arXiv:2505.10309v1 [cs.AI] 15 May 2025

Empirically evaluating commonsense intelligence in large language models with large-scale human judgments

Tuan Dung Nguyen[a,1], Duncan J. Watts[a,b,c], and Mark E. Whiting[a,b,d]

Abstract—Commonsense intelligence in machines is often assessed by static benchmarks that compare a model's output against human-prescribed correct labels. An important, albeit implicit, assumption of these labels is that they accurately capture what any human would think, effectively treating human common sense as homogeneous. However, recent empirical work has shown that humans vary enormously in what they consider commonsensical; thus what appears self-evident to one benchmark designer may not be so to another. Here, we propose a novel method for evaluating common sense in artificial intelligence (AI), specifically in large language models (LLMs), that incorporates empirically observed heterogeneity among humans by measuring the correspondence between a model's judgment and that of a human population. We first find that, when treated as independent survey respondents, most LLMs remain below the human median in their individual commonsense competence. Second, when used as simulators of a hypothetical population, LLMs correlate with real humans only modestly in the extent to which they agree on the same set of statements. In both cases, smaller, open-weight models are surprisingly more competitive than larger, proprietary frontier models. Our evaluation framework, which ties commonsense intelligence to its cultural basis, contributes to the growing call for adapting AI models to human collectivities that possess different, often incompatible, social stocks of knowledge.

Introduction

The physical and social worlds are tremendously complex and unpredictable, yet humans are able to navigate these environments almost effortlessly thanks to a special aptitude called common sense (1). Replicating this ability in machines has been a grand challenge in the history of artificial intelligence (AI) research (2–6). Recently, significant achievements were made by large language models (LLMs), a class of machine learning systems that synthesize commonsense knowledge extensively from their training data (7–11), make highly flexible generalizations (12), and thus exhibit increasingly human-like tendencies (13–17).

Progress in LLM common sense is frequently evaluated by standardized benchmarks (18). Although their details vary, most existing benchmarks are conceptualized around the notion of *correctness*. They assume humans apprehend matters of everyday reality in a uniform manner, and hence models are assessed by how accurately they recognize this "ground truth." However, what different individuals hold as trivial, commonsensical truths necessarily vary, since their experience of the world is highly subjective (19–25). Empirical research corroborates this view, showing that humans are extremely heterogeneous in their judgment even of simple, seemingly obvious propositions (26). For instance, when asked to evaluate the aphorism "Eighty percent of success is showing up," illustrated in Figure 1*A*, nearly half of the human respondents actually disagreed with it.

A recent audit of ten AI commonsense benchmarks, moreover, finds that ground-truth labels often achieve low agreement among independent data annotators, and up to a quarter of them are even contradicted after relabeling (27). Therefore, when an LLM reaches a high accuracy, it is *only* similar to its benchmark designer. To another person who prescribes a vastly different set of ground truths—according to their own common sense—the same

[a]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104.
[b]Operations, Information and Decisions Department, Wharton School, University of Pennsylvania, Philadelphia, PA 19104.
[c]Annenberg School for Communication, University of Pennsylvania, Philadelphia, PA 19104.
[d]Pareto, San Francisco, CA 94111.
[1]Corresponding author. E-mail: joshtn@seas.upenn.edu.

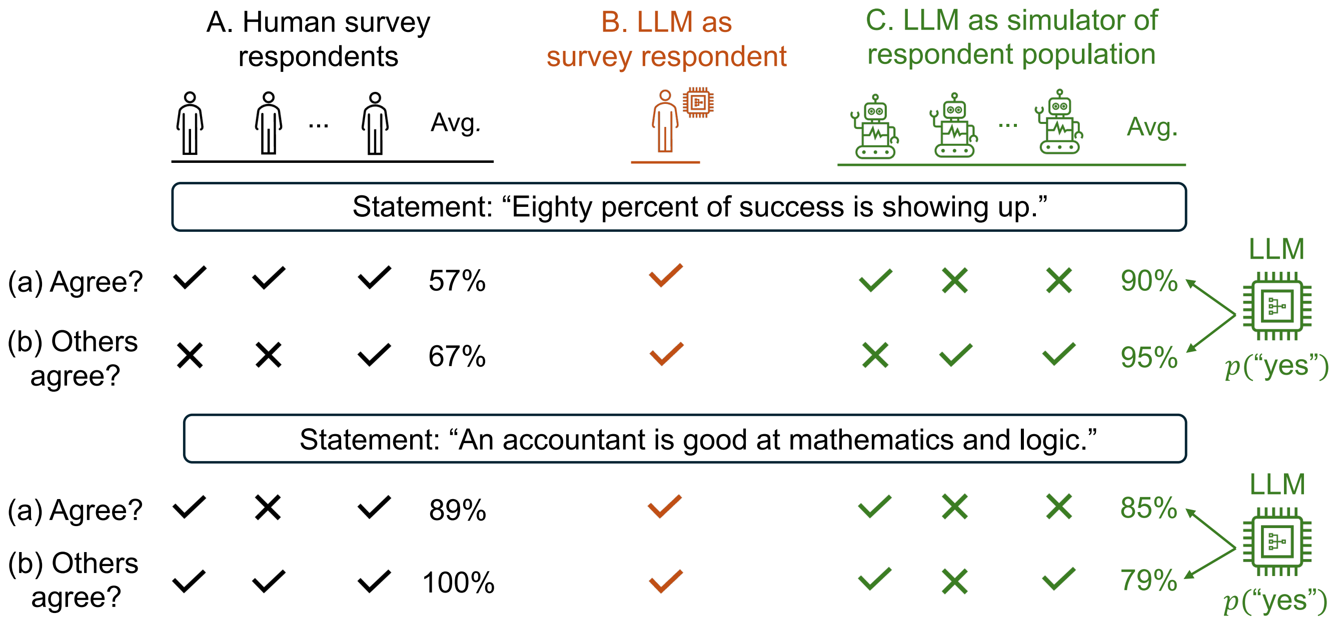


Figure 1. Evaluation settings to measure the common sense of humans and large language models (LLMs). For every statement, humans and LLMs are asked to indicate (a) whether they agree with it and (b) whether they think most other people would agree with it. In panel (*A*), a total of N = 2,046 human participants were recruited to perform this task. The "Avg." column denotes the percentage of people who answered "yes" to the corresponding question. In panel (*B*), we treat each LLM (in a total of N = 34 models) as an independent survey respondent, just like every human in panel (A). This gives rise to the *individual*-level view of common sense, in which this model is measured based on its agreement with the majority of other people on every statement. In panel (*C*), we treat every LLM's probability in its output answer as the average response of a hypothetical population of "silicon samples" (depicted as robots). For instance, if the LLM agrees with the statement "Eighty percent of success is showing up" with 90% probability, we interpret this as 90% of the silicon samples would agree with this statement. This gives rise to the *statement*-level metric of common sense which is used to measure the correlation between the human (*A*) and silicon sample (*C*) populations.

model could appear just as dubious. The challenge to evaluation, then, is in reconciling the aim for an informative metric with the pluralism in human judgment that permits almost no correct labels.

In the human sciences, commonsense knowledge is treated not as a set of universally irrefutable truths, but as a system of beliefs that are mutually upheld by people in a community (28). Thus, a person has common sense only if what they believe agrees with whatever their community collectively holds. Inspired by this relativist view, we address the lack of ground-truth labels by treating LLM evaluation as an empirical exercise. Conceptually, we start with a population—of humans and of LLMs alike—and collect their judgments toward a number of statements, from which consensus arises. Common sense is then measured by the degree to which members of this population agree with one another about these statements. In the present work, we elaborate this notion of *collective agreement* via two major but logically independent uses of an LLM, described in Figure 1.

First, an LLM can be treated as an independent survey respondent that is evaluated on an individual basis (Figure 1*B*). Under this agentic paradigm (29, 30), this respondent must both subjectively agree with the majority opinion of those around it (Figure 1*A*), and accurately predict this opinion regardless of what it subjectively holds. These two signals are combined to measure the commonsense competence of every individual respondent, including the LLM. The results show that both humans (N = 2,046) and models (N = 34) vary significantly in this respect. While the highest-ranked model is rated as commonsensical as 64.5% of recruited human participants, over two-thirds of LLMs are placed below the human median. Surprisingly, we find that smaller, open-weight models like Mistral-7B and Flan-T5-XXL are in fact comparable or even more competitive than larger, proprietary models like GPT-4o or Claude 3 Opus.

Second, viewed as a summarizer of social and cultural knowledge (31), an LLM can be evaluated by how well it reproduces aggregated human opinion via simulating a "silicon sample" collective (32–38). In this hypothetical population, silicon samples (depicted as robots in Figure 1*C*) provide ratings just like humans, and common sense is measured for each statement based on how widely agreed upon it is. We find that statement scores are modestly correlated between the two populations (Figure 1*A* and *C*); Pearson's r up to $.43$, compared to the average human internal reliability of $.60$. Some groups of silicon samples, in addition, exhibit traceable qualitative differences from humans; for instance, the population simulated by Gemini Pro 1.0 overwhelmingly associates common sense with figures of speech, while humans tend to prefer statements using only simple, literal language.

Prior work has expressed numerous concerns with current practices in AI commonsense evaluation. For example, as a result of large-scale crowdsourcing, some benchmarks contain very noisy human labels (39) or include many semantically incoherent stimuli (40)—thereby casting doubt on the precision of their performance metrics. Our argument here, however, is that not only can humans be noisy in their judgments, they may in fact hold conflicting beliefs about what is self-evident. Developing human-like AI requires explicitly acknowledging this pluralism, incorporating it in benchmarking domains where ground-truth labels likely do not exist, and appropriately describing what human performance is (6, 27). We address these by contributing a new framework to evaluate commonsense knowledge in LLMs that is grounded in its social basis, *i.e.*, in correspondence with real humans on a large scale. From an AI alignment perspective, this bottom-up framework allows for a fine-grained, empirical assessment of intelligent machines, especially as they are situated in highly diverse social contexts where questions of cultural awareness often arise (41).

Overview

We use a dataset introduced in ref. 26 which contains N = 4,407 statements taken from seven sources, including AI corpora and direct elicitations by online participants. Humans (N = 2,046) were shown a statement and asked to indicate (a) whether they agreed with it and (b) whether they thought most other people would agree with it. For example, in Figure 1*A*, 89% of humans agreed with the statement "An accountant is good at mathematics and logic," while 100% of the same people believed that others would agree with it. Each participant was assigned 50 statements selected at random, and on average every statement was labeled by 23 people. We determine a statement's human majority rating—agree or disagree with it, in response to question (a)—as the judgment that was held by at least half of those who were assigned to rate it (see the *Methods* and *Supplementary Information*, Section A).

We examine 34 autoregressive LLMs including proprietary, frontier models such as GPT-4, Claude 3, Gemini Pro 1.0, and Mistral-Large, as well as popular open-weight models such as LLaMA-2/3 and Falcon. Among the open-weight LLMs, the smallest model has 80M parameters, while the largest has 180B. The full list can be found in *Supplementary Information*, Table B.1. For each statement, we prompt a model with the same questions (a) and (b) above and record the probabilities with which it generates "yes" and "no" tokens (see the "Avg." column in Figure 1*C*). When binary answers are called for—such as in Figure 1*B* where the LLM is treated as an individual agent—we choose the answer that is associated with the higher probability. Otherwise, these probabilities are interpreted as the frequencies with which the simulated silicon samples (Figure 1*C*) would answer the given questions with a "yes" or "no." The *Methods* and *Supplementary Information*, Section B provide more information and precise definitions.

LLMs as Independent Survey Respondents

Upon its release, GPT-4 was demonstrated to score above 90% of the Uniform Bar Examination takers (42). This evaluation is notably unambiguous because exam questions are designed to have correct answers—and humans and machines can be objectively measured by how accurately they select these answers. Unfortunately, as we have argued above, this does not apply to common sense: what appears correct to someone may not be so to another (27).

Here we describe an alternative LLM evaluation strategy, previously illustrated in Figure 1*B* and now in more detail in Figure 2*A*, that can overcome this ground-truth deficiency. Since there is no guarantee of a correct answer, the common sense of humans and models should simply be measured by how much they agree with one another.

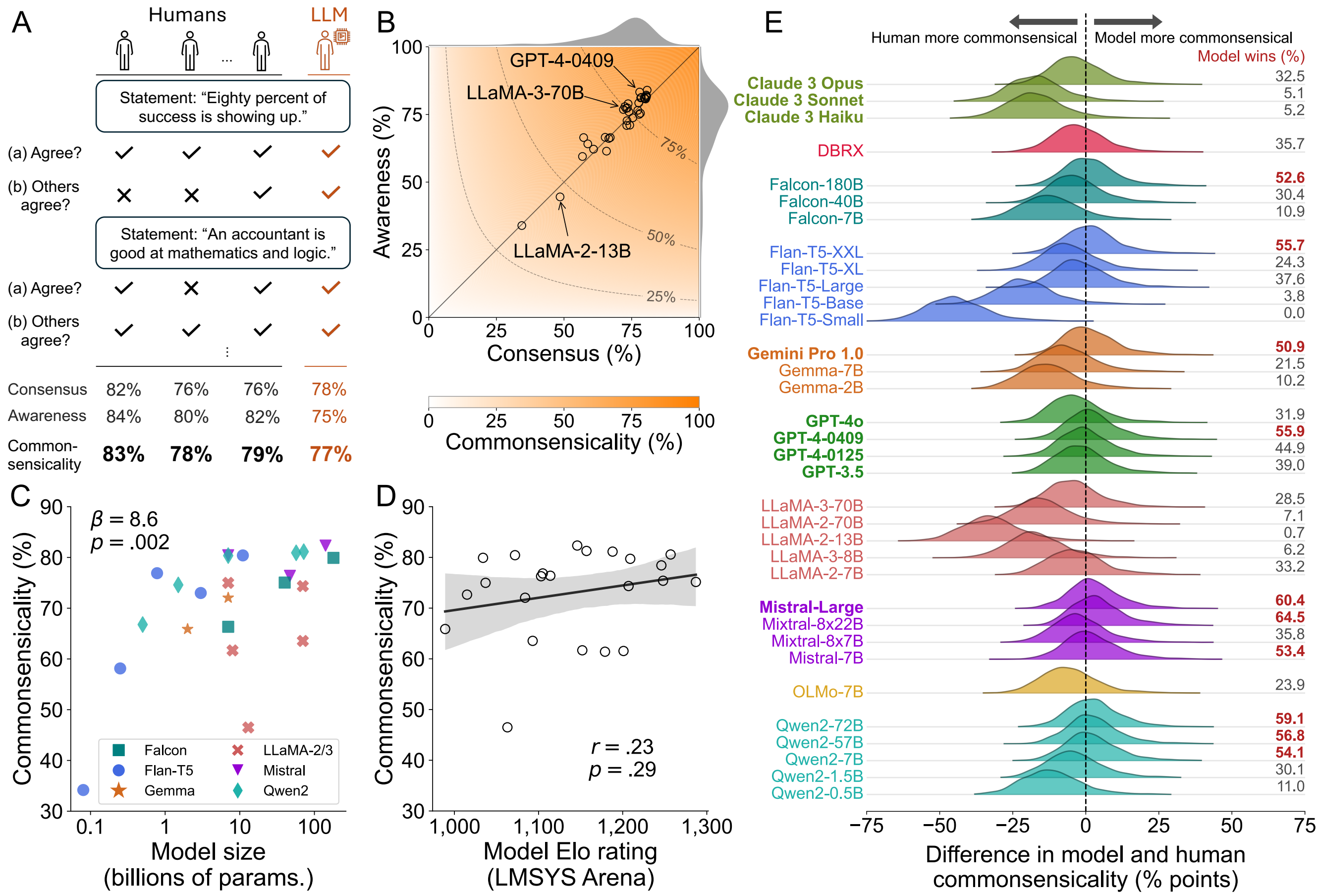


Figure 2. Individual-level commonsensicality of large language models. Panel (*A*) shows conceptually how individual commonsensicality, defined for every human and model, is calculated based on their judgments of each candidate statement. Panel (*B*) shows each model's consensus and awareness scores. The level curves depict combinations of consensus and awareness that produce three different values of commonsensicality scores: 25%, 50% and 75%. Panel (*C*) shows the relationship between a model's commonsensicality and its size, measured by the number of trainable parameters. Here, we only select six model families that each have at least two models of which we know the sizes. Also illustrated are the regression coefficient (β) and its two-sided p-value, estimated using a linear mixed-effect model predicting commonsensicality using an LLM's (log-)size, grouped by model family, such as Flan-T5. Panel (*D*) shows the relationship between a model's commonsensicality and its Elo rating on the LMSYS chatbot arena. Only 23 models with an Elo rating are shown in this figure. Pearson correlation r and its two-sided p-value are illustrated. Correlation is also displayed by the best-fit line and a 95% CI for the regression estimate (using 1,000 bootstrapped samples). Panel (*E*) compares commonsensicality between humans and LLMs. The x-axis represents the percentage-point difference in commonsensicality between a model and a person, where a positive difference indicates the model is more commonsensical. The y-axis represents the kernel density of this difference. The "model wins" column to the right is the frequency with which a model is judged more commonsensical than a person, which equals the area under the density curve to the right of the vertical dashed line at 0. Closed-source models' names are in bold.

In particular, we view each LLM as an independent survey respondent who, like any other human participant, provides binary ratings for every statement. The model's common sense is determined via two signals. First, does its judgment of a statement, via question (a), coincide with the human majority rating? Second, is its prediction of the human majority, via question (b), correct? Averaged over all $N = 4{,}407$ statements, these signals respectively give us the model's *consensus* and *awareness* scores. We take the geometric average of these two scores and call the result the model's *commonsensicality*. Figure 2*B* presents these three scores for all 34 models. See the *Methods* and *Supplementary Information*, Sections A and C and Table C.1 for precise calculations.

Intuitively, a model with high commonsensicality must both agree with the human majority (leading to high consensus) and accurately predict what most people think, regardless of its own judgment (high awareness). For example, in Figure 2*A*, the statement "Eighty percent of success is showing up" received ratings from 21 human participants, 12 of whom (57%) agreed with it. To have high commonsensicality, the model is expected both to agree with this statement and to predict that most people would agree with it. At one extreme we have a maximum score of 100%, where this always happens. A lower commonsensicality score could be driven by a model's lower consensus, awareness, or both. For instance, Claude 3 Opus disagrees with this statement, thereby lowering its consensus score, but it correctly predicts that most people would agree with it, thereby raising its awareness score.

The commonsensicality score achieves two important goals. First, it requires no prior ground truth for any candidate statement: what is "true" is entirely determined by what people believe. Second, it is calculated identically for humans and LLMs, since they perform the same rating task. This allows us to make empirically equivalent comparisons between them, as will be shown shortly. Our metric thus bears resemblance to the concept of cultural competence (43) or cultural consonance (44)—the propensity of an individual to know what their society collectively deems appropriate—which is found to vary among that society's members.

We observe in Figure 2*B* that most models lie close to the diagonal line, implying their consensus and awareness are roughly the same. All three scores are left-skewed; for instance, the average commonsensicality score among models is 71.9% (SD = 10.6). The most commonsensical LLM, with a score of 82.3%, is Mixtral-8x22B, Mistral AI's open-weight model based on the mixture-of-expert architecture with 141B parameters. Other top-performing models include both open- and closed-source LLMs such as Mistral-Large (81.3%, closed), Qwen2 (72B: 81.1%, 57B: 80.9%, open), GPT-4 (80.6%, 0409 version, closed), and Flan-T5-XXL (80.4%, open). The least commonsensical models also include open-weight and closed-source models such as Flan-T5 (Small: 34.2%, Base: 58.1%, open), LLaMA-2-13B (46.5%, open) and Claude 3 (Haiku: 61.4%, Sonnet: 61.6%, closed).

It is generally remarked that within the same LLM family (such as LLaMA), a model's benchmarking performance tends to increase with its size, or number of trainable parameters—a phenomenon dubbed the "scaling law" (45). We also find the same pattern on the commonsensicality test. Figure 2*C* depicts this relationship for 23 models belonging to 6 families in our collection, each of which contains at least 2 models whose sizes are known. As can be noticed in this figure, this correlation is the most pronounced for Flan-T5 and the least for Mistral families. It should be noted, however, that the scaling law does not only apply to model size; another significant factor is the quantity and quality of their training data, which might explain the large variance in commonsensicality among models of roughly 7 billion parameters, such as LLaMA-2/3, Qwen2, and Gemma. To quantify this relationship while accounting for heterogeneity across model families, we perform a mixed-effect regression analysis with commonsensicality as the outcome, the logarithm of the model size as the fixed effect, and the model family as the random effect. The result reveals that, on average, a ten-fold increase in an LLM's size is associated with an 8.6 percentage-point increase in its commonsensicality score ($p = .002$, 95% CI: [3.4, 13.4]; see the *Methods* and *Supplementary Information*, Section C).

Interestingly, an LLM's commonsensicality does not seem to correlate with its general appeal to humans. Figure 2*D* depicts the relationship between a model's commonsensicality and its Elo rating on the LMSYS Arena benchmark (46). This rating is calculated based on human preference signals, where online users interacted with a pair of randomly chosen models and decided which one they preferred. We choose this benchmark because it is based on real, on-the-fly interactions between LLMs and humans, which likely involve a lot of real-life commonsense reasoning unlike other static benchmarks. If a model scores 100 points higher than another, then the former is preferred to the latter about 64% of the time. See the *Methods* and *Supplementary Information*, Section C for more detail. Surprisingly, we do not find any significant correlational evidence between the commonsensicality and Elo scores; Pearson's $r(21) = .23$, $p = .29$, 95% CI: [-.20, .59]. For example, with a large gap in Elo rating of 141, GPT-4-0125 is preferred to GPT-3.5 about two-thirds of the time by humans when pitched side by side. Yet their commonsensicality scores are very similar, at 78.4% and 76.8%, respectively.

Based on their heterogeneous ratings, humans also vary significantly in their commonsensicality scores; see *Supplementary Information*, Figure A.2 for the human version of Figure 2*B*. Hence, there is no singular level of "human performance" to which LLMs can be compared, unlike existing AI benchmarks. Instead, we examine where exactly in this distribution of human scores each model is positioned. Note that during data collection, every human respondent was only tasked with labeling a random subset of 50 statements, whereas each model manages to label all $4,407$ of them. For fairness, when comparing each LLM with every human participant, we restrict the calculation of commonsensicality to the subset of 50 statements that the participant was asked to label. The *Methods* and *Supplementary Information*, Sections A and C provide more detail on this calculation.

Figure 2*E* presents the result of this comparison. The x-axis is the percentage-point difference between a model's and a human's commonsensicality scores; a positive difference indicates that the model is more commonsensical than the human. The y-axis depicts the estimated density of this difference across all $2,046$ humans. Therefore, the area under this curve, from 0 onwards, represents the model's percentile in the distribution of human commonsensicality scores—which is also reported in the "model wins" column.

Relative to humans, most LLMs are modest in their individual-level common sense. Over two-thirds of models (24 out of 34) are placed below the human median, meaning they would be judged less competent than a participant chosen at random. For instance, Claude 3 Opus, a frontier model, is only comparable to about a third of humans. The rest of these LLMs, which are indeed above the human median, include both closed-source models like GPT-4 (0409 version), Gemini Pro 1.0, and Mistral-Large, and open-weight models such as Falcon-180B, Mixtral-8x22B, and Flan-T5-XXL. The highest-ranked model according to this metric is Mixtral-8x22B, which is rated above 64.5% of human participants. Most surprisingly, Flan-T5-XXL is an LLM that was probably trained on orders of magnitude less data than today's models, yet it is comparable to GPT-4-0409 and even ranked higher than Falcon-180B, a model that is about 16 times its size.

The LLaMA family is also an interesting case. No LLM within this family manages to be more commonsensical than a third of humans. Moreover, even though the scaling law is observed in other model families, this is not the case for LLaMA-2. The highest-scoring model turns out to be the smallest variant, LLaMA-2-7B, with a winning rate of 33.2%. This rate drops to almost zero for the 13B version before coming back to 7.1% for the 70B variant. This observation is not even consistent with the inverse scaling law (47, 48), where models are expected to be *worse* in some metric as they grow in size. We therefore suspect that the training of LLaMA-2 models may not have been uniform, especially in terms of training datasets.

In summary, the commonsensicality score presented here measures an LLM's commonsense competence regarding its agreement with other humans. Unlike existing AI benchmarks, this evaluation is bottom-up: it is derived from the consensus of real humans as opposed to a researcher-prescribed correct label. Our metric can therefore be flexibly adapted to any other population of interest to analyze an LLM *relative* to that population. Moreover, the commonsensicality metric formalizes the notion of human performance which, as we remarked earlier, displays a non-trivial variation. This is also in contrast with current benchmarks, as human performance is often thought to be constant and (close to) perfect. As a result, we do not interpret commonsense intelligence as an LLM's ability to pass a rigid, predefined accuracy threshold, but rather we analyze this model by locating it in the heterogeneity of human competencies. Based on this, we show that smaller, open-weight models can be as competitive as their larger, proprietary, frontier counterparts.

LLMs as Aggregators of Human Opinion

We have analyzed one aspect of collective agreement by viewing LLMs as survey takers and comparing them with humans based on their (binary) responses. Here, we focus on another aspect of common sense in these models, one which is tied to their data-driven, generative nature. In particular, since LLMs synthesize the viewpoints of as many people as possible via their training data, these models can be considered a "distilled form of crowdsourcing" (49). In other words, when participating in a survey or study, the models can provide responses similar to those of the average human (37, 50, 51). A growing literature in the social sciences has shown that LLMs can be used to reproduce people's public opinion (33, 52, 53), voting preferences (32), moral judgments (34) and economic decisions (54, 55).

Similarly, if commonsense knowledge is captured widely in LLMs' training data, it may also be reflected in these models with reasonable fidelity. We clarify this notion via the following evaluation setting. Imagine a hypothetical society of "silicon samples" (32), previously illustrated in Figure 1*C* and now in more detail in Figure 3*A*. For each statement, every individual in this society is tasked with answering the same questions (a) and

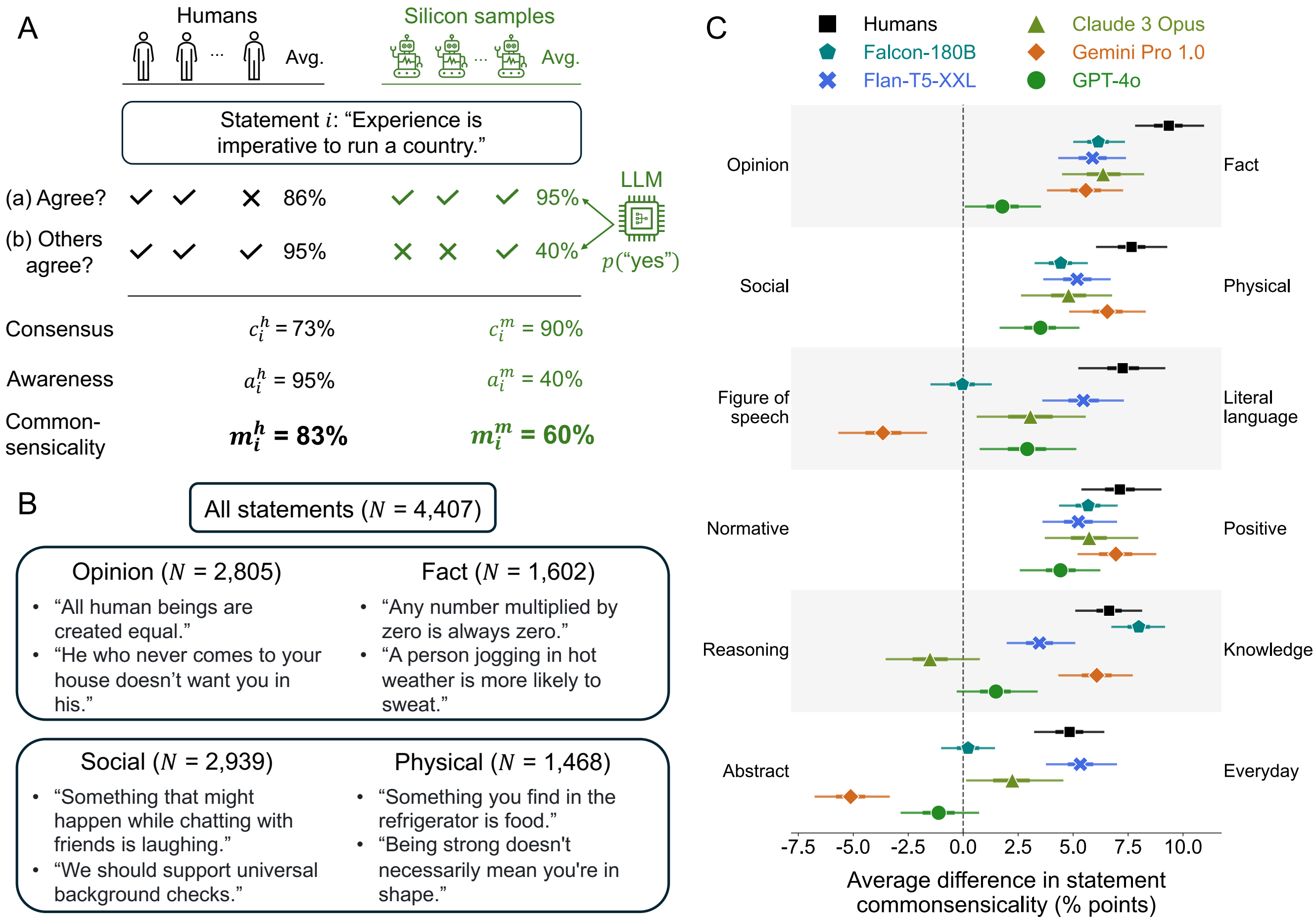


Figure 3. Commonsensicality of statements in different populations of raters. In panel (*A*), we depict two populations: one consisting of real humans and one of silicon samples generated by repeatedly sampling the responses of GPT-3.5. For each population and every statement i, we measure two quantities: how close participants in that population are to a unanimous judgment of the statement (consensus, c_i^h and c_i^m) and how well they can predict this majority opinion (awareness, a_i^h and c_i^m). The statement's commonsensicality, m_i^h and m_i^m, is the geometric average of its consensus and awareness scores. For example, the statement "Experience is imperative to run a country" is $m_i^h = 83\%$ commonsensical according to humans, but only $m_i^m = 60\%$ commonsensical according to GPT-3.5-simulated silicon samples. Panel (*B*) illustrates some features of these statements in our corpus. Every statement is labeled as either an objective fact (N = 1,602) or a subjective opinion (N = 2,805); to describe either the physical world (N = 1,468) or social reality (N = 2,939). In total, there are six such dichotomies, which are described further in the *Methods* and *Supplementary Information*, Section D. Panel (*C*) shows the difference in statement score within each dichotomy for several populations of raters. For instance, in the human population, each statement i receives a commonsensicality score of m_i^h. The black square in the top row represents the average difference in this score between statements labeled as a fact and those labeled as an opinion. Therefore, to humans, facts are on average 9.38 points more commonsensical than statements. Thick and thin bars depict the 50% and 95% confidence intervals from 1,000 bootstraps.

(b) in the *Overview*. These answers are created by repeatedly sampling an LLM's response to the same prompt, taking advantage of its probabilistic design. Assuming independence, the average response of a silicon sample in the limit is exactly the probability with which the LLM generates the "yes" token to a given question. Based on this signal, illustrated in the "Avg." column of Figure 3*A*, we calculate a commonsensicality score for this statement. Essentially, a higher score indicates that individuals in this population are close to unanimity in their judgment of this statement (thus increasing its consensus), and that they can well predict this judgment in one another (thus increasing its awareness). Precise calculations can be found in the *Methods* and *Supplementary Information*, Section D. Note that this *statement*-level measure is separate from the *individual*-level commonsensicality score (*cf.* Figure 2*A*) that was the subject of the previous section.

Figure 3*A* illustrates a statement (indexed by i): "Experience is imperative to run a country." Of the 22 humans who were assigned statement i, 19 (or 86%) agreed with it, while 21 (or 95%) believed other people would agree with it. This statement accordingly receives a consensus score of $c_i^h = 73\%$, an awareness score of $a_i^h = 95\%$, and a commonsensicality score of $m_i^h = 83\%$; see the *Methods* for the calculation. Thus, within this human group, statement i is 83% commonsensical. On the other hand, in the population of silicon samples generated by GPT-3.5, depicted by robots, 95% of them agree with this statement, while 40% of them believe others would agree with it. By the same calculation, we arrive at statement i's commonsensicality among this population, which is $m_i^m = 60\%$.

Features of Common Sense in Statements between Humans and Silicon Samples

What makes two populations of raters (*e.g.*, humans and silicon samples) similar in their common sense? We propose to analyze the commonsensicality scores for the same set of statements between these groups. First, we look at what types of statements tend to attract high agreement in each population. In particular, we note that our corpus is accompanied by six epistemological features: Each statement was rated to depict either an objective fact or a subjective opinion; to use either literal language or a figure of speech; to be about either an abstract rule or a description of an everyday experience; and so on. Figure 3*B* gives some example statements with respect to two features; the *Methods* and *Supplementary Information*, Section D provide a complete list as well as further examples.

For each feature dichotomy, we compare the two groups of statements separated by that dichotomy in terms of their average commonsensicality scores. The results for humans and five silicon sample populations are shown in Figure 3*C*. For example, according to humans—represented by a black square at the top of the figure—statements rated as facts (*e.g.*, "The Pope is the leader of the Catholic Church") are on average 9.38 points more commonsensical than statements rated as opinions (*e.g.*, "Never go on trips with anyone you do not love") (mean difference (MD) = 9.38, 95% CI: [7.72, 10.93]). For silicon samples generated by Falcon-180B (teal pentagon), this difference is 6.18 points (95% CI: [4.86, 7.45]).

Thus, Figure 3*C* shows that human common sense exhibits very clear tendencies: it strongly favors facts over opinions, descriptions of physical over social realities, literal expressions over figures of speech, *etc.* (26). When examining LLMs, we find that most of these tendencies are preserved by silicon samples. For instance, those generated by Flan-T5-XXL (blue crosses) are almost indistinguishable from humans across all six dichotomies.

The silicon samples simulated by Falcon-180B are also similar to humans except along two dimensions. First, this population displays no significant preference between statements that use literal language (*e.g.*, "A cat doesn't want to get wet") and those employing a figure of speech (*e.g.*, "Rudeness is the weak man's imitation of strength") (MD = 0.03, 95% CI: [-1.51, 1.46]). Second, it also does not favor statements that depict ordinary, everyday experiences (*e.g.*, "A grain of sand is very small") over abstract rules of thumb or aphorisms (*e.g.*, "Morality is just a concept that can change depending on the situation") (MD = 0.26, 95% CI: [-0.87, 1.51]).

The three most popular and capable closed-source LLMs today—Claude 3 Opus, Gemini Pro 1.0 and GPT-4o—all exhibit the same tendencies with humans in four out of six dichotomies. For instance, all three models generate a community that rates statements describing the physical world ("You are likely to find a shirt in closet") as significantly more commonsensical than those about social experience ("Justice without force is powerless, force without justice is tyrannical"). GPT-4o's simulated population (green circle), however, does not differentiate between claims about everyday reality with abstract statements (MD = -1.11, 95% CI: [-2.81, 0.65]). Whereas Gemini Pro 1.0 (orange diamond), like humans, favors statements that contain simple declarative knowledge about the world (*e.g.*, "Plants cannot survive without light") over those that involve logical reasoning (*e.g.*, "If we ask for someone else to explain things they will think about it against their own justification"), Claude 3 Opus (green triangle) and GPT-4o do not display such distinction.

More starkly, we observe that the population generated by Gemini Pro 1.0 significantly diverges from humans in two dichotomies of knowledge. According to this simulated group, statements that use a figure of speech (*e.g.*, "The clash of ideas brings forth the spark of truth") are on average 3.67 points (95% CI: [1.76, 5.57]) more commonsensical than those that use more literal expressions (*e.g.*, "A person doesn't want a low paying job"). Similarly, abstract statements (*e.g.*, "Our greatest misfortunes come to us from ourselves") are 5.17 points (95%

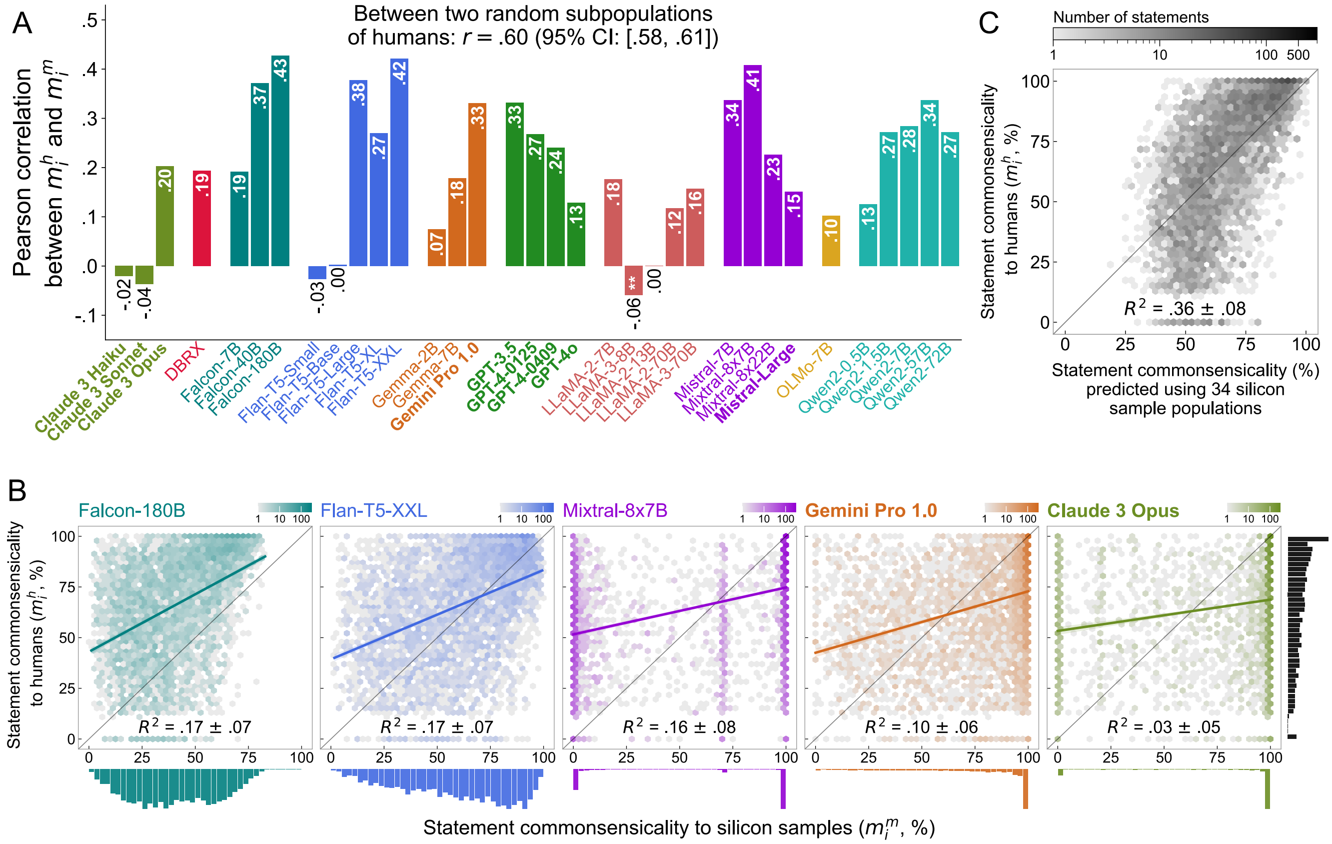


Figure 4. Correspondence between human and silicon sample populations with respect to statement scores. In each population, every statement receives a commonsensicality score. Panel (*A*) shows the Pearson correlation between m_i^h, the score in humans, and m_i^m, the score in silicon samples, for all models. All positive correlations are significant at the $p < .001$ level. All negative correlations are insignificant with $p > .05$, except for LLaMA-3-8B ($p = .003$, depicted with "**" in the figure). p-values are two-sided and Bonferroni corrected. As a baseline, we also display the same correlation between two randomly split subpopulations of humans, which is $r = .60$. The 95% CI is derived from 1,000 such splits. Panel (*B*) expands this correlation for some models. The shade of each hexagon represents its density, *i.e.*, the number of statements within that hexagon. We also illustrate a best-fit line in each plot, predicting m_i^h with m_i^m using a linear regression model. The out-of-sample R^2 for this model (mean and standard deviation) is calculated using 50-fold cross-validation. In panel (*C*), we combine the statement commonsensicality scores m_i^m in *all* 34 silicon sample populations to predict the same score in humans, m_i^h, using a multiple regression model. We also report the out-of-sample R^2 for this model (mean and standard deviation), calculated using 50-fold cross-validation.

CI: [3.36, 6.78]) more commonsensical than statements describing everyday reality (*e.g.*, "The last thing you do when you take a shower is dry off"). Although the reason for these surprising findings is unclear, they may suggest a significant difference in Gemini Pro 1.0's training strategy, such as involving an over-representation of abstract figures of speech in its training set.

Correlation in Statement Commonsensicality between Humans and Silicon Samples

We have seen in Figure 3 that LLM-simulated populations display largely similar qualitative tendencies to humans. Here, we provide a quantitative analysis of this correspondence. More specifically, if the populations of silicon samples and humans are similar with respect to their common sense, then the score of each statement should be the same in both populations. The "fidelity" of a model, therefore, can be calculated as the Pearson correlation between m_i^m and m_i^h for all statements i.

First, to establish a baseline for the internal consistency of human judgments—given their subjectivity—we perform a resampling analysis. We repeatedly split the full human sample (N = 2,046) into two random halves, calculate statement commonsensicality scores for each half, and compute the correlation in these scores between the two. Even among humans, the agreement on statement commonsensicality scores only reaches a moderate level; mean correlation $r = .60$ (95% CI from 1,000 repetitions: $[.58, .61]$).

Figure 4*A* shows the correlation in statement commonsensicality scores between humans and silicon samples generated by every model in our collection. The most faithfully representative models—Falcon-180B, Flan-T5-XXL and Mixtral-8x7B—correlate with people at the $r = .41$ to $.43$ level, well below the baseline of $r = .60$. Unsurprisingly, for some model families like Falcon and Gemini/Gemma, larger or newer models tend to be more representative. For instance, the correlation coefficient rises from .19 to .37 and .43 for Falcon-7B, 40B, and 180B, respectively. On the other hand, for some families like GPT-3.5/4 and Mistral, the opposite is observed. For example, the correlation decreases from .33 for the earliest model, GPT-3.5, down to .13 for the latest variant, GPT-4o. One model, LLaMA-3-8B, has a negative correlation with humans ($r = -.06$, $p = .003$), while five models (Claude 3 Haiku, Claude 3 Sonnet, Flan-T5-Small, Flan-T5-Base, and LLaMA-2-13B) demonstrate no significant correlation at all.

Five of these correlations are magnified in Figure 4*B*, through which we report two important observations. First, a considerable amount of variation in statement commonsensicality in humans (m_i^h) is not accounted for by the same score emerging from silicon samples (m_i^m). Using a linear model to regress m_i^h via m_i^m as the predictor, we depict the best-fit lines as well as their out-of-sample coefficients of determination (R^2) here. Only up to 17% of the variation in statement commonsensicality score according to humans can be explained, which is the case for Falcon-180B and Flan-T5-XXL. This figure is 10%, 3% and 1% for Gemini Pro 1.0, Claude 3 Opus and GPT-4o, respectively (the last of which is not shown in Figure 3*B*).[*]

Second, the best-fit lines, as well as the marginal distributions in Figure 4*B*, suggest some unique patterns among these silicon sample populations. For example, within the population simulated by Falcon-180B, statements are on average less commonsensical than they are to humans (most scores lie above the diagonal perfect-calibration line). In addition, while they have roughly the same correlation with humans, for Mixtral-8x7B statement commonsensicality scores tend to be collapsed to polar values of 0% and 100%, whereas for Falcon-180B and Flan-T5-XXL they tend to spread more evenly.

Similar to Mixtral-8x7B, the tendency of statement commonsensicality to be close to extreme values is common among many models like Gemini 1.0 and Claude 3 Opus, also shown in Figure 4*B*. In other words, statements are either very commonsensical or the opposite according to the ratings by the simulated populations from these models. This is partly a result of LLMs being extremely confident in their answers to our prompts by outputting probabilities close to 0 or 1. (Contrast this with what we observe in humans, depicted in the black histogram to the right of Figure 4*B*.) While the cause of this overconfidence is uncertain, this is likely because these models have gone through extensive instruction fine-tuning and human alignment, two phases of training that tend to collapse their outputs toward singular choices associated with high rewards (56, 57)—thereby sacrificing their distributional representativeness.

Finally, we explore the informativeness of *all* models at once by performing a regression analysis predicting a statement's score in humans using the same 34 scores in all silicon sample populations. As can be seen in Figure 4*C*, taking an ensemble of all models does indeed improve the prediction of statement scores; average out-of-sample $R^2 = .36$.

In summary, this section views every LLM as an aggregator of human knowledge on a large scale, and tests whether the population simulated by this model is an adequate representation of real humans. Once again, the heterogeneity in human commonsense beliefs is evidenced by a moderate split-half reliability of $r = .60$, but the best model-generated silicon samples can only correlate with humans at $r = .43$, well below this baseline. In addition, while a statement's score in a silicon-sample population is predictive of its score in the human population, the majority of the variation in the latter is left unaccounted for by the former.

[*]In *Supplementary Information*, Figure D.1, we illustrate this point further by showing in greater detail the difference between m_i^m and m_i^h. In particular, we calculate the mean absolute error (MAE) between these two scores, which represents the average difference in statement commonsensicality between humans and silicon samples. The highest-ranked model is Flan-T5-XXL, with (the smallest) MAE of 21.41. This means that on average, the commonsensicality score of a statement, according to Flan-T5-XXL's silicon samples, is 21.41 points away from the score of the same statement according to humans. This error is the highest for Falcon-40B, with an MAE of 52.69.

Discussion

One of AI's central objectives is to design machines that behave and interact with humans in meaningful and novel ways (49, 58–60). This requires AI systems to embody human commonsense knowledge and reasoning, and standardized benchmarking has allowed researchers to track developments in this area. In this paper, we argue that the notion of ground truth in benchmarking has likely misguided AI researchers, creating an illusion of machine common sense as LLMs pass a suite of tests (*e.g.*, refs. 61–64). Commonsense knowledge does not assume an objective basis for correctness; what a model "gets right" may be meaningful to some but not so much to others (65, 66). This "one-truth myth" (67) manifests itself in many domains involving subjective interpretation (68–70) such as toxicity detection (71), image classification (72), moral judgment (73–75), rhetoric decoding (76), and even medical relations understanding (67).

In response, researchers have called for treating empirically observed human judgments as a normative standard for AI benchmarking (77). We contribute an evaluation framework that specifically focuses on this aspect, *i.e.*, by measuring collective agreement empirically. Because of the differences in their judgments, human commonsense competence—how much they agree with one another—exhibits a non-trivial variation that is unlikely a result of noise alone. Accordingly, instead of reporting an accuracy score, we treat each LLM as a survey taker and examine its position in this empirical distribution of human competence (Figure 2). We show that LLMs also vary substantially, from being as commonsensical as no participant (Flan-T5-Small) to being placed above about two-thirds of humans (Mixtral-8x22B). Notably, there is correlational evidence that scaling up a model in its size could improve its competence.

Our proposed framework is also compatible with a novel interpretation of an LLM—not as an autonomous, intelligent agent but as a faithful aggregator of social and cultural artifacts (31). Such a system can thus be assessed by how well it reproduces elements of human culture, including commonsense knowledge. Qualitatively, we find that certain characteristics of human common sense, such as a strong preference for facts over opinions, are largely preserved by LLM-simulated silicon samples (Figure 3). However, silicon samples and humans only correlate with each other modestly, and the majority of variation in a statement's commonsensicality within humans remains unexplained (Figure 4). Thus, LLM representation of humans in this respect remains limited.

The current study has two limitations. First, the measures of LLM common sense introduced here are tied to a human population which, in our case, consists mostly of U.S. residents recruited on Amazon Mechanical Turk. It is possible that researchers who focus on a different population or demographic group may reach separate conclusions about these models. This, though, is by design: if one has reason to believe another population diverges significantly from Americans in their cultural beliefs, then the same models should be expected to appear differently in this new context. What this work has not provided, however, is a holistic assessment of LLMs by involving more diverse judgments collected from humans across the globe. Second, it is possible that our corpus of choice has not touched upon certain areas of knowledge posited to be in the domain of common sense. Curating a larger statement collection, guided by a certain taxonomy (*e.g.*, refs. 18, 78), can address this potential incomprehensiveness.

The foregoing suggests two directions worthy of investigation in future work. First, it remains unclear whether there are systematic characteristics in a human population that allow one to make accurate predictions about an LLM's common sense relative to that population. For instance, could the performance of GPT-4o, which ranks below two-thirds of humans in the current sample, be expected to change dramatically if it were considered alongside only American college students? Future work, in examining these characteristics, can help propose elements of knowledge that are (near-)unanimously shared, or, when they are only locally shared, how exactly they vary from one group to another (79).

Second, the fact that statement-level commonsensicality correlates poorly between silicon samples and humans in some cases may not strike as surprising. The fidelity of an LLM-simulated population could be improved with more fine-grained descriptions of individual personas (32, 53, 80–82), although this approach is not without caveats (83–85). (In our work, though, silicon samples are simply identical copies of one another, which allows for a straightforward interpretation of a model's probabilistic output.) For example, past research found that social perceptiveness—the ability to “read” other people’s emotions—can reliably predict commonsensicality (26). Silicon samples, therefore, could be enhanced by having this variable encoded in their personas when engaging in role-playing (30). In addition, future work could examine certain controllable settings of these models—especially their training data and human alignment strategies—that may contribute to observable biases that differentiate these two populations, one real and the other simulated.

Methods

We use the dataset in ref. 26 which contains N = 4,407 statements sourced from seven domains: the news media via Google News (N = 290), political campaign emails during U.S. elections (N = 668), AI corpora like ConceptNet (86, N = 581) and ATOMIC (87, N = 697), aphorisms taken from books (N = 709), and statements elicited by online participants—either via completing a short sentence (N = 630) or in response to a prompted domain of knowledge (N = 832). See *Supplementary Information*, Table A.1 for some examples. Most statements are short with a median of 11 words. After collection, these statements were labeled by Amazon Mechanical Turk workers (N = 2,046 in total) who answered the following questions: (a) “Do you agree with this statement?” and (b) “Do you think most other people would agree with this statement?” Each participant was tasked with labeling 50 randomly chosen questions, and on average each statement received ratings from 23 participants.

Large Language Models and Their Ratings

We make use of 34 widely used instruction-finetuned large language models. These include open-weight models (*i.e.*, those whose weights are openly accessible) and closed-source models. The full list can be found in *Supplementary Information*, Table B.1. All open-weight models are loaded via the Hugging Face library in Python. For closed-source models, we directly use the APIs provided by their creators.

For every statement, we ask each LLM the same two questions (a) and (b). Each question is asked in a separate chat session, in order to eliminate any influence of the chat history on an answer. We design each prompt so that the model’s answer is expected to start with a definitive “yes” or “no.” Exact prompts and conversation settings are the same for all models and can be found in *Supplementary Information*, Section B.

To generate an answer, the model uses a parametric probability distribution over its vocabulary to sample new tokens, one at a time. We extract the probabilities of the tokens “yes” and “no” and discard all other tokens, then rescale these two probabilities so they add up to one. For all open-weight models, this distribution can be accessed directly. For GPT-3.5/4, we can access the probabilities for up to the top 20 tokens. For Gemini, Claude 3 and Mistral-Large, we perform repeated sampling, where we ask a model the same questions multiple times and report the empirical frequencies of their generated answers. See *Supplementary Information*, Section B.

The probability that the model answers “yes” to question (a) is called its rating distribution; see *Supplementary Information*, Equation (12). This can be considered as the model’s inherent uncertainty in agreeing with a statement. When binary decisions are called for, we choose the answer—“yes” or “no”—that is associated with the higher probability, also called the “argmax” of the model’s rating distribution. These prompting settings were preregistered prior to data collection on AsPredicted, project number 162475 (88).

Measuring LLMs’ Individual Commonsensicality

We adopt the intuition in ref. 26 and calculate an LLM’s degree of common sense as if it were a real human participant (see Figure 2*A*). In particular, to have common sense, the model must agree with the human majority opinion, and correctly predict this consensus irrespective of its subjective opinion. These two criteria are reflected in the model’s answer to questions (a) and (b) above, respectively.

Precisely, for every statement let the human majority rating (agree or disagree) be the opinion held by at least half of the participants. If the model’s answer to question (a) (whether it agrees with the statement) coincides with this majority rating, this counts as a correct answer. The accuracy of the model, averaged over all 4,407 statements, is called its *consensus* score, which depicts how often its subjective rating agrees with the human majority. Similarly, the same accuracy with respect to the model's answer to question (b) (whether it thinks most people would agree with this statement) is called its *awareness* score, representing how often it predicts the position of human majority, notwithstanding how it subjectively rates a statement previously via question (a). We combine a model's consensus and awareness scores by taking their geometric average into its *commonsensicality* score. See *Supplementary Information*, Equations (16) to (18). These three scores, each ranging between 0 and 100%, are presented in Figure 2*B* above.

To analyze the relationship between a model's size and its commonsensicality score, we choose 23 models from 6 model families in our collection: Falcon, Flan-T5, Gemma, LLaMA-2/3, Mistral and Qwen2. These are families that contain at least 2 models whose sizes are available to us. The models are presented in Figure 2*C*. We then perform a linear mixed-effects regression analysis with commonsensicality as the outcome, model size as the fixed effect and model family as the random effect. A detailed setting can be found in *Supplementary Information*, Equation (19).

To investigate the relationship between an LLM's general performance on a popular benchmark and its commonsensicality, we extract the LMSYS Arena Elo scores (46), last updated on July 8, 2024. Essentially, LMSYS tests a total of 114 LLMs side by side and uses crowd-sourced human ratings to rank these models. A participating user interacts with two randomly chosen LLMs at the same time, asking them the same question and receiving answers from both. Then, the user indicates which model is better. From approximately 1.5 million such comparisons, models are assigned an Elo score that depicts their relative performance. More detail can be found in *Supplementary Information*, Section C, especially Equation (20). We are able to find Elo scores for 23 models, and in Figure 2*D*, we report the Pearson correlation between a model's Elo score and its commonsensicality.

Since the calculation of model commonsensicality is exactly how human participants are measured on their common sense; see *Supplementary Information*, Equations (9) to (11), this allows us to compare every LLM against every human directly. One major difference is that while every model can label all 4,407 statements, every person was only asked to label a random subset of 50 statements. To make a fair comparison, for each LLM-human pair, we restrict the calculation of their commonsensicality scores to the 50 questions that the participant previously rated. More detail can be found in *Supplementary Information*, Section C. Figure 2*E* depicts how often a model is judged to be more commonsensical than a human.

Analyzing LLMs as Simulators of Silicon Sample Populations

Recall that when we ask an LLM questions (a) and (b) above, we record the probability with which it answers "yes" or "no." In the limit, this probability can be interpreted as the frequency of answers by the silicon samples simulated by this model. For example, if an LLM answers "yes" with 70% probability, then among 1,000 silicon samples that are generated by this model, 700 of them are expected to say "yes."

For each population (humans or silicon samples), we analyze common sense on the statement level. Here we describe this calculation with respect to the human population, more detail of which can be found in *Supplementary Information*, Section A. Let $d_i^{h,a}$, called the human rating distribution, be the proportion of people that indicated they agreed with statement i, via question (a). The majority opinion for this statement is then $\texttt{majority}_i^h = \mathbb{1}\left[d_i^{h,a} \geq 0.5\right]$, which is 1 if at least half of the people agreed with it, and 0 otherwise. The consensus score measures how far people are to a unanimous opinion: $c_i^h = 2 \times |d_i^{h,a} - 0.5|$. It is 0 when exactly half of the people agree with it, and 1 when everyone agrees or disagrees with it. The awareness score measures how accurately people predicted the majority rating via question (b). If the $\texttt{majority}_i^h = 1$, then the statement's awareness, a_i^h, is the proportion of people that predicted most others would agree; otherwise, it is the proportion that predicted most others would disagree. The commonsensicality score is defined as $m_i^h = \sqrt{c_i^h \times a_i^h}$. We denote the same scores for the model-simulated population with the superscript m instead of h, such as m_i^m.

For example, in Figure 3*A*, we observe that 86% of humans agree with the statement, "Experience is imperative to run a country," while 95% of the same people thought most others would agree with it. Hence, $d_i^{h,a} = 0.86$, which means the majority rating is 1, or "agree." The consensus score is thus $c_i^h = 2 \times |0.86 - 0.5| = 0.73$, and the awareness score is $a_i^h = 0.95$, because 95% of the people correctly predicted that it would receive an "agree."

The commonsensicality score is, therefore, $m_i^h = \sqrt{0.73 \times 0.95} = 0.83$. The same scores ($c_i^m$, a_i^m and m_i^m) are identically defined for the population of LLM-based silicon samples.

Each statement in our corpus was labeled using six epistemological dichotomies. Online participants were asked to rate whether a statement depicts a *fact* (something that can be objectively demonstrated to be true) *vs.* an *opinion* (something that might be held true by some but not by others); describes *physical* reality (objective features of the world according to, say, physics and biology) *vs. social* beliefs, perceptions and rules that govern human experience; uses *literal language* that means exactly what it says *vs.* a *figure of speech* such as an aphorism or metaphor; conveys *positive* empirical regularities in the world (*e.g.*, "Hot things will burn you.") *vs.* a *normative* judgment, belief, value or convention (*e.g.*, "Treat others how you want them to treat you."); conveys declarative *knowledge* about the world *vs.* a *reasoning* that involves both knowledge and logic; and depicts *everyday* experiences of humans *vs. abstract* rules and regularities that must be synthesized from experience. See *Supplementary Information*, Section D for more detail.

In Figure 3*C*, we calculate the reported differences as follows. First, we separate the entire corpus of N = 4,407 into two categories: those that depict a fact and those that portray a personal opinion. Each statement, again, has a commonsensicality score m_i^h (assuming we are considering the human population). We calculate the average difference in this score between the fact-like statements and the opinion-like statements. To quantify uncertainty, we perform this calculation via 1,000 random bootstraps. The mean difference and the 95% bootstrapped CI are depicted in the figure. This applies to other dichotomies as well as the populations of silicon samples by LLMs.

In Figure 4*A*, we compute the Pearson correlation between m_i^h and m_i^m—the same commonsensicality scores for every statement within two populations: humans *vs.* silicon samples. All two-sided p-values are Bonferroni-corrected. In Figure 4*B*, we show this result in more detail via scatter plots for five models: Falcon-180B, Flan-T5-XXL, Mixtral-8x7B, Gemini Pro 1.0 and Claude 3 Opus. We also fit a linear model, predicting m_i^h using m_i^m, and plot the best-fit line within these plots. The out-of-sample coefficient of determination, R^2, for this linear model is also depicted. Finally, in Figure 4*C*, we combine all 34 statement scores m_i^m in 34 silicon sample populations to predict the same score in humans, m_i^h.

References

1. A. S. Gordon, J. R. Hobbs, *A Formal Theory of Commonsense Psychology: How People Think People Think* (Cambridge University Press, 2017).
2. J. McCarthy, "Programs with Common Sense" in *Mechanisation of Thought Processes* (1959), vol. 1, pp. 75–91.
3. M. Minsky, *The Emotion Machine: Commensense Thinking, Artificial Intelligence, and the Future of the Human Mind* (Simon & Schuster, 2007).
4. E. Davis, G. Marcus, Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence. *Commun. ACM* **58**, 92–103 (2015).
5. R. J. Brachman, H. J. Levesque, *Machines like Us: Toward AI with Common Sense* (MIT Press, 2023).
6. M. Kejriwal *et al.*, Can AI Have Common Sense? Finding out Will Be Key to Achieving Machine Intelligence. *Nature* **634**, 291–294 (2024).
7. F. Petroni *et al.*, "Language Models as Knowledge Bases?" in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (2019), pp. 2463–2473.
8. J. Davison, J. Feldman, A. Rush, "Commonsense Knowledge Mining from Pretrained Models" in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (2019), pp. 1173–1178.
9. Z. Jiang, F. F. Xu, J. Araki, G. Neubig, How Can We Know What Language Models Know? *Transactions Assoc. for Comput. Linguist.* **8**, 423–438 (2020).
10. Z. Jiang, J. Araki, H. Ding, G. Neubig, How Can We Know *When* Language Models Know? On the Calibration of Language Models for Question Answering. *Transactions Assoc. for Comput. Linguist.* **9**, 962–977 (2021).
11. X. L. Li *et al.*, "A Systematic Investigation of Commonsense Knowledge in Large Language Models" in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (2022), pp. 11838–11855.
12. Y. Choi, The Curious Case of Commonsense Intelligence. *Daedalus* **151**, 139–155 (2022).
13. S. Bubeck *et al.*, Sparks of Artificial General Intelligence: Early Experiments with GPT-4. arXiv [Preprint] (2023). https://arxiv.org/abs/2303.12712 (Accessed 2024-10-10).

14. M. Binz, E. Schulz, Using Cognitive Psychology to Understand GPT-3. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2218523120 (2023).
15. Q. Mei, Y. Xie, W. Yuan, M. O. Jackson, A Turing Test of Whether AI Chatbots Are Behaviorally Similar to Humans. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2313925121 (2024).
16. J. W. A. Strachan *et al.*, Testing Theory of Mind in Large Language Models and Humans. *Nat. Hum. Behav.* **8**, 1285–1295 (2024).
17. M. Kosinski, Evaluating Large Language Models in Theory of Mind Tasks. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2405460121 (2024).
18. E. Davis, Benchmarks for Automated Commonsense Reasoning: A Survey. *ACM Comput. Surv.* **56**, 1–41 (2024).
19. A. Schutz, Common-Sense and Scientific Interpretation of Human Action. *Philos. Phenomenol. Res.* **14**, 1–37 (1953).
20. P. L. Berger, T. Luckmann, *The Social Construction of Reality: A Treatise in the Sociology of Knowledge* (Anchor Books, 1966).
21. R. K. Merton, *Social Theory and Social Structure* (Free Press, 1968).
22. G. J. Fletcher, Psychology and Common Sense. *Am. Psychol.* **39**, 203–213 (1984).
23. F. van Holthoon, D. R. Olson, eds. *Common Sense: The Foundations for Social Science*, Sources in Semiotics (University Press of America, 1987).
24. S. A. Rosenfeld, *Common Sense: A Political History* (Harvard University Press, 2011).
25. D. J. Watts, *Everything Is Obvious: Once You Know the Answer* (Crown Business, 2011).
26. M. E. Whiting, D. J. Watts, A Framework for Quantifying Individual and Collective Common Sense. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2309535121 (2024).
27. L. Ying *et al.*, On Benchmarking Human-Like Intelligence in Machines. arXiv [Preprint] (2025). https://arxiv.org/abs/2502.20502 (Accessed 2025-03-12).
28. C. Geertz, Common Sense as a Cultural System. *The Antioch Rev.* **33**, 5–26 (1975).
29. J. Andreas, "Language Models as Agent Models" in *Findings of the Association for Computational Linguistics: EMNLP* (2022), pp. 5769–5779.
30. M. Shanahan, K. McDonell, L. Reynolds, Role Play with Large Language Models. *Nature* **623**, 493–498 (2023).
31. H. Farrell, A. Gopnik, C. Shalizi, J. Evans, Large AI Models Are Cultural and Social Technologies. *Science* **387**, 1153–1156 (2025).
32. L. P. Argyle *et al.*, Out of One, Many: Using Language Models to Simulate Human Samples. *Polit. Anal.* **31**, 337–351 (2023).
33. S. Santurkar *et al.*, "Whose Opinions Do Language Models Reflect?" in *Proceedings of the 40th International Conference on Machine Learning* (2023), vol. 202, pp. 29971–30004.
34. D. Dillion, N. Tandon, Y. Gu, K. Gray, Can AI Language Models Replace Human Participants? *Trends Cogn. Sci.* **27**, 597–600 (2023).
35. P. Hämäläinen, M. Tavast, A. Kunnari, "Evaluating Large Language Models in Generating Synthetic HCI Research Data: A Case Study" in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023), pp. 1–19.
36. B. J. Jansen, S.-g. Jung, J. Salminen, Employing Large Language Models in Survey Research. *Nat. Lang. Process. J.* **4**, 100020 (2023).
37. I. Grossmann *et al.*, AI and the Transformation of Social Science Research. *Science* **380**, 1108–1109 (2023).
38. E. Durmus *et al.*, "Towards Measuring the Representation of Subjective Global Opinions in Language Models" in *Conference on Language Modeling* (2024).
39. M. Kejriwal, H. Santos, K. Shen, A. M. Mulvehill, D. L. McGuinness, A Noise Audit of Human-Labeled Benchmarks for Machine Commonsense Reasoning. *Sci. Rep.* **14**, 8609 (2024).
40. S. Palta *et al.*, "Plausibly Problematic Questions in Multiple-Choice Benchmarks for Commonsense Reasoning" in *Findings of the Association for Computational Linguistics: EMNLP 2024* (2024), pp. 3451–3473.
41. S. Pawar *et al.*, Survey of Cultural Awareness in Language Models: Text and Beyond. arXiv [Preprint] (2024). http://arxiv.org/abs/2411.00860 (Accessed 2025-04-23).
42. J. Achiam *et al.*, GPT-4 Technical Report. arXiv [Preprint] (2023). https://arxiv.org/abs/2303.08774 (Accessed 2024-03-20).
43. A. K. Romney, S. C. Weller, W. H. Batchelder, Culture as Consensus: A Theory of Culture and Informant Accuracy. *Am. Anthropol.* **88**, 313–338 (1986).
44. W. W. Dressler, *Culture and the Individual: Theory and Method of Cultural Consonance* (Routledge, 2018).
45. J. Kaplan *et al.*, Scaling Laws for Neural Language Models. arXiv [Preprint] (2020). https://arxiv.org/abs/2001.08361 (Accessed 2024-10-23).
46. W.-L. Chiang *et al.*, "Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference" in *Proceedings of the 41st International Conference on Machine Learning* (2024).
47. I. R. McKenzie *et al.*, Inverse Scaling: When Bigger Isn't Better. *Transactions on Mach. Learn. Res.* (2023).
48. S. Lin, J. Hilton, O. Evans, "TruthfulQA: Measuring How Models Mimic Human Falsehoods" in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2022), pp. 3214–3252.

49. J. W. Burton *et al.*, How Large Language Models Can Reshape Collective Intelligence. *Nat. Hum. Behav.* **8**, 1643–1655 (2024).
50. T. Heyman, G. Heyman, The Impact of ChatGPT on Human Data Collection: A Case Study Involving Typicality Norming Data. *Behav. Res.* **56**, 4974–4981 (2023).
51. C. A. Bail, Can Generative AI Improve Social Science? *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2314021121 (2024).
52. J. Kim, B. Lee, AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction. arXiv [Preprint] (2023). https://arxiv.org/abs/2305.09620 (Accessed 2024-10-10).
53. A. C. Kozlowski, H. Kwon, J. A. Evans, In Silico Sociology: Forecasting COVID-19 Polarization with Large Language Models. arXiv [Preprint] (2024). https://arxiv.org/abs/2407.11190 (Accessed 2025-04-06).
54. J. Horton, *Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?*, Working Paper 31122, National Bureau of Economic Research (2023).
55. J. Brand, A. Israeli, D. Ngwe, *Using LLMs for Market Research*, Working Paper 23-062, Harvard Business School (2023).
56. Y. Bai *et al.*, Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv [Preprint] (2022). https://arxiv.org/abs/2204.05862 (Accessed 2024-03-01).
57. S. Kadavath *et al.*, Language Models (Mostly) Know What They Know. arXiv [Preprint] (2022). https://doi.org/10.48550/arXiv.2207.05221 (Accessed 2023-08-01).
58. I. Rahwan *et al.*, Machine Behaviour. *Nature* **568**, 477–486 (2019).
59. K. M. Collins *et al.*, Building Machines That Learn and Think with People. *Nat. Hum. Behav.* **8**, 1851–1863 (2024).
60. M. Tsvetkova, T. Yasseri, N. Pescetelli, T. Werner, A New Sociology of Humans and Machines. *Nat. Hum. Behav.* **8**, 1864–1876 (2024).
61. H. Levesque, E. Davis, L. Morgenstern, "The Winograd Schema Challenge" in *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning* (2012), pp. 552–561.
62. T. Mihaylov, P. Clark, T. Khot, A. Sabharwal, "Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering" in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (2018), pp. 2381–2391.
63. A. Talmor, J. Herzig, N. Lourie, J. Berant, "CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge" in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (2019), pp. 4149–4158.
64. R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, Y. Choi, "HellaSwag: Can a Machine Really Finish Your Sentence?" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (2019), pp. 4791–4800.
65. T.-P. Nguyen, S. Razniewski, A. Varde, G. Weikum, "Extracting Cultural Commonsense Knowledge at Scale" in *Proceedings of the ACM Web Conference 2023* (2023), pp. 1907–1917.
66. J. Huang, D. Yang, "Culturally Aware Natural Language Inference" in *Findings of the Association for Computational Linguistics: EMNLP 2023* (2023), pp. 7591–7609.
67. L. Aroyo, C. Welty, Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Mag.* **36**, 15–24 (2015).
68. C. O. Alm, "Subjective Natural Language Problems: Motivations, Applications, Characterizations, and Implications" in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (2011), pp. 107–112.
69. B. Plank, "The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation" in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (2022), pp. 10671–10682.
70. F. Cabitza, A. Campagner, V. Basile, Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. *AAAI* **37**, 6860–6868 (2023).
71. N. Goyal, I. D. Kivlichan, R. Rosen, L. Vasserman, Is Your Toxicity My Toxicity? Exploring the Impact of Rater Identity on Toxicity Annotation. *Proc. ACM Hum.-Comput. Interact.* **6**, 1–28 (2022).
72. M. Miceli, M. Schuessler, T. Yang, Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision. *Proc. ACM Hum.-Comput. Interact.* **4**, 1–25 (2020).
73. K. C. Fraser, S. Kiritchenko, E. Balkir, "Does Moral Code Have a Moral Code? Probing Delphi's Moral Philosophy" in *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)* (2022), pp. 26–42.
74. T. Sorensen *et al.*, Value Kaleidoscope: Engaging AI with Pluralistic Human Values, Rights, and Duties. *AAAI* **38**, 19937–19947 (2024).
75. L. Jiang *et al.*, Investigating Machine Moral Judgement through the Delphi Experiment. *Nat. Mach. Intell.* **7**, 145–160 (2025).
76. J. Mendelsohn, R. Le Bras, Y. Choi, M. Sap, "From Dogwhistles to Bullhorns: Unveiling Coded Rhetoric with Language Models" in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2023), pp. 15162–15180.
77. F. J. Pelletier, R. Elio, What Should Default Reasoning Be, by Default? *Comput. Intell.* **13**, 165–187 (1997).
78. F. Ilievski *et al.*, Dimensions of Commonsense Knowledge. *Knowledge-Based Syst.* **229**, 107347 (2021).
79. D. Sperber, *Explaining Culture: A Naturalistic Approach* (Blackwell, 1996).

80. J. S. Park *et al.*, Generative Agent Simulations of 1,000 People. arXiv [Preprint] (2024). https://arxiv.org/abs/2411.10109 (Accessed 2024-12-10).
81. H. Jiang *et al.*, "PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits" in *Findings of the Association for Computational Linguistics: NAACL 2024* (2024), pp. 3605–3627.
82. J. R. Anthis *et al.*, LLM Social Simulations Are a Promising Research Method. arXiv [Preprint] (2025). https://arxiv.org/abs/2504.02234 (Accessed 2025-04-06).
83. J. Bisbee, J. D. Clinton, C. Dorff, B. Kenkel, J. M. Larson, Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. *Polit. Anal.* **32**, 401–416 (2024).
84. T. Hu, N. Collier, "Quantifying the Persona Effect in LLM Simulations" in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2024), pp. 10289–10307.
85. A. Wang, J. Morgenstern, J. P. Dickerson, Large Language Models That Replace Human Participants Can Harmfully Misportray and Flatten Identity Groups. *Nat. Mach. Intell.* **7**, 400–411 (2025).
86. R. Speer, J. Chin, C. Havasi, ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. *AAAI* **31**, 4444–4451 (2017).
87. J. D. Hwang *et al.*, COMET-ATOMIC 2020: On Symbolic and Neural Commonsense Knowledge Graphs. *AAAI* **35**, 6384–6392 (2021).
88. T. D. Nguyen, D. J. Watts, M. E. Whiting, Testing Common Sense in Large Language Models. (2024). https://aspredicted.org/XXL_ZPR (Accessed 2024-02-16).

Supplementary Information for
Empirically evaluating commonsense intelligence in large language models with large-scale human judgments

Tuan Dung Nguyen, Duncan J. Watts, and Mark E. Whiting

Corresponding Author: Tuan Dung Nguyen
Email: joshtn@seas.upenn.edu

Table of Contents

A Measuring Common Sense in Humans and Statements

Here we describe the commonsensicality framework by Whiting and Watts (26), whose data we adopt to study common sense in LLMs. Suppose we have a set of $n_s \geq 1$ statements (see Table A.1 for an example) and a set of $n_p \geq 1$ people. (In this case, $n_s = 4,407$ and $n_p = 2,046$.) Given a statement, such as "If you want to play a *[sic]* guitar then you should take lessons," a human participant was asked two questions:

(a) "Do you agree with this statement?"

(b) "Do you think most people would agree with this statement?"

Common sense is measured for both humans and statements. The underlying rationale is that a person with common sense knows what is true (in the sense that their belief coincides with the general human consensus), and also knows what other people believe to be true (in the sense that they can accurately predict what the majority of other people believe). Similarly, a statement is common sense if it is (nearly) unanimously agreed upon by people in this population, who must also accurately predict what the human consensus on this statement is. All of this depends on participants' answers to the two questions above.

To make this precise, define A be the $n_s \times n_p$ matrix such that

$$A_{i,j} = \begin{cases} 1 & \text{if the } j\text{-th participant answered "yes" to question (a) about the } i\text{-th claim} \\ 0 & \text{if the } j\text{-th participant answered "no" to question (a) about the } i\text{-th claim.} \end{cases} \tag{1}$$

In other words, the matrix A contains participants' answers to question (a) above. Similarly, let B be the $n_s \times n_p$ matrix containing answers to question (b) such that

$$B_{i,j} = \begin{cases} 1 & \text{if the } j\text{-th participant answered "yes" to question (b) about the } i\text{-th claim} \\ 0 & \text{if the } j\text{-th participant answered "no" to question (b) about the } i\text{-th claim.} \end{cases} \tag{2}$$

One complication in this study is that not every claim was labeled by everyone, because every participant was required to label only a small number of claims, 50 to be exact. Every (i, j) entry in matrices A and B, therefore, is only valid if the i-th claim was indeed assigned to the i-th participant; otherwise, participant j would have never seen and labeled claim i in the first place, and hence the entry must not be taken into account. Hence, A and B are partially filled matrices.

Define $\Omega_i \subseteq \{1, 2, \ldots, n_p\}$ to be the set of participants who were indeed assigned to rate the i-th statement, and $\Phi_j \subseteq \{1, 2, \ldots, n_s\}$ to be the set of statements that were indeed assigned to the j-th participant. We hence have $|\Phi_j| = 50$, where $|\cdot|$ denotes the size of a finite set.

Before we go on to define commonsensicality, define the *human rating distribution* for each statement i as the proportion of people that indicated they agreed with statement i (*i.e.*, by answering "yes" to question (a)):

$$d_i^{h,a} = \frac{1}{|\Omega_i|} \sum_{j \in \Omega_i} A_{i,j}. \tag{3}$$

Here, the subscript i denotes the index of a statement; the superscripts h and a indicate that this is the distribution based on human ratings, and with respect to question (a), respectively. For example, if 70% of people who were assigned to rate statement i agreed with it, then $d_i^{h,a} = 0.7$. If we take the mode of this distribution, the *human majority rating* each statement i is

$$\texttt{majority}_i^h = \mathbb{1}\left[d_i^{h,a} \geq 0.5\right], \tag{4}$$

where the indicator function $\mathbb{1}[\cdot]$ is equal to 1 if the given argument is true and 0 otherwise. Unlike the human rating distribution, the human majority rating is a binary variable and is equal to 1 if and only if at least half of the participants who rated i agreed with it. Note that we resolve ties, where exactly half of the participants agreed with a statement, with a final label of 1, or "agree." Note also that we only count ratings by participants in Ω_i because they were the only ones assigned to rate statement i.

Data source	N	**Examples**
News media	290	- A low covid positivity rate should allow for restrictions to be loosened. - Children should not be taught to be ashamed of their own skin color. - Providing access to loans for producers boosts agricultural economy during crisis. - We should aim to put our responsibilities before any personal or political agenda. - The proliferation of charter schools affects the ability of states to efficiently distribute tax dollars.
Campaign emails	668	- Abolish the police is an extreme stance taken by democrats. - Access to healthcare should be available for all without worry of debt. - Partisan politics and the democrats have held up real change. - The health care system can help a lot of americans that need jobs. - We should support freedom transparency and accountability in our elections.
ConceptNet	581	- A ball is round. - A desire to share knowledge would make you want to teach other people. - If you want to compete against someone then you should enter a competition. - Something that might happen as a consequence of having a conversation is exchange of information. - The first thing you do when you surf the web is turn on your computer.
ATOMIC	697	- If max accepts sam invitation then max intends to to gain person's friendship. - If max [*sic*] about to get married then max wants to be happily married. - If max accepts the job then max intends to to be employed. - If max accompanies sam far then max intends to to keep sam company. - If max attracts sam's attention then sam feels curious about max.
Aphorisms	709	- A doctor and a farmer know more than a doctor alone. - It is always the best policy to speak the truth unless of course you are an exceptionally good liar. - Let us never negotiate out of fear but let us never fear to negotiate. - Live your life as though your every act were to become a universal law. - Without craftsmanship inspiration is a mere reed shaken in the wind.
Situational response	630	- If a person digs themselves into a hole they will be in a bad mood. - If alex calls sam's mother they would have a conversation about sam. - If alex lands a job they will arrive early the first day of work to make a good impression. - If alex sees a bat he will poke the bat. - If alex works the night shift they will sleep during the day so they are awake for work.
Categorical response	832	- 5 is alot bigger than 1. - Advancements in technology can help save lives and improve the quality of life for everyone in the world. - Computers can never be an actual human. - If john wants to be a teacher he would apply for a teaching position. - Pausing before proceeding can lead to better results.

Table A.1. Seven sources of statements in our corpus, taken from Whiting and Watts (26). For ATOMIC (2), people's names such as "max" and "sam" were used to replace entities such as "PersonX" and "PersonY".

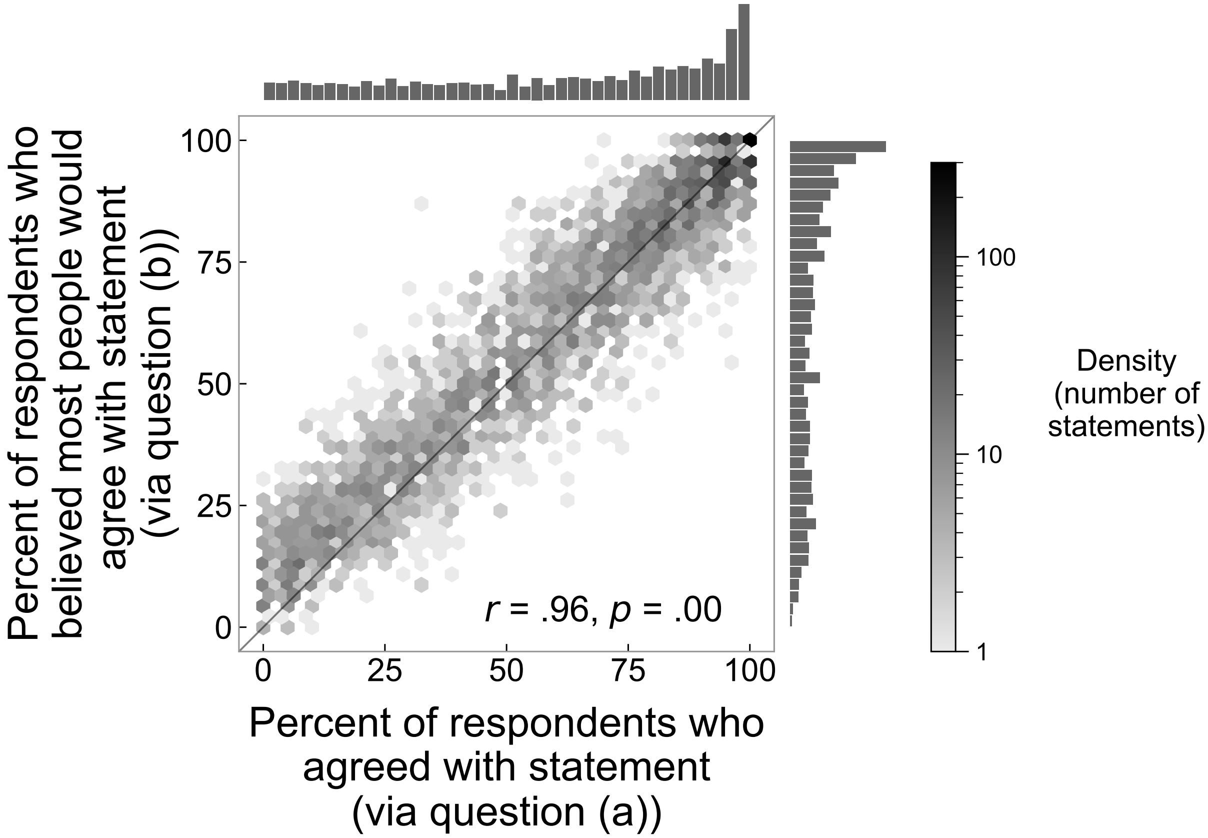


Figure A.1. Summary of the human ratings collected by Whiting and Watts (26). Participants (N = 2,046) were given a statement (N = 4,407) and asked to indicate (a) whether they agreed with the statement and (b) whether they thought most people would agree with the statement. Each participant assigned only 50 randomly chosen statements, and on average, each statement received 23 unique ratings. Every statement is depicted in this figure, where the x-axis represents the percentage of people who agreed with it, $d_i^{h,a}$ in Equation (3), and the y-axis represents the percentage of people who believed most others would agree with it, $d_i^{h,b}$ in Equation (5).

Similarly, the proportion of human participants who think that most other people would agree with statement i (*i.e.*, by answering "yes" to question (b)) is

$$d_i^{h,b} = \frac{1}{|\Omega_i|} \sum_{j \in \Omega_i} B_{i,j}. \tag{5}$$

Figure A.1 on page 21 shows the proportions of human participants who responded "yes" to questions (a) and (b) for all statements.

A.1 Statement-Level Commonsensicality

The commonsensicality of a statement is a combination of its two scores, consensus and awareness.

Given a statement i, its *consensus* score measures how close people's judgments on this statement are to absolute unanimity.

$$c_i^h = 2 \times \left| d_i^{h,a} - 0.5 \right|. \tag{6}$$

High consensus requires that the human rating distribution $d_i^{h,a}$ to be very close to 0 (everyone disagrees with i) or 1 (everyone agrees with i). Note that this definition is symmetric with respect to the polarity of opinions. For example, a statement which was agreed by 60% of participants would have the same consensus score as a statement which was disagreed by 60% of participants.

In addition, given statement i, its *awareness* score measures how accurately people perceived its majority rating. Formally, we have

$$a_i^h = \frac{1}{|\Omega_i|} \sum_{j \in \Omega_i} \mathbb{1}\left[B_{i,j} = \texttt{majority}_i^h\right]. \tag{7}$$

High awareness requires that people know what most others believe, which is recorded in question (b) above.

For a statement to be common sense, it must score high in both consensus and awareness. The *commonsensicality* score of statement i is defined as the geometric mean of its consensus and awareness scores:

$$m_i^h = \sqrt{c_i^h \times a_i^h}. \tag{8}$$

The commonsensicality score m_i^h is also between 0 and 1 and values closer to 1 indicate higher degrees of common sense.

Figure A.2 (bottom) on page 23 presents statements on two dimensions: consensus and awareness. Generally, statements received a higher awareness scores than consensus scores. The figure also shows that most statements received a high commonsensicality score, with a median of 70.12%.

A.2 Human-Level Commonsensicality

Commonsensicality is defined also for people, comprising consensus and awareness scores.

Given a person j, their *consensus* score is defined as the fraction of times they agreed with the majority rating across all statements they rated:

$$C_j = \frac{1}{|\Phi_j|} \sum_{i \in \Phi_j} \mathbb{1}\left[A_{i,j} = \texttt{majority}_i^h\right]. \tag{9}$$

To have high consensus, person j's own ratings of the statements, via their answer to question (a) above, must coincide with the overall human majority ratings.

The *awareness* score of person j is the fraction of time they accurately predicted the majority rating across all statements:

$$A_j = \frac{1}{|\Phi_j|} \sum_{i \in \Phi_j} \mathbb{1}\left[B_{i,j} = \texttt{majority}_i^h\right]. \tag{10}$$

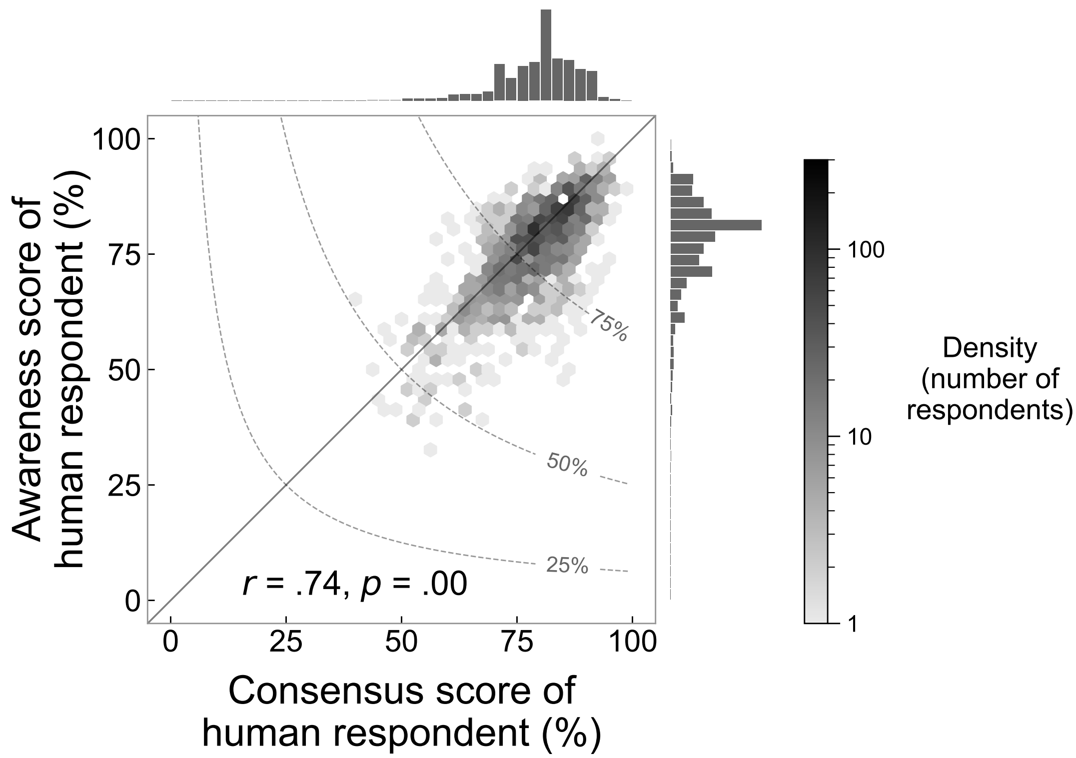
To have a high awareness score, person j's prediction of what most other people think—via question (b) above, and regardless of what they chose for question (a)—must also coincide with the overall human majority ratings.

Consensus C_j and awareness A_j scores are also between 0 and 1. Similarly, we take the geometric mean of these scores to give participant j's *commonsensicality* score as

$$M_j = \sqrt{C_j \times A_j}. \tag{11}$$

This score is also between 0 and 1 with higher values indicating higher degrees of common sense in a person.

Figure A.2 (top) on page 23 also presents humans on two dimensions: consensus and awareness. Different from statements, humans tended to exhibit very similar scores on both dimensions (most of them lie close to the diagonal line). The median commonsensicality for humans is 79.43%.



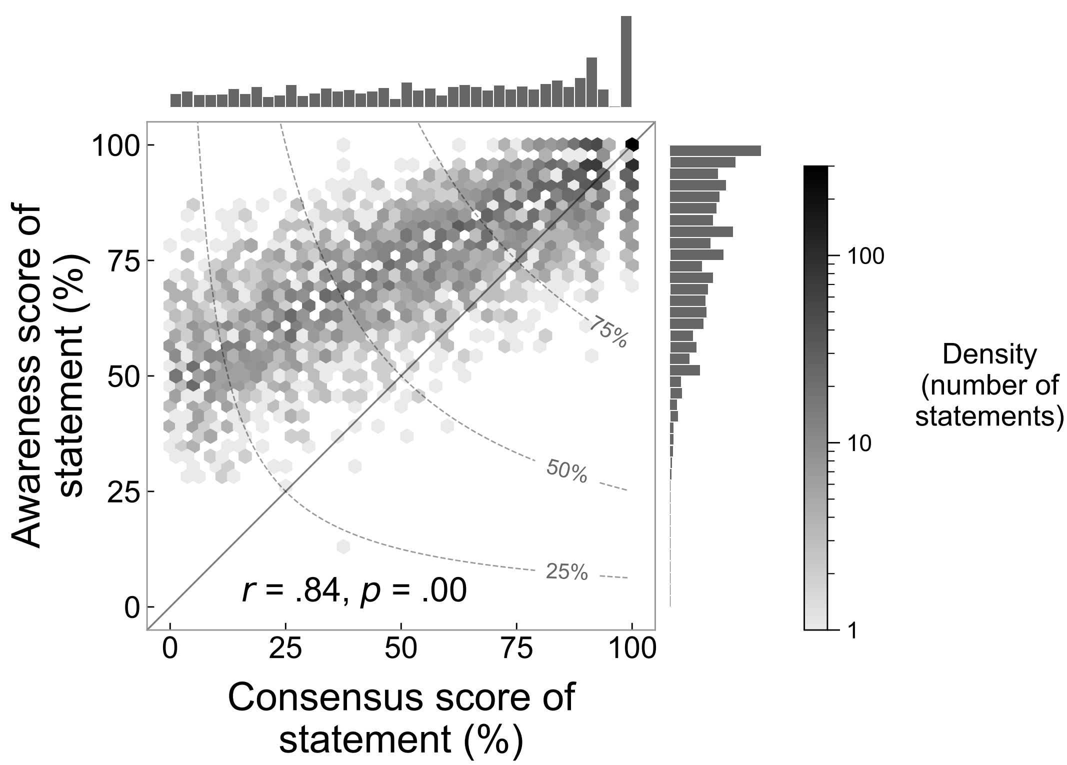


Figure A.2. Top: Individual-level commonsensicality, M_j in Equation (11), defined for every human participant. Bottom: collective-level commonsensicality, m_i^h in Equation (8) defined for every statement. The commonsensicality score is the geometric average of the consensus and awareness scores, which are depicted in these figures. The three level curves depict combinations of consensus and awareness that produce three levels of commonsensicality: 25%, 50%, and 75%.

B Collecting Responses from Language Models

To assess common sense in LLMs, we select a number of models that have been trained to follow conversations (instruction-tuned models). The full list of models we experiment with is found in Table B.1 on page 27. A typical conversation with an LLM is as follows.

```
[SYSTEM PROMPT]
[USER PROMPT]
[ASSISTANT RESPONSE]
[USER PROMPT]
[ASSISTANT RESPONSE]
...
```

The conversation starts with an optional *system prompt*, which is used to steer the general behavior of the LLM. For example, one can set this to "You are a debate expert" if the model is intended to be used to make arguments. Then, the system prompt is followed by a series of alternating *user prompts* and *assistant responses*. All messages in this conversation are concatenated into a *chat history*, which is a string of the form `[SYSTEM PROMPT] [SEP] [USER PROMPT] [SEP] [ASSISTANT RESPONSE] [SEP] [USER PROMPT] [SEP] [ASSISTANT RESPONSE]...` The `[SEP]` symbol is a special token denoting the separation of messages, and is specific to each model.

To use an LLM to generate an assistant response, one prepares a chat history that ends with a user prompt, and hence expects an assistant response to follow from it. The model takes the chat history as input and autoregressively generates new tokens one by one, until a finishing condition is reached. These tokens are then appended to the chat history, thereby completing the assistant response. Instruction-tuned models are specifically trained in this manner, via many such conversations, and hence we choose these models for assessment.

Code Block 1 shows a concrete example of our procedure, which is described in more detail in the below subsections.

B.1 Designing Prompts and Expected Responses

We decide not to use any specific system prompt. For some models, this is set to their default system prompts, such as "You are a helpful assistant." For some other models, this can be left blank, so the chat history actually starts with a user prompt.

For each of the $n_s = 4,407$ statements, we prompt a model to respond to the following two use prompts:

(a) `Consider the following statement, "[STATEMENT]" Do you agree with this statement? Start your answer with a "yes" or "no".`

(b) `Consider the following statement, "[STATEMENT]" Do you think most people would agree with this statement? Start your answer with a "yes" or "no".`

These are the same as questions (a) and (b) described in Section A. We note three details of these prompts. First, questions (a) and (b) are used in two separate chat sessions. This is because any chat history can modify a model's answer in unexpected ways, which we do not like to experience. Second, we include the sentence "Start your answer with a 'yes' or 'no'." at the end of each prompt to instruct the model to output a "yes" or "no" directly. Without this explicit instruction, the model may start its response with, for example, "I agree with the answer," or give some reasoning before giving its final answer. Either way, extracting the binary answer from would be more complicated and noisy. And third, in framing the questions in this way, we need the model to output only one token—"yes" or "no"—and we can extract the probability with which it generates either token, which represents the model's inherent uncertainty in its answer (more below).

```
1 # Load tokenizer and model
2 tokenizer = AutoTokenizer.from_pretrained(MODEL_NAME)
3 model = model = AutoModelForCausalLM.from_pretrained(
4     "databricks/dbrx-instruct",
5     device_map="auto",
6     trust_remote_code=True,
7     torch_dtype=torch.bfloat16,
8 )
9 model.eval(); model.tie_weights()
10
11 # These are all the IDs of the tokens in DBRX's vocabulary that map to
12 # the answers "yes" and "no".
13 answer2tokid = {
14     "yes": [9642, 60844, 14331, 58841, 20137, 60665, 7566, 85502, 77830, 95934, 10035, 86508, 98171, 41898, 14410, 9891],
15     "no": [72719, 6673, 42257, 86176, 43983, 67579, 2822, 79027, 18847, 40305, 34200, 58749, 5782, 38089, 39522, 73204, 9173, 99076, 2360, 38557, 31415, 12674, 85298, 29466, 2201, 912, 61559, 17184, 51899, 9278]
16 }
17
18 # Formulate the prompt
19 messages = [
20     {"role": "user",
21      "content": 'Consider the statement, "A desire to share knowledge would make you want to teach other people." Do you agree with this statement? Start your answer with a "yes" or "no".'}
22 ]
23
24 # Encode the message
25 inputs = tokenizer.apply_chat_template(messages, add_generation_prompt=True)
26
27 # Feedforward the input
28 inputs = torch.tensor(inputs, device=model.device, dtype=torch.long).reshape(1, -1)
29 outputs = model(inputs, output_hidden_states=False)
30
31 # Extract the probability of immediate next token
32 probs = torch.softmax(outputs[0][0, -1], axis=0)
33 probs = probs.cpu().detach().float().numpy()
34
35 # Map token IDs to answers
36 answer_probs = {}
37 # The probability that the model answers "yes" is the sum of the
38 # probabilities of all tokens that map to "yes".
39 answer_probs["yes"] = float(probs[answer2tokid["yes"]].sum())
40 # Similar for "no".
41 answer_probs["no"] = float(probs[answer2tokid["no"]].sum())
42 # The probability that the model answers with something else.
43 answer_probs["other"] = 1 - answer_probs["yes"] - answer_probs["no"]
```

Code Block 1. Python code used to extract an open-source model's answer to an example prompt. Here, we demonstrate this process using the model DBRX. Note that there could be multiple tokens within this model's vocabulary that can be mapped to the answers "yes" and "no," hence we need to enumerate them in the `answer2tokid` dictionary on lines 14–17. Based on the prompt (lines 19–22), the model is supposed to start its response with a definitive "yes" or "no". Therefore, we perform one feedforward operation to get the probabilities of all tokens that can be generated next (lines 29–33). We then sum the probabilities of all tokens that map to "yes" and "no" to get the model's probability of answering with either option (lines 36–43).

B.2 Extracting Responses from LLMs

Questions (a) and (b) in Section B.1 are formulated as user prompts for each model, which is then expected to generate a response starting with a "yes" or "no." As explained above, the model does so by autoregression: it maintains a multinomial probability distribution over all possible tokens and randomly chooses a token to generate via sampling from this distribution. In other words, the tokens "yes" and "no" are each assigned a probability with which they are generated. This also means that every other token may be generated, as long as its probability is not zero, leading to "illegal" answers. We address this problem in the following.

First, a review of how we can obtain tokens' probabilities for each model.

- For GPT-3.5 and GPT-4, OpenAI's API allows us to have access to up to 20 tokens associated with the highest sampling probabilities via the `logprobs` parameter.
- For Gemini 1.0 Pro, Claude 3 and Mistral-Large, their APIs do not allow access to these probabilities. We repeatedly ask the same question multiple times and report the empirical frequencies of the generated tokens, which give us an unbiased estimate of the probabilities. We perform 23 repetitions for Claude 3 and Mistral-Large, which matches the average number of people that rated each question in the original dataset (26). For Gemini 1.0 Pro, we perform 50 repetitions.
- For all other models, which are open-weight and loaded via the Hugging Face library in Python, we have access to the probabilities of all tokens.

In generating tokens, we set all sampling parameters to default; most notably, we keep the sampling temperature at 1.0. Once we have the probabilities of generated tokens, we perform the following processing steps.

- We discard all tokens that do not represent "yes" or "no" answers. This usually happens when a model wants to avoid giving an answer, for example, "As an AI model I cannot. . . " which leads to the first generated token being "As," an invalid token.
- For some models, their tokenizers have different variants of the same answer, due to formatting reasons. For example, the answer "yes" can be generated from the tokens `Yes`, `yes`, `␣yes`, `"Yes`, *etc.*, which should all be valid. In this case, we compile a list of all such valid tokens and sum over their probabilities to get the final score.
- After this, we have probabilities for the "yes" and "no" answers. Since we have discarded all illegal tokens, these two probabilities do not necessarily add up to one. We hence rescale them by dividing each by their sum. See Code Block 1.

The resulting binomial distribution of answers is called the *model rating distribution* and denoted

$$d_i^{m,a} = p(\text{“yes”} \mid \text{prompt (a) for statement } i). \tag{12}$$

where i denotes a statement and m denotes a model ($m \in \{1, \ldots, n_m\}$, n_m being the number of models). Compare this with the human rating distribution in Equation (3) on page 19.

Similarly, the probability that the model thinks most people would agree with statement i is

$$d_i^{m,b} = p(\text{“yes”} \mid \text{prompt (b) for statement } i). \tag{13}$$

Figure B.1 (top) on page 28 shows the frequencies of the human and model rating distributions. As depicted, human judgments are widely varied with many statements receiving close to 50/50 ratings. For most models, on the other hand, these probabilities are predominantly driven to either 0 and 1. The same is observed for humans' and models' answers to question (b).

Model family	Model name	Model size	Access	Source	Released	Comments
Claude	Claude 3 Haiku	–	Closed	(3)	03/24	Most capable
	Claude 3 Sonnet	–	Closed			Medium
	Claude 3 Opus	–	Closed			Least capable
DBRX	dbrx-instruct	132B	Open	(4)	03/24	
Falcon	Falcon-7b-Instruct	7B	Open	(5)	06/23	
	Falcon-40b-Instruct	40B	Open			
	Falcon-180B-Chat	180B	Open		09/23	
Flan-T5	Flan-T5-Small	80M	Open	(6)	10/22	
	Flan-T5-Base	250M	Open			
	Flan-T5-Large	780M	Open			
	Flan-T5-XL	3B	Open			
	Flan-T5-XXL	11B	Open			
Gemini	Gemma-2b-it	2B	Open	(7)	02/24	
	Gemma-7b-it	7B	Open			
	Gemini Pro 1.0	–	Closed	(8)		Original release
GPT	GPT-3.5-Turbo-0125	175B	Closed	(9)	01/24	Original model 2020
	GPT-4-Turbo-0125	–	Closed	(42)	01/24	Original model 2023
	GPT-4-Turbo-0409	–	Closed		04/24	
	GPT-4o	–	Closed	(11)	05/24	
LLaMA	LLaMA-2-7B-Chat	7B	Open	(12)	07/23	
	LLaMA-2-13B-Chat	13B	Open			
	LLaMA-2-70B-Chat	70B	Open			
	LLaMA-3-8B-Instruct	7B	Open	(13)	04/24	
	LLaMA-3-70B-Instruct	70B	Open			
Mistral	Mistral-7B-Instruct	7B	Open	(14)	09/23	
	Mixtral-8x7B-Instruct	47B	Open	(15)	11/23	
	Mixtral-8x22B-Instruct	141B	Open	(16)	04/24	
	Mistral-Large	–	Closed	(17)	02/24	
OLMo	OLMo-7B-Instruct	7B	Open	(18)	02/24	
Qwen	Qwen2-0.5B-Instruct	0.5B	Open	(19)	06/24	
	Qwen2-1.5B-Instruct	1.5B	Open			
	Qwen2-7B-Instruct	7B	Open			
	Qwen2-57B-Instruct	57B	Open			
	Qwen2-72B-Instruct	72B	Open			

Table B.1. Language models used for our evaluation of commonsensicality. All closed-source models (Claude 3, GPT-3.5/4 and Mistral-Large) are accessed via their developers' APIs. All open-weight models are accessed via Hugging Face in Python. For closed-source models, the "–" denotes their unknown sizes (number of parameters).

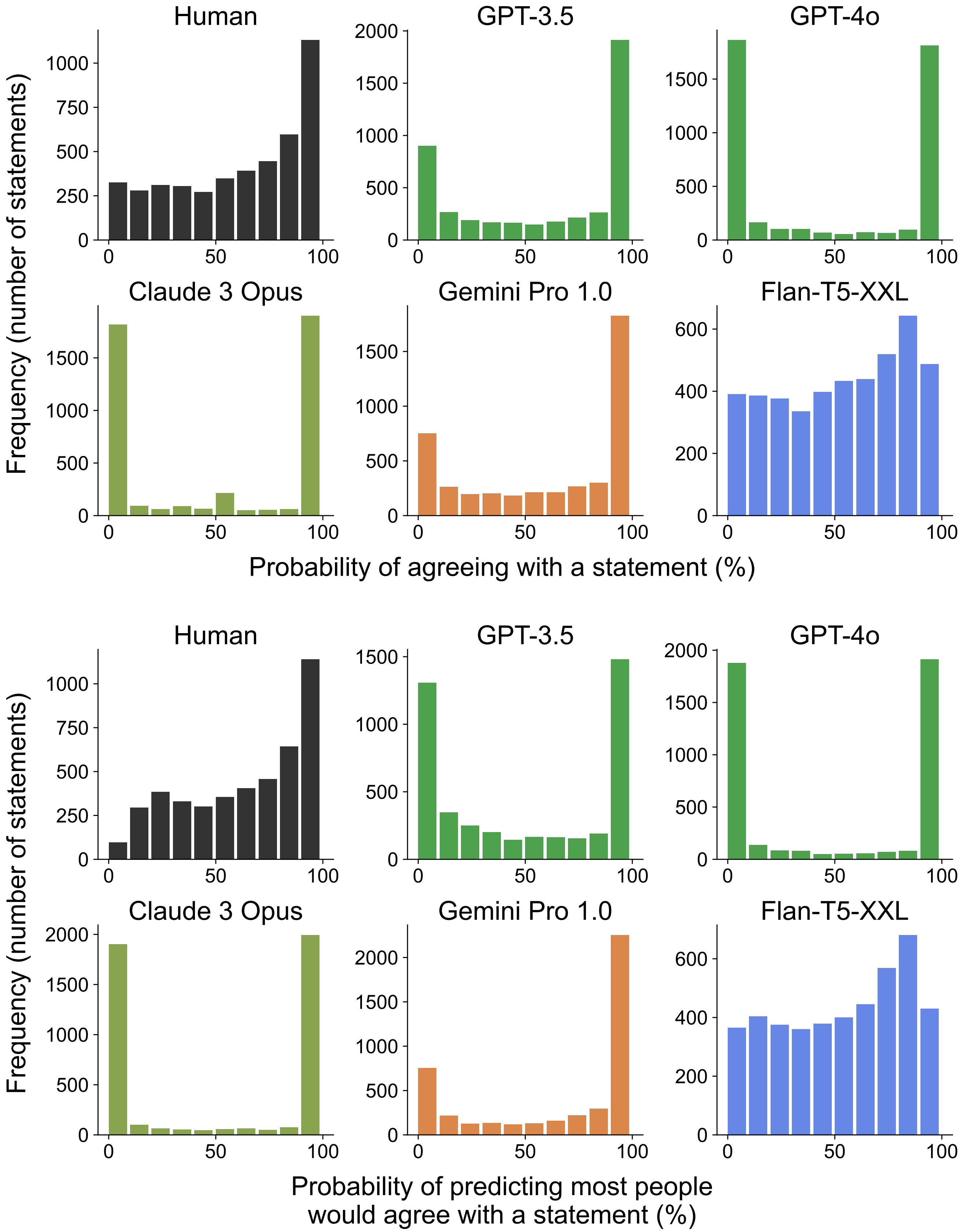


Figure B.1. Top: frequencies of the human rating distribution, $d_i^{h,a}$ (Equation (3)), and the model rating distribution $d_i^{m,a}$ (Equation (12)) for some models. These are derived from responses to question (a). Bottom: distribution of the same quantities, but for question (b).

C Measuring Individual-Level Commonsensicality in LLMs

In this section, we consider each language model as a hypothetical human survey respondent, just like those described in Section A.2.

Start from $d_i^{m,a}$ (Equation (12)), which is the probability that the model agrees with statement i, via its answer to question (a). We binarize this probabilistic output by taking its mode

$$\alpha_i^m = \mathbb{1}[d_i^{m,a} \geq 0.5]. \tag{14}$$

In other words, we consider that the model agrees with statement i (*i.e.*, $\alpha_i^{m,a} = 1$) if and only if it agrees with this statement with a probability of at least 0.5.

Similarly, the model believes that most people would agree with this statement if and only if it answers "yes" to question (b) with a probability of at least 0.5:

$$\beta_i^m = \mathbb{1}[d_i^{m,b} \geq 0.5]. \tag{15}$$

Based on these two binary answers, the model's *consensus* and *awareness* scores, respectively, are

$$C_m = \frac{1}{n_s} \sum_{i=1}^{n_s} \mathbb{1}\left[\alpha_i^m = \texttt{majority}_i^h\right], \tag{16}$$

$$A_m = \frac{1}{n_s} \sum_{i=1}^{n_s} \mathbb{1}\left[\beta_i^m = \texttt{majority}_i^h\right], \tag{17}$$

where $\texttt{majority}_i^h$ is the human majority rating for statement i (Equation (4)). Compare these calculations with the consensus and awareness scores for humans, in Equations (9) and (10).

Finally, similar to Equation (11), the model's *commonsensicality* score is the geometric mean of its consensus and awareness scores:

$$M_m = \sqrt{C_m \times A_m}. \tag{18}$$

All three scores are between 0 and 1, and higher values indicate closer proximity to absolute common sense. The goal of this calculation is to compare M_m with M_j, the commonsensicality of person j defined in Equation (11).

Table C.1 shows every model's consesus, awareness and Commonsensicality scores.

C.1 Model Size and Commonsensicality

To investigate the relationship between a model's size and its commonsensicality score, we select from Table B.1 (on page 27) 6 model families: Falcon, Flan-T5, Gemma, LLaMA-2/3, Mistral and Qwen2. Our selection criterion is that the family must include at least 2 models for which we know the sizes. See Table C.1. We also exclude Gemini 1.0 Pro in the Gemma/Gemini family and Mistral-Large in the Mistral family because these models are proprietary with unknown sizes. For GPT, although we know that GPT-3.5 has 175 billion parameters, we do not know this information for GPT-4.

These 6 families give us 23 models in total. Since model commonsensicality varies substantially across families, we perform a mixed-effects regression analysis to analyze the relationship between model size and commonsensicality. In R's syntax, the model is described by the following formula:

$$\texttt{commonsensicality} \sim \log_{10}(\texttt{model size}) + (1 \mid \texttt{model family}). \tag{19}$$

We also log-transform model size since these models can be orders of magnitude different from each other. In addition, we establish the 95% confidence interval of the effect of (log) model size using the `confint` function in R.

C.2 Model General Performance and Commonsensicality

There are numerous general benchmarks of LLMs available publicly. We choose the LMSYS Chatbot Arena benchmark (46) to represent a model's general performance, relative to other models.

LMSYS compares 114 LLMs models side by side, which include 23 models in our collection. When a user visits their website,[†] two models are randomly chosen. The user then interacts with these models simultaneously, asking the same question and getting corresponding answers from both. Then the user decides whether they prefer the left or the right model. (They also can choose to tie them, or to indicate both models are bad.)

Models in this collection are then assigned an Elo score, with higher values indicating a higher chance of being preferred by humans. Essentially, the Elo scores are used to approximate the win-rate matrix, and these scores are estimated by Chiang *et al.* (46) using the Bradley-Terry maximum likelihood method.

Refer to the original paper (46) for the exact calculation. The Elo score by LMSYS can be interpreted as follows. Suppose two models, a and b, receives their respective Elo scores of R_a and R_b. Then the probability that model a is chosen over model b is

$$\frac{1}{1 + 10^{\frac{R_b - R_a}{400}}}. \tag{20}$$

For example, GPT-4o and DBRX have scores of 1,287 and 1,103, respectively. The probability that DBRX is preferred over GPT-4o is

$$\frac{1}{1 + 10^{\frac{1{,}287 - 1{,}103}{400}}} \approx 0.257.$$

In other words, GPT-4o is expected to win about $100\% - 25.7\% = 74.3\%$ of the time. Elo scores are invariant to translation, *i.e.*, if all scores shift up or down by the same amount, the interpretation above does not change. In this paper, we keep the original scores that are reported on LMSYS Arena.

Since this calculation is updated over time on LMSYS Arena—as more models are added to the benchmark and more users participate in evaluation—we report the scores that are the most updated at the time of writing, July 8, 2024. Models' Elo scores are calculated based on 1,530,830 pairwise comparisons.

Finally, for the 23 models in our collection that also exist on LMSYS Arena (see Table C.1), we also extract their commonsensicality scores and evaluate the Pearson correlation between the two types of scores.

C.3 Comparing Models and Humans Side by Side on Commonsensicality

When comparing a model and a human directly on their commonsensicality scores, note that each person was only tasked with labeling 50 randomly chosen statements, whereas a model labels all 4,407 of them. For a fair comparison, we restrict the calculation of consensus, awareness and commonsensicality of the model only to the 50 statements that the person labeled.

Specifically, for person j the calculations in Equations (9) to (11) do not change because they are already with respect to Φ_j, the subset of statements that were labeled by j. To compare model m with person j, we modify Equations (16) and (17) above so that the average is with respect to Φ_j:

$$C_m = \frac{1}{|\Phi_j|} \sum_{i \in \Phi_j} \mathbb{1}\left[\alpha_i^m = \texttt{majority}_i^h\right],$$

$$A_m = \frac{1}{|\Phi_j|} \sum_{i \in \Phi_j} \mathbb{1}\left[\beta_i^m = \texttt{majority}_i^h\right].$$

[†] https://chat.lmsys.org

Model	Consensus (%)	Awareness (%)	Commonsensicality (%)	Size (B)	LMSYS Elo
Claude 3 Haiku	58.8	64.1	61.4	–	1,179
Claude 3 Sonnet	60.9	62.2	61.5	–	1,201
Claude 3 Opus	73.4	77.4	75.4	–	1,248
DBRX	73.7	79.0	76.3	132	1,103
Falcon-7B	66.6	66.1	66.3	7	–
Falcon-40B	73.0	77.2	75.1	40	–
Falcon-180B	78.6	81.3	79.9	180	1,034
Flan-T5-Small	34.4	33.9	34.2	.08	–
Flan-T5-Base	56.8	59.5	58.1	.25	–
Flan-T5-Large	77.3	76.5	76.9	.78	–
Flan-T5-XL	73.3	72.7	73.0	3	–
Flan-T5-XXL	79.9	80.9	80.4	11	–
Gemma-2B	65.2	66.6	65.9	2	989
Gemma-7B	73.2	70.9	72.0	7	1,084
Gemini Pro 1.0	78.4	81.1	79.7	–	1,209
GPT-3.5	78.3	75.4	76.8	175	1,105
GPT-4-0125	77.6	79.2	78.4	–	1,246
GPT-4-0409	78.0	83.3	80.6	–	1,257
GPT-4o	72.5	77.9	75.2	–	1,287
LLaMA-2-7B	74.0	76.0	75.0	7	1,037
LLaMA-2-13B	48.5	44.5	46.5	13	1,063
LLaMA-2-70B	65.7	61.4	63.5	70	1,093
LLaMA-3-8B	57.2	66.5	61.7	8	1,152
LLaMA-3-70B	72.0	76.8	74.4	70	1,207
Mistral-7B	80.2	80.7	80.4	7	1,072
Mixtral-8x7B	77.8	75.0	76.4	47	1,114
Mixtral-8x22B	80.7	84.0	82.3	141	1,146
Mistral-Large	80.4	82.2	81.3	–	1,157
OLMo-7B	74.3	71.0	72.7	7	1,015
Qwen2-0.5B	67.1	66.5	66.8	.5	–
Qwen2-1.5B	75.4	73.8	74.6	1.5	–
Qwen2-7B	79.7	81.1	80.4	7	–
Qwen2-57B	80.4	81.4	80.9	57	–
Qwen2-72B	80.5	81.8	81.1	72	1,188

Table C.1. Individual-level consensus (C_m, Equation (16)), awareness (A_m, Equation (17)) and commonsensicality (M_m, Equation (18)) scores for all models. This table also presents the sizes of these models (in billions of parameters, also shown in Table B.1) as well as their Elo scores on the LMSYS Chatbot Arena benchmark (46).

D Measuring Commonsensicality when LLMs are Simulators of Silicon Samples

In this section, we use each language model not as an individual survey participant (*cf.* Section C), but rather as a generator of "silicon samples" which, together, form a hypothetical population raters. Common sense in this case is measured with respect to this entire population, rather than for each silicon sample within.

To make this clearer, imagine a population of silicon samples, each represented by the same language model. When we ask each respondent in this population questions (a) and (b), we record its binary answers ("yes" or "no"). Each respondent is treated as independent from one another, and so collecting answers for this entire population amounts to resampling the answers to the same questions. In the limit—*i.e.*, if the size of the population goes to infinity—the proportion of respondents that agree with statement i is exactly the probability with which the language model agrees with this statement, $d_i^{m,a}$ (defined in Equation (12) on page 26). Similarly, the proportion of silicon respondents that believe most people would agree with this statement is exactly $d_i^{m,b}$ (Equation (13) on page 26).

Mimicking the calculations in Section A.1, we first define the majority rating within this population of silicon samples for statement i as

$$\texttt{majority}_i^m = \mathbb{1}[d_i^{m,a} \geq 0.5]. \tag{21}$$

The consensus score of statement i with respect to this population is

$$c_i^m = 2 \times |d_i^{m,a} - 0.5|. \tag{22}$$

The statement's awareness score is the proportion of silicon samples that accurately predict the majority opinion via their answer to question (b):

$$a_i^m = \begin{cases} d_i^{m,b} & \text{if} \quad \texttt{majority}_i^m = 1 \\ 1 - d_i^{m,b} & \text{if} \quad \texttt{majority}_i^m = 0. \end{cases} \tag{23}$$

Which leads to its commonsensicality score of

$$m_i^m = \sqrt{c_i^m \times a_i^m}.. \tag{24}$$

The objective is to compare m_i^m with m_i^h (Equation (8) on page 22). First, we look at the correlation between these two scores, taken over $n_s = 4,407$ statements. Specifically, we calculate the Pearson correlation between m_i^m with m_i^h as well as its two-sided p-values (after Bonferroni correction).

We also make this comparison through a different lens: how well does a statement's commonsensicality score in the population of silicon samples predict its same score in the population of humans? To do so, we fit a linear regression model, predicting m_i^h using m_i^m. We report the regression coefficient as well as its two-sided p-values.

D.1 Comparing Statement Commonsensicality Ccores within Humans and Silicon Samples by Statement Type

Each statement in our corpus was labeled by MTurk workers according to six epistemological features:

1. *Behavior*
   - *Social*: it refers to beliefs, perceptions, preferences, and socially constructed rules that govern human experience; it can be "real" or opinion, but is intrinsically of human origins. *e.g.*, "I exist and am the same person I was yesterday," "He yelled at me because he was angry," "There are seven days in the week."
   - *Physical*: it refers to objective features of the world as described by, say, physics, biology, engineering, mathematics or other natural rules; it can be measured empirically, or derived logically. *e.g.*, "Men on average are taller than women," "The Earth is the third planet from the Sun," "Ants are smaller than Elephants."

2. *Everyday*
    - *Everyday*: people encounter, or could encounter, situations like this in the course of their ordinary, everyday experiences, *e.g.*, "Touching a hot stove will burn you," "Commuting at rush hour takes longer," "It is rude to jump the line."
    - *Abstract*: this claim refers to regularities or conclusions that cannot be observed or arrived at solely through individual experience, *e.g.*, "Capitalism is a better economic system than Communism," "Strict gun laws save lives," "God exists."
3. *Figure of speech*
    - *Figure of speech*: it contains an aphorism, metaphor, hyperbole, *e.g.*, "Birds of a feather flock together," "A friend to all is a friend to none."
    - *Literal language*: it is plain and ordinary language that means exactly what it says. *e.g.*"The sky is blue," "Elephants are larger than dogs," "Abraham Lincoln was a great president."
4. *Judgment*
    - *Normative*: it refers to a judgment, belief, value, social norm or convention. *e.g.*, "If you are going to the office, you should wear business attire, not a bathing suit," "Treat others how you want them to treat you," "Freedom is a fundamental human right."
    - *Positive*: it refers to something in the world such as an empirical regularity or scientific law, *e.g.*, "Hot things will burn you," "The sun rises in the east and sets in the west."s
5. *Opinion*
    - *Opinion*: it is something that someone might think is true, or wants others to think is true, but can't be demonstrated to be objectively correct or incorrect; it is inherently subjective. *e.g.*, "FDR was the greatest US president of the 20th Century," "The Brooklyn Bridge is prettier than the Golden Gate," "Vaccine mandates are a tolerable imposition on individual freedom."
    - *Factual*: it is something that can be demonstrated to be correct or incorrect, independently of anyone's opinion, *e.g.*, "the earth is the third planet from the sun," (this is correct and we know it is correct), "Obama was the 24th president of the United States" (this is incorrect, but we know it's incorrect) "It will be sunny next Tuesday" (we don't yet know if this is correct, but we will be able to check in the future).
6. *Reasoning*
    - *Knowledge*: the claim refers to some observation about the world; it may be true or false, opinion or fact, subjective or objective *e.g.*, "The sun rises in the east and sets in the west," "Dogs are nicer than cats," "Glasses break when they are dropped."
    - *Reasoning*: the claim presents a conclusion that is arrived at by combining knowledge and logic, *e.g.*, "The sun is in the east, therefore it is morning," "My dog is wagging its tail, therefore it is happy," "The glass fell off the table, therefore it will break and the floor will become wet."

For every dichotomy, the corpus of $n_s = 4,407$ statements is divided into two subsets corresponding to the two opposing dimensions in that dichotomy. Our goal is to compare the commonsensicality scores (either by humans or by models) between these two dimensions. For instance, a statement is either about a fact or an opinion, corresponding to dichotomy 5 above. Are statements depicting facts more commonsensical than those depicting opinions?

To do so, we compute the difference in mean commonsensicality score between the two subsets of statements. This is done for every population: of humans and of silicon samples created by 34 models. To establish a confidence interval, we construct 1,000 boostrapped samples and calculate this difference for each sample.

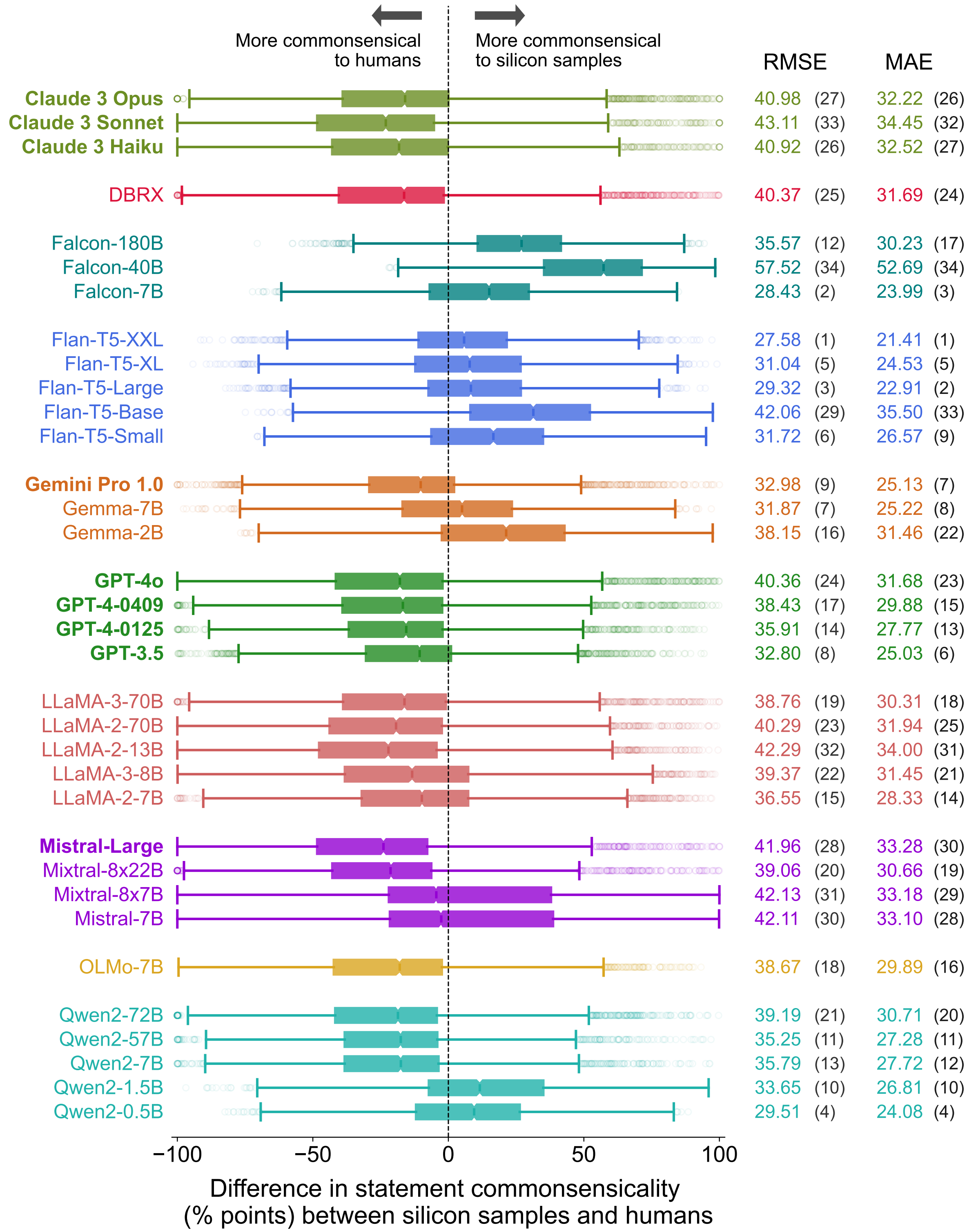


Figure D.1. Comparison of statement commonsensicality scores between humans and silicon samples generated by LLMs. The x-axis represents the difference in a statement's scores between these two populations, where a positive difference indicates that the statement is more commonsensical to silicon samples than it is to humans (*i.e.*, $m_i^m > m_i^h$). Each box indicates the inter-quartile range, and the median notch is contained within it. The root mean-squared error (RMSE) and mean absolute error (MAE) are calculated based on the difference between m_i^m and m_i^h, averaged over 4,407 statements. The numbers in parentheses are the rankings of the RMSE and MAE (in ascending order); for example, Claude 3 Opus has an RMSE ranking of 27 out of 34 models, whereas its MAE ranking is 26 out of 34.

SI References

1. M. E. Whiting, D. J. Watts, A Framework for Quantifying Individual and Collective Common Sense. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2309535121 (2024).
2. M. Sap *et al.*, Atomic: An Atlas of Machine Commonsense for If-Then Reasoning. *AAAI* **33**, 3027–3035 (2019).
3. Anthropic, The Claude 3 Model Family: Opus, Sonnet, Haiku. (2024). https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf (Accessed 2024-05-28).
4. M. R. Team, Introducing DBRX: A New State-of-the-Art Open LLM. (2024). https://www.databricks.com/blog/introducing-dbrx-new-state-art-open-llm (Accessed 2024-06-15).
5. E. Almazrouei *et al.*, The Falcon Series of Open Language Models. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2311.16867 (Accessed 2024-01-01).
6. H. W. Chung *et al.*, Scaling Instruction-Finetuned Language Models. arXiv [Preprint] (2022). https://doi.org/10.48550/arXiv.2210.11416 (Accessed 2023-04-05).
7. Google, Gemma: Open Models Based on Gemini Research and Technology. (2024). https://storage.googleapis.com/deepmind-media/gemma/gemma-report.pdf (Accessed 2024-02-03).
8. Google, Gemini 1.5: Unlocking Multimodal Understanding across Millions of Tokens of Context. (2024). https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf (Accessed 2024-03-01).
9. T. Brown *et al.*, Language Models Are Few-Shot Learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
10. J. Achiam *et al.*, GPT-4 Technical Report. arXiv [Preprint] (2023). https://arxiv.org/abs/2303.08774 (Accessed 2024-03-20).
11. OpenAI, GPT-4o System Card. (2024). https://cdn.openai.com/gpt-4o-system-card.pdf (Accessed 2024-08-01).
12. H. Touvron *et al.*, Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2307.09288 (Accessed 2023-09-12).
13. A. Grattafiori *et al.*, The Llama 3 Herd of Models. arXiv [Preprint] (2024). https://arxiv.org/abs/2407.21783 (Accessed 2025-01-06).
14. A. Q. Jiang *et al.*, Mistral 7B. arXiv [Preprint] (2023). https://arxiv.org/abs/2310.06825 (Accessed 2023-11-13).
15. A. Q. Jiang *et al.*, Mixtral of Experts. arXiv [Preprint] (2024). https://doi.org/10.48550/arXiv.2401.04088 (Accessed 2024-06-07).
16. M. AI, Cheaper, Better, Faster, Stronger. (2024). https://mistral.ai/news/mixtral-8x22b/ (Accessed 2024-06-15).
17. M. AI, Au Large. (2024). https://mistral.ai/news/mistral-large/ (Accessed 2024-05-28).
18. D. Groeneveld *et al.*, "OLMo: Accelerating the Science of Language Models" in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2024), pp. 15789–15809.
19. A. Yang *et al.*, Qwen2 Technical Report. arXiv [Preprint] (2024). https://doi.org/10.48550/arXiv.2407.10671 (Accessed 2024-12-11).
20. W.-L. Chiang *et al.*, "Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference" in *Proceedings of the 41st International Conference on Machine Learning* (2024).