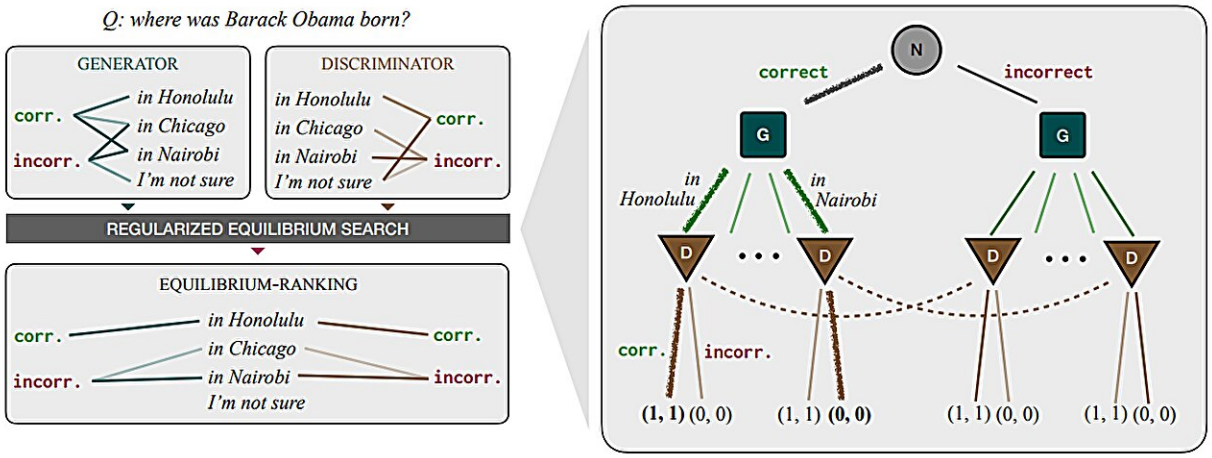
Using ideas from game theory to improve the reliability of language models

May 15 2024, by Rachel Gordon



(Left) Overview of our approach. (Right) Structure of the CONSENSUS GAME, a twoplayer sequential signaling game with imperfect information. Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2310.09139

Imagine you and a friend are playing a game where your goal is to communicate secret messages to each other using only cryptic sentences. Your friend's job is to guess the secret message behind your sentences. Sometimes, you give clues directly, and other times, your friend has to guess the message by asking yes-or-no questions about the clues you've given. The challenge is that both of you want to make sure you're understanding each other correctly and agreeing on the secret message.

MIT Computer Science and Artificial Intelligence Laboratory (CSAIL) researchers have created a similar "game" to help improve how AI understands and generates text. It is known as a "consensus game" and it involves two parts of an AI system—one part tries to generate sentences (like giving clues), and the other part tries to understand and evaluate those sentences (like guessing the secret message).

The researchers discovered that by treating this interaction as a game, where both parts of the AI work together under specific rules to agree on the right message, they could significantly improve the AI's ability to give correct and coherent answers to questions.

They tested this new game-like approach on a variety of tasks, such as reading comprehension, solving math problems, and carrying on conversations, and found that it helped the AI perform better across the board. Their paper is published on the *arXiv* preprint server.

Traditionally, large language models answer one of two ways: generating answers directly from the model (generative querying) or using the model to score a set of predefined answers (discriminative querying), which can lead to differing and sometimes incompatible results.

With the generative approach, "Who is the president of the United States?" might yield a straightforward answer like "Joe Biden." However, a discriminative query could incorrectly dispute this fact when evaluating the same answer, such as "Barack Obama."

So how do we reconcile mutually incompatible scoring procedures to achieve coherent, efficient predictions?

"Imagine a new way to help language models understand and generate text, like a game. We've developed a training-free, game-theoretic method that treats the whole process as a complex game of clues and

signals, where a generator tries to send the right message to a discriminator using natural language. Instead of chess pieces, they're using words and sentences," says Athul Jacob, an MIT Ph.D. student in electrical engineering and computer science and CSAIL affiliate.

"Our way to navigate this game is finding the 'approximate equilibria,' leading to a new decoding algorithm called 'equilibrium ranking.' It's a pretty exciting demonstration of how bringing game-theoretic strategies into the mix can tackle some big challenges in making language models more reliable and consistent."

When tested across many tasks, like reading comprehension, commonsense reasoning, math problem-solving, and dialogue, the team's algorithm consistently improved how well these models performed. Using the ER algorithm with the LLaMA-7B model even outshone the results from much larger models.

"Given that they are already competitive, that people have been working on it for a while, but the level of improvements we saw being able to outperform a model that's 10 times the size was a pleasant surprise," says Jacob.

Game on

"Diplomacy," a strategic board game set in pre-World War I Europe, where players negotiate alliances, betray friends, and conquer territories without the use of dice—relying purely on skill, strategy, and interpersonal manipulation—recently had a second coming.

In November 2022, computer scientists, including Jacob, developed "Cicero," an AI agent that achieves human-level capabilities in the mixed-motive seven-player game, which requires the same aforementioned skills, but with natural language. The math behind this

partially inspired the Consensus Game.

While the history of AI agents long predates when OpenAI's software entered the chat in November 2022, it's well documented that they can still cosplay as your well-meaning, yet pathological friend.

The consensus game system reaches equilibrium as an agreement, ensuring accuracy and fidelity to the model's original insights. To achieve this, the method iteratively adjusts the interactions between the generative and discriminative components until they reach a consensus on an answer that accurately reflects reality and aligns with their initial beliefs. This approach effectively bridges the gap between the two querying methods.

In practice, implementing the consensus game approach to language model querying, especially for question-answering tasks, does involve significant computational challenges. For example, when using datasets like MMLU, which have thousands of questions and multiple-choice answers, the model must apply the mechanism to each query. Then, it must reach a consensus between the generative and discriminative components for every question and its possible answers.

The system did struggle with a grade school right of passage: math word problems. It couldn't generate wrong answers, which is a critical component of understanding the process of coming up with the right one.

"The last few years have seen really impressive progress in both strategic decision-making and language generation from AI systems, but we're just starting to figure out how to put the two together. Equilibrium ranking is a first step in this direction, but I think there's a lot we'll be able to do to scale this up to more complex problems," says Jacob.

An avenue of future work involves enhancing the base model by integrating the outputs of the current method. This is particularly promising since it can yield more factual and consistent answers across various tasks, including factuality and open-ended generation. The potential for such a method to significantly improve the base model's performance is high, which could result in more reliable and factual outputs from ChatGPT and similar language models that people use daily.

"Even though modern language models, such as ChatGPT and Gemini, have led to solving various tasks through chat interfaces, the statistical decoding process that generates a response from such models has remained unchanged for decades," says Google Research Scientist Ahmad Beirami, who was not involved in the work.

"The proposal by the MIT researchers is an innovative game-theoretic framework for decoding from language models through solving the equilibrium of a consensus game. The significant performance gains reported in the research paper are promising, opening the door to a potential paradigm shift in language model decoding that may fuel a flurry of new applications."

More information: Athul Paul Jacob et al, The Consensus Game: Language Model Generation via Equilibrium Search, *arXiv* (2023). DOI: 10.48550/arxiv.2310.09139

Provided by Massachusetts Institute of Technology