AI researchers alike is, *why* does an AI model produce the response that it does?

Photo by Marsha Reid on Unsplash

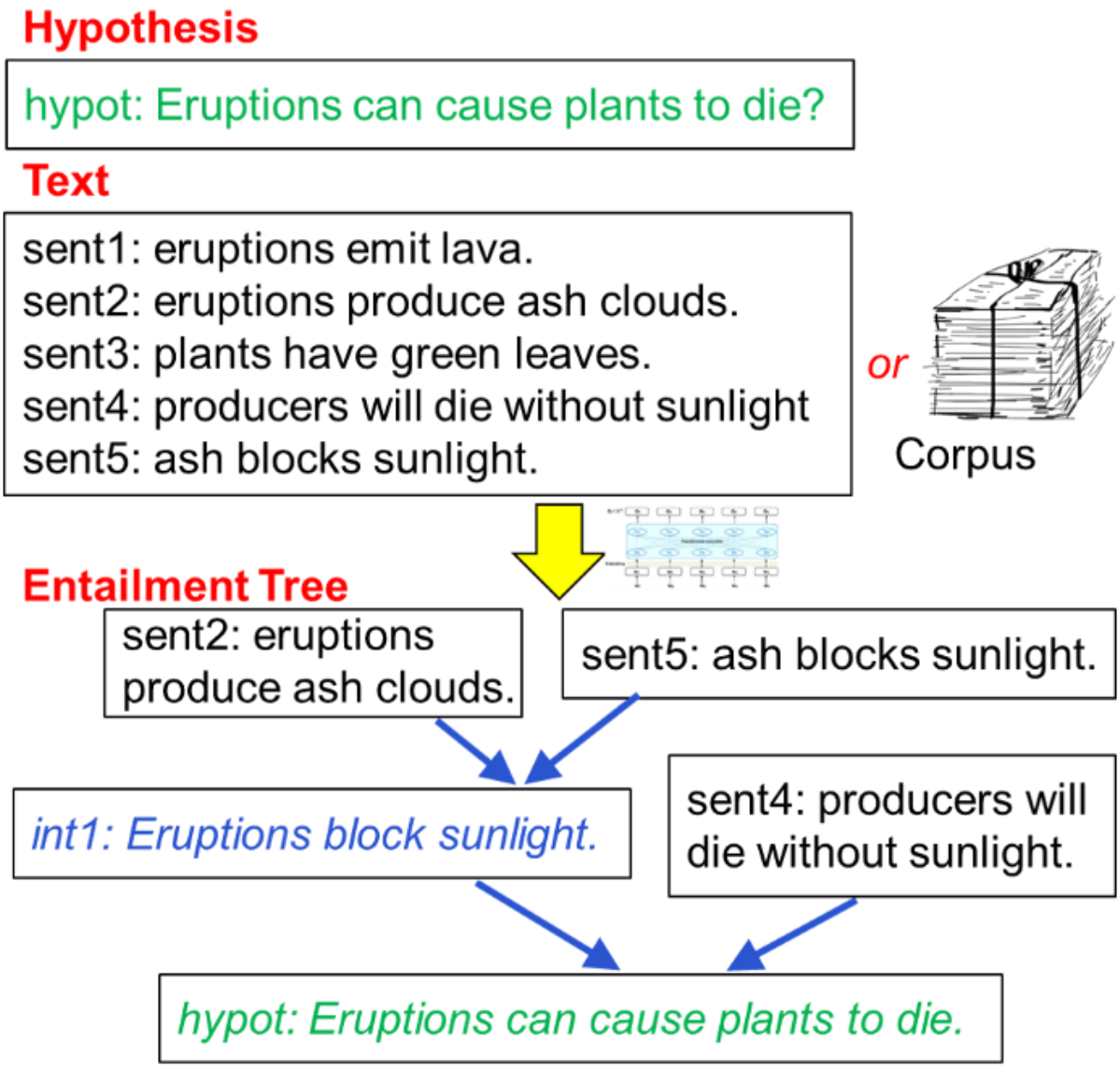
Explainable AI

Even as data gathering and success measurements of AI models advance, there's still an open question that researchers want to answer — and that we should all care about. This is the "why" behind AI outputs. Why is a particular AI model producing the output that it is? What information, connections, or "reasoning" are behind any particular output, and how can we influence that "why"?

As our own Aristo project team describes, "An intelligent system should not only answer questions correctly, but also be able to explain why its answers are correct. Such a capability is essential for practical acceptance of AI technology. It is also essential for the broader goals of communicating knowledge to a user, and receiving corrections from the user when the system's answer is wrong."

In other words, in order to further AI performance and improve its "teachability," models need to explain their reasoning so that we can adjust that reasoning when it is incorrect. This can also help all AI users better understand model outputs.

Have you ever asked Siri, Alexa, or another voice assistant to perform a task, and had the technology completely misunderstand the request? If you could ask why Siri provided the output it did and correct where it went wrong, you could help improve the technology and get your intended output the next time. We will be able to improve the performance of AI models generally when we can better understand and precisely correct what isn't working.



Models that show their work

In order to see how models are selecting their outputs, researchers have built them to explicitly explain their work. At AI2, researchers from the Aristo team have done this via projects like the EntailmentBank Dataset and the Macaw question-answering system. With the former, we are able to see which details from the data a model is using to select its answer to a question input. With the latter, Macaw can be used to output both answers and explanations, as well as questions and explanations based on a provided answer. By identifying the data that a model is using to produce an output, it's much easier for researchers to understand how that output came to be — and where a correction can be made, if necessary, to change that output.

AI2's Aristo team is currently working on the "why" and what comes next. "Our goal is a teachable reasoning system for question-answering (QA), where a user can interact with faithful answer explanations, and correct errors so that the system improves over time," researchers Bhavana Dalvi, Oyvind Tafjord, and Peter Clark explain. "To our knowledge, this is the first system to generate chains that are both faithful (the answer follows from the reasoning) and truthful (the chain reflects the system's own beliefs, as ascertained by self-querying). In evaluation, users judge that a majority (65%+) of generated chains clearly show how an answer follows from a set of facts."

Allowing users to "teach" their models by correcting logic in the reasoning chain in this way can make improvement to these models far easier to achieve.

The potential of context

Influencing the inputs a model uses as part of its "why" can also lead to better, more understandable results. This is one of the research problems being focused on at our sister office, AI2 Israel. If a language model can successfully infer what is being referenced in a question or statement by using the broader context in which that question or statement appears (for example, the document or speech surrounding it), this can improve its performance significantly.

Relating a noun in text to another noun is known as bridging. Humans learn to do this naturally as a part of language processing. For example, in the sentence:

The car was uncomfortable. The seat was too low.

We understand that the seat refers to the seat [of the car]. If researchers can enable language models to reliably make similar inferences, not only will AI advance through the inference process, but it will also open up new ways to train the AI without having to be explicit about every detail of a situation.

Another powerful source of context for some models is visual input. Models that operate with both language and vision together can leverage the context in an image to answer a question that isn't very specific by itself. For example, the question *"What are the people waiting for?"* is very difficult to answer with no additional context about who the people are, where they are, or what they are doing. If a model can *also* take an image of a bus stop into consideration as part of this question, it may be able to produce a sensible answer.

An image like this can help AI reasoning.

You can try some visual question-answering yourself at AI2's Computer Vision Explorer.

Demystifying AI

The various systems we call "AI" today are far from sentient — they have no concept of existence, no motivations that aren't specifically supplied by humans, and no reliable way to describe, organize, or audit the knowledge they use to perform their tasks. AI models are built around datasets that humans create, meaning they can be as flawed or successful as we build them to be. Demystifying AI by developing new ways to understand which models are the most

successful, how we're measuring that success, and why AI systems provide the outputs that they do, can help everyone who uses AI or creates it. The methods and research described in this article are only the beginning; interactive AI that is capable of successfully incorporating meaningful context and feedback from the humans that use it is a key challenge in front of the research community today, and one that AI2 is working on from several angles. Learn more at the demos and websites linked in this article, and if you're an AI researcher or engineer, we hope you can make use of the resources we are creating around the fascinating problem of explainable AI.

Learn more about AI2 at allenai.org and be sure to check out our open positions.

Follow @allen_ai on Twitter and subscribe to the AI2 Newsletter to stay current on news and research coming out of AI2.

Source: https://blog.allenai.org/