Artificial Intelligence System with Specialized Personas and Methods of Training and Evaluating Thereof

Inventor: Jack Felix

BACKGROUND OF THE INVENTION

[0001] There is criticism of certain artificial intelligence computer systems such as large language models that the computer systems are trained with inherent bias that may lead to inaccurate responses to prompts and questions. These computer systems are able to accurately recognize patterns and use this pattern recognition to generate responses to prompts and questions. However, pattern recognition does not always equate to a deep understanding of the prompts and questions.

[0002] Inherent bias in computer system training may be present due to the demographics of those who create and train said computer systems. Many computer scientists are educated Asian and Caucasian men between the ages of 25 and 45. Therefore, these computer scientists' understanding of the world may be from only a certain perspective, and therefore the initial training data provided to computer systems created and trained by these computer scientists may be from only this same specific perspective.

[0003] Research, such as Argyle et al., 2023, has shown that conditioning computer systems with human backstories can lead to a measurable decrease in algorithmic bias. This research can be understood to implement training data in a computer system wherein the training data comprises demographic data defining personas, along with example responses to prompts and questions from those defined personas. If the example responses to prompts and questions are solicited from actual humans with diverse demographic qualities, the computer systems may overcome inherent bias present when creating and training the computer systems.

SUMMARY OF THE INVENTION

[0004] Embodiments of the present invention may comprise a computer system and method of training and evaluating thereof. A memory may contain instructions that, when executed by a processor, cause the processor to perform a series of steps. The instructions may include data describing demographic qualities that may be organized to define personas within the memory. Machine answers to questions may be solicited from the processor based on the defined personas. Human answers to the same questions may be solicited from humans with demographic qualities representing those of the defined personas. Variances between the machine answers and human answers may be calculated using Human-AI Variance Scores. The Human-AI Variance Scores may be provided to the memory in the form of feedback data. The steps of soliciting machine answers, soliciting human answers, comparing the machine answers and human answers using Human-AI Variance Scores, and providing the Human-AI Variance scores to the memory in the form of feedback data may be iteratively performed until the Human-AI Variance scores are calculated to be a desired value.

DETAILED DESCRIPTION OF THE INVENTION

[0005] The description of the present invention provided herein describes example embodiments of the present invention and is not intended to limit the present invention to any particular embodiment, feature, component, method step, calculation, hardware, software, or any other property. The embodiments described herein are non-limiting implementations of the present invention.

[0006] Embodiments of the present invention may comprise a computer system. The computer system may have a processor. The processor may be any component or mechanism capable of receiving computer-readable instructions and producing an output accordingly. The processor may be a single-core processor, a dual-core processor, or any other form of computer processor known in the art. The computer system may also have a memory comprising a non-transitory computer-readable medium containing instructions that, when executed by the processor, cause the processor to perform a series of steps. The memory may be any component or mechanism capable of storing computer-readable instructions in a non-transitory format. The memory may be a hard drive such as but not limited to a solid-state hard drive or a hard drive disk.

[0007] When the processor executes the instructions contained in the non-transitory computer-readable medium of the memory, the processor may provide machine answers to questions using a numerical scale based on a defined persona. The numerical scale may be a scale from 0-100, where 0 represents the most negative answer to the questions, and where 100 represents the most positive answer to the questions. For example, if a question is, “Are financial stimulus packages helpful for an economy?”, an answer of 0 may connotate a strong “No”, an answer of 100 may connotate a strong “Yes”, and an answer of 50 may connotate a completely neutral opinion.

[0008] The defined persona may be a compilation of demographic data. The demographic data may be a combination of one or more of: age, gender, political affiliation, location, occupation, marital status, number of children, income, religious affiliation, race/ethnicity, and education level. This list of demographic qualities is non-limiting; additional demographic qualities may be used to define personas.

[0009] Upon providing the machine answers to the questions, the processor may be provided feedback regarding the machine answers. In embodiments wherein the computer system is a large language model, the feedback may be provided in the form of written or spoken natural language. In embodiments wherein the computer system is not a large language model, the feedback may be provided in the form of computer-readable instructions. The processor may be configured to update the instructions stored in the memory based on the feedback. If the processor is then caused to provide machine answers to the same questions again, the machine answers provided by the processor may be different if the feedback had caused the processor to update the instructions stored in the memory accordingly.

[0010] The instructions stored in the memory may comprise training data. The training data may comprise information on demographic qualities of defined personas that may be used to create defined personas within the computer system. The defined personas may be broad by defining only one or two demographic qualities (ex: Republicans, or Republican women). Alternatively, the defined personas may be relatively specific by defining many demographic qualities (ex: Hispanic, Republican, married women who live in Texas, have at least one child, work in healthcare, and earn between $80,000 and $100,000 per year). The training data may also comprise human answers to questions using a numerical scale, such as from 0-100 as described previously herein. The human answers to questions may be solicited from humans with demographic qualities that represent those of the defined personas. For example, if a persona of a Republican married woman is defined within the memory, then human answers to questions may be solicited from actual women who are married and identify politically as Republicans. These human answers may be stored in the memory as training data so the processor has a reference when asked to provide machine answers to questions based on a defined persona.

[0011] The training data may also comprise feedback data comprising Human-AI Variance Scores showing calculated variances between the machine answers to questions and the human answers to questions. The Human-AI Variance Scores may be calculated to show variances only between machine answers based on defined personas and human answers solicited from humans with demographic qualities that represent those of the same defined personas. For example, Human-AI Variance Scores may be calculated using machine answers based on the persona of Republican women and human answers from actual Republican women. However, the Human-AI Variance Scores may not be calculated using machine answers based on the persona of Republican women and human answers from actual Democratic men, since this would result in an inaccurately low Human-AI Variance Score.

[0012] The Human-AI Variance Score may determine how closely the machine answers mimic the human answers. A high Human-AI Variance Score may signify that the machine answers closely mimic the human answers. A low Human-AI Variance Score may signify that the machine answers do not closely mimic the human answers. The Human-AI Variance Scores may be calculated by taking differences between each of the human answers from humans having demographic qualities of a specific persona and each of the machine answers based on the same specific persona. These differences may then be squared, and the sum of all the squared differences may be taken. The square root of this sum may then be taken. This square root may then be divided by the total number of differences between human answers and machine answers.

[0013] The Human-AI Variance Scores may be calculated by the processor upon executing some of the instructions stored in the non-transitory computer-readable medium of the memory. The Human-AI Variance Scores may be automatically stored as feedback data in the memory upon being calculated by the processor. The feedback data may then be provided to the processor automatically such that when machine answers to the same questions are solicited from the processor a second time, the machine answers solicited the second time are different from the original machine answers, with the goal of achieving a higher Human-AI Variance score using the machine answers solicited the second time. Machine answers may then be solicited a third time, fourth time, etc. and the Human-AI Variance scores may be calculated and provided as feedback data a third time, fourth time, etc. until desired Human-AI Variance scores are calculated, showing that the computer system has been trained to closely mimic actual human answers to questions.

[0014] In some embodiments, the computer system of the present invention may include a second processor and a second memory comprising a non-transitory computer-readable medium containing instructions that, when executed by the second processor, cause the second processor to perform a series of steps. Instructions for calculating the Human-AI Variance Scores may be stored in the second memory, and the Human-AI Variance Scores may be calculated using the second processor. The feedback data, which may include the calculated Human-AI Variance Scores, may then be provided to the memory manually. Providing the feedback data to the memory manually may include providing the feedback data in the form of computer-readable instructions or in the form of written or spoken natural language.

[0015] The instructions contained in the memory may be organized in a neural network such that feedback data causes the instructions to be automatically updated without manual intervention from a human. The instructions may be organized in a recurrent neural network (RNN), a generative adversarial network (GAN), a liquid neural network (LNN), or any other neural network known in the art. In this manner, the computer system may be defined as an artificial intelligence (AI) system in that it is capable of "learning" by adjusting its own processing instructions based on feedback rather than relying on a human to adjust its processing instructions.

[0016] Though the present invention has been described mainly as a computer system thus far, it is understood that embodiments of the present invention may comprise a method for training a computer system and/or a method for evaluating a computer system using steps and principles described thus far. The computer system trained and/or evaluated by these methods may be the same computer system described thus far.

[0017] The benefit of the system and methods of the present invention is to provide a computer system capable of closely mimicking human answers to questions. While it is understood that the methods described herein may be used to compare human answers to questions with other human answers to questions, this deviates from the nature of the present invention. The computer-implemented nature of the present invention is not merely an implementation of a general system or method, but is rather a means for achieving human-like output from a non-human machine.

Claims

What is claimed is:

1. A computer system comprising:

    a processor;

    a memory comprising a non-transitory computer-readable medium containing instructions that, when executed by the processor, cause the processor to perform a series of steps, the steps comprising:

        providing machine answers to questions using a numerical scale based on a defined persona; and

        updating the instructions stored the memory based on feedback received regarding the machine answers to the questions,

    wherein the instructions stored in the memory comprise:

        training data comprising:

            information on demographic qualities of defined personas;

            human answers to questions using a numerical scale, the human answers to questions being solicited from humans with demographic qualities that represent those of the defined personas; and

        feedback data comprising Human-AI Variance Scores showing calculated variances between the machine answers to questions and the human answers to questions.

2. The system of Claim 1, wherein the Human-AI Variance Scores are calculated using the processor,

    wherein instructions for calculating the Human-AI Variance Scores are stored in the memory,

    wherein the Human-AI Variance Scores are stored as feedback data in the memory automatically,

    and wherein the feedback data is provided to the processor automatically.

3. The system of Claim 1, further comprising:

    a second processor; and

a second memory comprising a non-transitory computer-readable medium containing instructions, that, when executed by the second processor, cause the second processor to perform a series of steps, the steps comprising calculating the Human-AI Variance Scores,

wherein instructions for calculating the Human-AI Variance Scores are stored in the second memory,

wherein the feedback data is provided to the processor and the memory manually.

4. The system of Claim 1, wherein the computer system is a large language model configured to generate content in the form of human-readable text representing natural language.
5. A method of training a computer system comprising:

providing the computer system, the computer system comprising:

a processor;

a memory comprising a non-transitory computer-readable medium containing instructions that, when executed by the processor, cause the processor to perform a series of steps;

providing training data to the computer system, the training data comprising:

information on demographic qualities of defined personas;

human answers to questions using a numerical scale, the human answers to questions being solicited from humans with demographic qualities that represent those of the defined personas;

soliciting machine answers to questions from the computer system using a numerical scale based on a defined persona; and

calculating Human-AI Variance Scores based on variances between the machine answers to questions and the human answers to questions; and

providing the Human-AI Variance Scores to the memory as feedback data.

6. The method of Claim 5, further comprising iteratively performing the steps of:

soliciting machine answers to questions from the computer system using a numerical scale based on a defined persona;

calculating Human-AI Variance Scores based on variances between the machine answers to questions and the human answers to questions; and

providing the Human-AI Variance Scores to the memory as feedback data,

wherein these steps are iteratively performed until the Human-AI Variance Scores are calculated to be a desired value.

7. A method of evaluating a computer system comprising:

defining two or more personas using demographic qualities;

finding humans with demographic qualities that match those of the two or more personas;

classifying each of the humans into one of the two or more personas based on the humans' demographic qualities;

soliciting human answers to questions using a numerical scale;

soliciting machine answers to questions from the computer system using a numerical scale based on a defined persona;

calculating numerical differences between each of the human answers from humans of a specific persona and each of the machine answers from the computer system based on the same specific persona;

squaring the differences between the human answers and the machine answers;

taking the sum of all the squared differences;

taking the square root of the sum of all the squared differences; and

dividing the square root of the sum of all the squared differences by the total number of numerical differences calculated between human answers and machine answers to calculate a Human-AI Variance Score.