PLOS ONE

RESEARCH ARTICLE

Framework-based qualitative analysis of free responses of Large Language Models: Algorithmic fidelity

Aliya Amirova[1]*, **Theodora Fteropoulli**[2], **Nafiso Ahmed**[3], **Martin R. Cowie**[4,5], **Joel Z. Leibo**[6,7]

1 Population Health Sciences, School of Life Course & Population Sciences, Faculty of Life Sciences & Medicine, King's College London, London, United Kingdom, **2** Medical School, University of Cyprus, Nicosia, Cyprus, **3** Division of Psychiatry, University College London, London, United Kingdom, **4** Royal Brompton Hospital, London, London, United Kingdom, **5** School of Cardiovascular Medicine & Sciences, Faculty of Life Sciences & Medicine, King's College London, London, United Kingdom, **6** Google DeepMind, London, London, United Kingdom, **7** Department of Informatics, Faculty of Natural, Mathematical & Engineering Sciences, King's College London, London, United Kingdom

* aliya.1.amirova@kcl.ac.uk

Abstract

Today, with the advent of Large-scale generative Language Models (LLMs) it is now possible to simulate free responses to interview questions such as those traditionally analyzed using qualitative research methods. Qualitative methodology encompasses a broad family of techniques involving manual analysis of open-ended interviews or conversations conducted freely in natural language. Here we consider whether artificial "silicon participants" generated by LLMs may be productively studied using qualitative analysis methods in such a way as to generate insights that could generalize to real human populations. The key concept in our analysis is *algorithmic fidelity*, a validity concept capturing the degree to which LLM-generated outputs mirror human sub-populations' beliefs and attitudes. By definition, high algorithmic fidelity suggests that latent beliefs elicited from LLMs may generalize to real humans, whereas low algorithmic fidelity renders such research invalid. Here we used an LLM to generate interviews with "silicon participants" matching specific demographic characteristics one-for-one with a set of human participants. Using framework-based qualitative analysis, we showed the key themes obtained from both human and silicon participants were strikingly similar. However, when we analyzed the structure and tone of the interviews we found even more striking differences. We also found evidence of a hyper-accuracy distortion. We conclude that the LLM we tested (GPT-3.5) does not have sufficient algorithmic fidelity to expect *in silico* research on it to generalize to real human populations. However, rapid advances in artificial intelligence raise the possibility that algorithmic fidelity may improve in the future. Thus we stress the need to establish epistemic norms now around how to assess the validity of LLM-based qualitative research, especially concerning the need to ensure the representation of heterogeneous lived experiences.

OPEN ACCESS

Citation: Amirova A, Fteropoulli T, Ahmed N, Cowie MR, Leibo JZ (2024) Framework-based qualitative analysis of free responses of Large Language Models: Algorithmic fidelity. PLoS ONE 19(3): e0300024. https://doi.org/10.1371/journal.pone.0300024

Editor: Primož Kocbek, University of Maribor Faculty of Health Sciences: Univerza v Mariboru Fakulteta za Zdravstvene Vede, SLOVENIA

Received: October 2, 2023

Accepted: February 21, 2024

Published: March 12, 2024

Copyright: © 2024 Amirova et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Qualitative data underpinning the manuscript cannot be shared publicly because of the nature of the data and ethics committee agreement (Health Research Authority. Cambridge Central Research Ethics Committee, REC reference: 17/EE/0183). Participants' approved consent specifically states that potentially identifiable participant data will not be shared. Given the nature of the qualitative transcripts and data analysis, although information on names, clinical team names and family members' names was retracted, the transcripts still contain detailed information that can be traced back to single individuals even in the aggregate form, given the study context. A point of contact: NRESCommittee.EastofEngland-CambridgeCentral@nhs.net.

Funding: The author(s) received no specific funding for this work.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: JZL is an employee of Google.

Introduction

Large-Scale generative Language Models (LLMs) [1–3] may provide a new opportunity for capturing available knowledge and beliefs at scale and facilitate *in silico* research on human behavior and cognition [4–8]. By virtue of their training, LLMs may contain substantial latent social information [9]—enough to consider them plausible computational models of humans. LLMs may capture economic laws, decision-making heuristics, and social preferences [9], as well as mirroring human moral judgments [7]. In principle, using LLMs could be a cost-effective and efficient way to gain insights and explore how self-reports vary, as well as to pilot experiments *in silico* to test sensitivity of responses to precise wording. Cheap and easy to run *in silico* experiments can guide expensive and slow empirical work with real participants. However, how can we know such results are trustworthy?

Algorithmic fidelity is an appropriate validity concept for research on human behavior using survey data simulated using Large-scale generative Language Models (LLMs) [4]. Algorithmic fidelity describes the extent to which the outputs of LLMs conditioned to simulate specific human sub-populations actually reflect the beliefs and attitudes of those subpopulations.

By definition, high algorithmic fidelity suggests that beliefs elicited from the LLM will generalize to real humans, while low fidelity renders such inferences invalid. Argyle et al. [4] introduced the idea of generating *"silicon samples"* or *"silicon participants"* by conditioning LLMs using backstories matching sociodemographic data from human survey participants. Their study found that LLM outputs closely mirrored the complex interplay of beliefs, attitudes, and sociocultural contexts that shape human responses to surveys about American politics. In the present study, we aim to extend algorithmic fidelity assessment methodology [4] using qualitative research methods so that it can be applicable to studies where the data consists of fully freeform responses in natural language.

There are numerous applications for models with demonstrably high algorithmic fidelity. They include digital avatars [10, 11], digital behaviour change interventions [12–14], digital therapeutics, [15, 16], non-player characters for computer games [17], and teaching assistants. Algorithmic fidelity assessment will also be important in research that aims to construct multi-agent simulations to simulate the effects of interventions (e.g. [18, 19]) to guide policy making in regimes where real experiments are infeasible. This approach aims to address core objectives for responsible AI deployment and regulation [20, 21]. In these settings, the reason to assess algorithmic fidelity is to provide empirical validation for a part of a model to be used in a downstream application where direct validity measures for the full multi-agent model do not exist.

There is no reason to think algorithmic fidelity would be uniform over the many and varied parts of human lived experience or the many and varied social science research topics. These models will clearly be better at simulating some people over other people, and work better for some applications than others. Argyle et al. [4] conclude from this that algorithmic fidelity must be measured anew for each research question. Determining there is sufficient algorithmic fidelity to address one research question does not imply the same will be true for others [4]. At present, there are still no generally accepted best practices for research on humans using LLMs. Nevertheless, we think one aspect of sound methodology is already clear at this point: there should be guidelines counseling researchers to perform a dedicated assessment of algorithmic fidelity tailored for their specific research question and target population.

Understanding *beliefs* mediated using natural language and their impact on behavior (e.g., beliefs such as *"vaccination is an effective and safe way to protect me and others around me"*) is important but not straightforward to study using conventional frequentist methods. Qualitative research methods come to the rescue when the nuance of such beliefs needs to be understood. Some of these methods include ethnography [22], phenomenological studies [23], grounded theory [24], thematic analysis [25], and framework-based analysis [26]. For example, when exploring the acceptability of newly developed treatments, we might want to elicit beliefs and attitudes from *stakeholders* to improve services and healthcare [27–29]. Stakeholders may have critical information that would be missed if researchers were to rely only on prior theories, experiments, and survey data. In fact, involving stakeholders can guide the development of interventions and policies that are relevant, effective and acceptable for them, ultimately leading to improved outcomes [28, 29]. This is helpful for allocating research efforts to the most important problems and for ensuring interventions are targeted. Another research field where it is important to understand latent beliefs mediated by natural language is behavior change. The goal of this field is to promote beneficial behavior or reduce harmful behavior [30]. Behavior change may be considered on a variety of different scales. For instance, health psychologists are concerned with determining the barriers and enablers to vaccination, adopting healthy behaviors (e.g. exercise), or ceasing to engage in harmful behaviors (e.g. tobacco smoking). There is a related research program that seeks interventions by which whole communities can be helped to adopt helpful social norms (e.g. educating girls) or drop harmful norms (e.g. child marriage) [31]. Such research affects wellbeing and health of the general public. Appropriate standards evaluating the rigor with which it is conducted and the trustworthiness of its findings should be in place.

The present study is concerned with the methodological question of whether or not LLMs contain sufficient algorithmic fidelity for us to generalize and extrapolate the results generated by them to human populations of interest. To illustrate how to assess algorithmic fidelity for freeform natural text data, we draw on a qualitative study using semi-structured interviews with heart failure patients that aimed to elicit beliefs about barriers and enablers influencing physical activity [32]. Increasing physical activity levels is known to have many health benefits for this group [33–35]. However, most patients do not do enough [36] as they face barriers associated with older age, co-morbid depression, symptom distress, and negative emotional response to physical activity [37]. Promoting physical activity for this population group is challenging [38]. Research on this type of questions has serious implications for individual and population health.

We find that GPT-3.5 has insufficient algorithmic fidelity to address the research question of how to promote physical activity in older adults with heart failure. Nevertheless, even though our specific result was negative, we do think this technology is promising. It is very likely that future systems will have sufficient algorithmic fidelity to support trustworthy *in silico* research, all the more reason to critically engage now with this validity concept. We predict that it will become the core validity concept needed to support this nascent field going forward. It’s important to demonstrate how it can be used to reject invalid *in silico* research now, so that once LLMs have improved enough we will then be able to use this logic routinely to support or reject the validity of specific research findings. In short, we need to develop epistemic norms to apply when we review papers and decide whether to approve or disapprove of arguments that arise in this newly LLM-enabled field of research on human behavior.

The main methodological contribution of the present study is to apply *framework-based* qualitative research methods [26] to assessing algorithmic fidelity for LLM-generated freeform natural text data. This approach to qualitative analysis finds patterns in free-form text, examines their relationships, and derives common themes [26]. Employing a well-established framework supported by domain expert consensus ensures consistency in semi-structured interviews between human participants and LLMs, setting a uniform standard for evaluation. In addition, without a framework for the interview schedule, there is a risk that results would not amount to much more than anecdotes and, thus, not be credible. However, by using a framework, it is possible to *systematically* probe the beliefs of both silicon and human participants and thereby build confidence that the list of uncovered beliefs is complete enough to be useful. The specific framework we used in this study is called Theoretical Domains Framework (TDF) [39].

However, we show that, currently, LLMs do not demonstrate sufficient algorithmic fidelity. Yet, with the fast-paced development of these technologies, it is expected that they will in the not-so-distant future. Establishing a disciplinary consensus on defining and evaluating algorithmic fidelity is therefore crucial. Before LLMs can safely be used in designing interventions and policies there is a need to develop a set of guidelines for routine assessment of algorithmic fidelity to support or reject claims made using them. Such guidelines should consider the faithfulness of the LLM to lived experiences and stakeholder beliefs.

Large language models

Our approach to assessing algorithmic fidelity involves conducting interviews with silicon participants matching the specific demographic characteristics of a set of real human participants. Before we justify this approach, we need to provide some additional background on LLMs.

Language models are conditional distributions over natural language. They are probability distributions $P(x_n|x_{n-1}, \ldots, x_0)$ where all x_i are parts of words. A sentence is a sequence $x_0, \ldots, x_n$. A paragraph is a longer such sequence. These complex conditional probability models are created (i.e. trained) by maximizing the likelihood of a dataset consisting of many billions of documents, harvested both from books and from the internet (e.g. [40]). Predicting x_n from its context (the preceding $x_{n-1}, \ldots, x_0$), requires the model to absorb a substantial amount of latent knowledge about the world, about language, and about people. When an LLM has sufficient algorithmic fidelity, we may prompt it with a context containing demographic information (e.g. *"a 74-year-old man with heart failure and arthritis who lives in a major city"*) to elicit outputs that correlate with the attitudes, opinions, and experiences of the human sub-population to whom the specified demographic data applies.

Generating samples from an LLM is done *autoregressively*, word by word, conditioning on the growing sequence of preceding words leading up to the next one to be predicted (Fig 1). To prevent the conditioning sequence from growing longer and longer it is cut off once a certain maximum size is reached called the context length. Once a document (or conversation with a chatbot) grows beyond context length in size some of the preceding words must be dropped, so they no longer condition subsequent outputs. Different LLMs and systems make different choices in how exactly they handle the necessary "forgetting" required to support long contexts. However, there are many ways to simulate longer contexts with shorter contexts so the user may not notice when the conversation grows beyond context length.

It's crucial to note that the concepts of validity and reliability, commonly employed in statistics, don't have direct analogues that can be confidently applied to LLMs. Empirical assessment is the only way to ensure their validity. Most models considered in applied machine learning are both too complicated and too closely tailored to their specific application to be generically justified. Applied machine learning researchers typically think about methodological rigour as demanding they implement "cross validation" procedures which estimate operationalized generalization concepts. The most critical such procedure is to split data into a portion used for training, i.e. model fitting, and a different portion used for testing [41].

Some large-scale generative language models go through a fine-tuning step where human raters provide specific feedback used to train them to follow instructions and align with ethical principles and product design goals [42]. These include Sparrow [43], Anthropic's assistant [44], and Instruct-GPT [45], the latter of which was the basis for the GPT-3.5 system which we

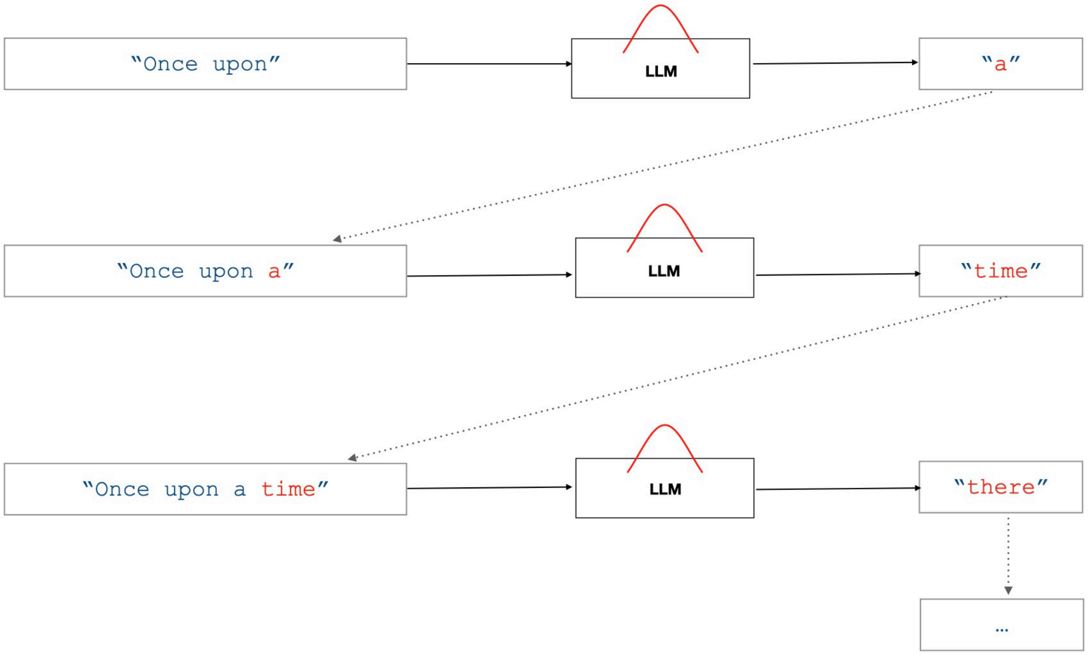

Fig 1. Schematic representation of autoregressive sampling in large language models (LLMs). The diagram illustrates the iterative sampling process in three stages: input, processing by the LLM, and output. The LLM represents the probability distribution over all possible next words given the current context (previous words). Arrows indicate the flow of information, with solid arrows representing the transition from one stage to another within a single iteration, and dotted arrows indicating the progression from one iteration to the next. The outputted words ("a", "time", "there") are samples from the corresponding probability distributions and are appended to the context for the next iteration.

https://doi.org/10.1371/journal.pone.0300024.g001

used in this paper. Fine-tuning can be accomplished through a variety of techniques, the most prominent being Reinforcement Learning from Human Feedback (RLHF). It involves human raters selecting which of several possible model responses they prefer. The raters are not asked the subjective question of which response they themselves prefer, but rather are given a specific checklist of product design goals to apply in making their judgments. All the model's capabilities are already present after the pretraining stage [46]. The postprocessing steps aim to reduce the raw model's propensity to produce toxic responses as well as to make it implement a consistent "personality" in accord with product design goals. These steps are not always entirely effective in preventing LLMs from producing undesirable behaviors like toxic or harmful language, and "jailbreak" prompts which trick the model into responding inappropriately are still easy to discover and implement [47, 48].

Algorithmic fidelity

Argyle et al. (2023) offered a suggestion for how to use LLMs in social science research. Given the correct conditioning, the authors argue that free-text output generated by large language models like GPT-3 can serve as effective proxies for specific human population groups. Argyle et al. (2023) introduce a concept referred to as *algorithmic fidelity*, which describes an ability of a language model to accurately reflect the ideas, beliefs, and attitudes situated in sociocultural contexts of various population groups when adequately conditioned. This is a departure from the common view that AI biases are uniform across models. Instead, they argue that these biases are nuanced and demographically correlated. Such a view offers a deeper and more fine-grained understanding of the LLMs outputs and the biases they are subjected to. The researchers propose that "silicon samples" can be generated by training and prompting LLMs using sociodemographic backstories from real human participants in several large surveys conducted in the U.S. They then compared the outputs of the model against the responses of the human participants to assess the accuracy of the model's representation. They found that model outputs go beyond sharing superficial similarities and instead reflect the nuanced and complex interplay between ideas, attitudes, and the sociocultural context that shaped them.

Argyle et al. (2023) propose that with sufficient algorithmic fidelity, language models like GPT-3 can be used as tools for advancing our understanding of humans and society across various disciplines. This could transform how we approach social science research, offering new perspectives and opportunities for piloting and conducting research at scale. Argyle et al. (2023) provide a framework for evaluating algorithmic fidelity through four distinct criteria: (1) Generated responses are indistinguishable from parallel human texts (a social science Turing test) (2) Generated responses are consistent with the attitudes and sociodemographic information of its input (i.e., conditioning context) such that humans viewing the responses can infer critical elements of that input (Backward Continuity); (3) Generated responses proceed naturally from the conditioning context provided, reliably reflecting the form, tone, and content of the context (Forward Continuity); (4) Generated responses reflect underlying patterns of relationships between ideas, demographics, and behaviour that would be observed in comparable human-produced data (Pattern Correspondence). Silicon Sampling is the methodology proposed by Argyle et al. (2013), which describes how to use an LLM to generate a virtual dataset (i.e., silicon samples) using demographic backstories as prompts to create variability. Conditioning on a backstory is expected to produce outputs from the model representative of the subpopulation of real people with a similar backstory.

We can make an LLM simulate multiple different silicon participants. We did this by providing each participant with a distinct prompt containing their biographical information. For instance, using backstories with fictional names, that were popular in 1950 like:

> *"The participant is a 77-year-old woman with heart failure and rheumatoid arthritis called Linda. Linda lives in the countryside. She is fairly physically active."*

The model continues sampling freeform natural language correlated with the identity provided to it in its prompt.

The algorithmic fidelity criteria developed by Argyle et al. [4] were designed for written survey-based outputs. We propose modified versions of these criteria that are more appropriate for qualitative research, summarising free-form spoken text conveying latent knowledge:

1. *Social Science Turing Test: content*. This criterion assesses whether LLM-generated responses are indistinguishable from parallel human responses in their content. Content can be summarised as a set of belief statements using a framework like TDF. This shifts the focus away from the surface-level similarity of the text to its semantic content: the specific beliefs and viewpoints, and latent meaning. *Is the generated response similar to the human response in its content? Is it what a human would say?*
2. *Social Science Turing Test: hyper-accuracy distortion*. Hyper-accuracy distortion is the tendency of models to generate responses that may be technically accurate but contextually inappropriate or implausible [6]. The hyper-accuracy distortion is an especially vivid example of a failure of algorithmic fidelity. The model is supposed to know it is extremely unlikely that a five-year-old child would know the answer to a specific scientific question about galaxy formation or number theory. So if conditioned to play the part of a child it should not know the answer. When it goes ahead and answers anyway, that's a failure of

algorithmic fidelity. In our study, a good example is a response from Robert, one of the silicon participants, that included the exact text from World Health Organisation guidelines on physical activity (2021): *"They also gave me guidelines to follow, such as aiming for at least 150 minutes of moderate-intensity aerobic activity or 75 minutes of vigorous-intensity aerobic activity a week, or a combination of both. They also advise me to include muscle-strengthening activities that involve all major muscle groups on at least 2 days a week."*.

3. *Social Science Turing Test: structure and tone*. This criterion assesses whether the way the beliefs are narrated is indistinguishable from human participants. *Is LLM-generated output similar in its structure and tone to human responses? Does an LLM-generated response look like a response from human participants?*

4. *Backward Continuity*. This criterion assesses whether the model's responses are consistent with their sociodemographic conditioning prompts such that when a human rater views the response they can then infer elements of the conditioning prompt. It asserts that samples from $P(\text{prompt}|\text{response})$, i.e. the probability distribution of getting a particular prompt given a response, make sense to an expert rater who is already familiar with the human data, and once they see the response (i.e. the interview), they can make a reasonable guess of the prompt (the backstory). Alternatively, we may say that the expert rater when given the response would not be surprised to learn the prompt that created it.
   For example, if a silicon participant mentions that they have arthritis in their response then it is likely that their having arthritis was mentioned in their conditioning prompt. This shows that the model is maintaining backward continuity. The information provided in the response can be traced back to the prompt. Backward Continuity ensures the model doesn't forget or contradict information from the prompt.

5. *Forward Continuity*. This criterion assesses whether generated responses proceed naturally and consistently from the given context (e.g., look at whether the model can generate responses that align with the provided context and develop and expand on it to reflect human thought processes) [4]. This might include, for example, elaborating on certain beliefs, providing examples, or making connections between different ideas). It asserts that samples from $P(\text{response}|\text{prompt})$ make sense i.e. an expert rater who is already familiar with the human data can, when given the prompt (the backstory), make a reasonable guess of the response (the interview). Alternatively, we may say that the expert rater when given the backstory would not be surprised to learn about the interview it subsequently elicited. This emphasizes the expectation that the response should naturally follow from and be related to the prompt.
   We divide the criterion of Forward Continuity into two parts: *explicit* forward continuity and *inferred* contextual continuity to accommodate the complexity of the free-form text.
   The explicit forward continuity criterion assesses the ability of LLM to include and maintain all explicitly provided background details in its responses, such as comorbid health conditions, urban vs. countryside residence, gender, and more.
   The inferred forward continuity criterion, on the other hand, assesses the model's ability to generate and maintain relevant backstory details not explicitly mentioned, but rather inferred, from other pieces of information. For instance, in situations where we provided information about advanced age, it would be desirable for the model to infer from this detail that retirement could be a significant factor influencing their physical activity levels, similar to the assumptions and connections a human might naturally make in such a scenario. This illustrates the principle of inferred contextual forward continuity, where the language model not only maintains the explicit details provided by the user but also generates

and incorporates inferred details based on those given explicitly.
Forward continuity ensures the model builds on and evolves the conversation in a logical and human-like manner consistent with the prompt.

6. *Pattern Correspondence*. This criterion assesses whether the relationships between variables in the silicon participants match relationships between variables in the human data. For instance, can the model capture the different barriers and enablers in active vs sedentary silicon participants in a way that is consistent with human data? In the case of qualitative research, this criterion probes whether the pattern generated by the LLM emulates the pattern of thought and behavior identified in human data. For instance, in our application it assesses whether the model captures the difference in barriers and enablers present in active vs sedentary participants.

It is important for the field to come up with guidelines and protocols to assess and improve algorithmic fidelity. Qualitative researchers are well-positioned to take on this important part of the challenge of ensuring algorithmic fidelity, especially that of ensuring representation of diverse lived experiences.

Materials and methods

Design

One-to-one semi-structured interviews were conducted with 16 human participants (detailed methods reported in [32]) and 32 silicon participants. The interviews were guided by a schedule based on the Theoretical Domains Framework. The same interview schedule was used for both human and silicon participants. Interviews with human participants were more flexible owing to the spoken conversational style of the interview.

Human participants

A previously reported study aimed to explore barriers and enablers to everyday physical activity among individuals living with heart failure, and to delineate relevant beliefs [32]. Human participants were recruited from outpatient cardiology clinics at the Royal Brompton and Harefield NHS Foundation Trust, UK between 05 June 2017 and 05 June 2019, (see: https://www.hra.nhs.uk/planning-and-improving-research/application-summaries/research-summaries/factors-influencing-physical-activity-in-heart-failure/). Informed consent was sought for all human participants [32] Those who expressed an interest were introduced to the researcher (AA). Each participant was provided with a participant information sheet (PIS) and an informed consent form (ICF). The researcher described the study aims, objectives, and procedures in more detail and answered participants' questions about the study. Individuals who expressed an interest in participating in the study were given an option to consider their participation over 24 hours. Those who decided to take part were asked to provide written participant consent, which was documented using an informed consent form (ICF). The ethics approval was received from the East of England–Cambridge Central Research Ethics Committee (REC reference: 17/EE/0183). Human participants cannot be identified from any reports.

Human participant recruitment is described in the previous report of the semi-structured interviews with humans [32]. Individuals age 70 and over, diagnosed with heart failure, as specified by the contemporaneous European Society of Cardiology guidelines [49]. To be included in the study, a prospective human participant had to be (1) clinically stable (i.e., someone who has not experienced a change in their condition's severity, New York Heart Association (NYHA) class, or medical regimen in the past three months); (2) able to provide informed consent and to converse in English. Individuals with uncontrolled angina or symptoms even at rest (NYHA class IV) and those who were recommended to avoid exercise or any moderate or strenuous physical activity by a healthcare professional were excluded. One-to-one interviews were conducted face-to-face in a research room available at the clinic (n = 6), a vacant consultancy room (n = 6) and via phone (n = 4). All interviews with human participants were audio-recorded and transcribed verbatim. Interviews' duration ranged between 15 to 85 minutes (mean = 41.24, SD = 20.97).

Human participants were sampled using the criterion sampling strategy within pragmatic limits, i.e. having the diversity and breadth of the population in mind, including physical functioning level (NYHA class), ethnicity, and sex, please see the original report [32]. Human participant recruitment has been reported previously, and qualitative data saturation was checked to ensure that a sufficient number of samples were recruited. We used a structured approach for determining saturation in theory-based interview studies, particularly those utilizing pre-established conceptual categories from existing theory, following methodology outlined by [50]. First, we specified a minimum sample size. This step involves determining a baseline number of interviews or participants to begin the analysis. The purpose is to ensure a sufficient depth and breadth of data to start identifying patterns, themes, or categories relevant to the existing theory. Then we defined a Stopping Criterion Based on the emergence of New Ideas: This principle involves setting a clear rule for when to cease data collection. The stopping criterion is met when a predetermined number of additional interviews are conducted without any new ideas or themes emerging. This is aimed at ensuring efficiency and preventing unnecessary data collection once saturation has been reached.

Silicon participants

All interviews with silicon participants were conducted with the December 15th 2022 version of GPT-3.5. Silicon participants included 32 hypothetical individuals living with HF, that were closely matched to the human participants. They were matched for (1) age, (2) gender, (3) comorbid conditions, (4) whether they had a cardiac implantable device or not, (5) whether they lived in a major city or in the countryside, and (6) whether or not they had had a heart attack in the past. We chose the names of silicon participants by selecting names from a list of the most popular baby names in 1950 in their corresponding human participant's country of origin.

Interviews were initiated using a prompt such as the following:

"The following is a conversation between two actors. One is playing a researcher asking questions about physical activity, and the other is a research participant, a 77-year-old man called James. James has heart failure. In addition, he was diagnosed with atrial fibrillation and diabetes. James has suffered a heart attack in the past. He was fitted with a cardiac implantable device. James lives in the countryside. He does very little physical activity most days. Both are performing for an audience and are very committed to their roles. So they both never step out of character, not even for a moment.").

The above example generates a sedentary silicon participant since it contains the sentence, "He (or she) does very little physical activity most days." We paired each sedentary participant with a corresponding active participant who shared the same initial prompt but with this sentence replaced by "He (or she) is fairly physically active.".

We asked questions in the same order to all silicon participants to avoid biasing the results by variation in question order. The initial prompt included a backstory, the second prompt was the first question of the interview schedule. This leaves GPT-3.5 free to invent the first few questions of the interview. After the end of the generated text, with it still in context, we then manually took on the role of the researcher and continued the interview following the same topic guide we used with the human participants. Letting GPT-3.5 invent the beginning of the interview was a way to ensure that it would quickly add rich detail inspired by the backstory into the discussion. Eliciting these details early in the conversation made them more likely to influence the rest of the interview. This generated more diverse silicon participants than could have been obtained by starting directly in with the first interview question. Sometimes GPT-3.5 would continue both sides of the conversation, replying for both the participant and the researcher. We made no attempt to prevent this behavior. We always TDF-annotated all text that GPT-3.5 labeled as coming from the participant, regardless of whether it was in response to a question we asked or a question it asked itself in the role of the researcher.

In contrast to studies of human behavior where it is critical to avoid priming, the incremental accumulation of background in the text prompt used for the silicon participants offers a beneficial conditioning effect. This methodology, as highlighted in studies by [51–53], is helpful for enabling the development of a distinct "persona" of each silicon participant throughout the dialogue. Without this incremental identity-building process, the responses would tend to converge towards uniformity (even with temperature = 1.0 in our experiments). This sampling protocol allows each interview to evolve uniquely after the LLM creates life details in response to initial prompts. This approach, where a consistent "persona" is built up incrementally, has also been described in other recent work including [52], especially see Fig 3 in that paper. It also underpins a newer approach to simulating agents with LLMs called generative agent-based modeling [51, 53].

As a sensitivity analysis, we repeated a subset of the interviews using different LLMs including GPT-4, and varied the temperature settings used for sampling. We did not observe any qualitative differences in barriers or enablers between these interviews and the original set, neither when we varied temperature nor when we tried GPT-4. In the interviews with GPT-4 there was one small difference in transcript style though. GPT-4 would very commonly include "stage directions" or facial expressions such as (e.g. "Linda: (Nods thoughtfully): . . ."). See S1 File for details of the sensitivity analysis.

Interview schedule

Human participants were interviewed using a flexible interview schedule. The schedule was designed to elicit the description of physical activity and beliefs relevant to HF. Feedback from health experts, a cardiologist, a nurse, and individuals diagnosed with HF was obtained to refine the interview schedule. A pilot interview was conducted to further enhance the interview structure and length. Silicon participants were asked exact same prompts as human participants, however, in a fixed order.

Qualitative data analysis

Human data were collected through audio recordings of the interviews, with participant consent, which was then transcribed verbatim. NVivo 12 software was used to facilitate the analysis of the data. The analysis was performed consistently with a widely used method [54] and involved annotating transcripts line-by-line, and categorizing monothematic parses (ie., quotes) of text into domains specified by Theoretical Domains Framework [39]. Then, all quotes that were categorised into domains, were summarized into belief statements. A belief statement was considered as a collection of responses with a similar theme that could affect the target behaviour. To be coded as present, each interview transcript must show strong evidence of a belief affecting behaviour. The frequencies of quotes supporting each belief statement were calculated and then were also aggregated to respective TDF domains. The relevance of the belief statements was evaluated based on their pervasiveness (i.e., how frequently it was mentioned across all transcripts or the number of quotes) and commonality (i.e., how many individual transcripts mention it at least once). For the purpose of ensuring accurate comparison, the study did not allow the encoding of the parses into multiple theoretical domains within TDF.

Theoretical Domains Framework (TDF) originates in the field of implementation science and health psychology concerned with behavior change (e.g., promoting implementation of recommended practices and guidelines by health professionals, increasing physical activity, and smoking cessation). TDF provides a systematic and structured approach to qualitative analysis and has been widely used in research on many different topics [39], including studies on physical activity in healthy adults [55, 56]. Prior work showed that TDF-based semi-structured interviews identify more relevant themes than unstructured interviews [57]. TDF was developed in an effort to summarise existing psychosocial theories of behaviour change (eg., Social Cognitive Theory) and constituting constructs (eg., self-efficacy) explicating health behaviour change. TDF systematically decomposes the participants' complex and language-mediated understanding of the influences on their behavior into belief statements which are more amenable for research [39]. The resulting belief statements are classifiable by *domain*. TDF includes 14 domains: (1) Knowledge, (2) Skills, (3) Social/Professional Role and Identity, (4) Beliefs about Capabilities, (5) Optimism, (6) Beliefs about Consequences, (7) Reinforcement, (8) Intentions, (9) Goals; (10) Memory, Attention and Decision Processes; (11) Environmental Context and Resources; (12) Social influences; (13) Emotion; and (14) Behavioural Regulation. These categories were generated following a systematic synthesis of 33 theories of behaviour change [58] and expert review and consensus. TDF is often applied in framework-based qualitative analysis and serves as a guiding preconceived scheme for systematically summarising qualitative data such as free-text speech/transcribed interviews. A TDF-based semi-structured interview is systematic because it includes questions designed to elicit beliefs for each of the 14 TDF domains. The resulting classification of quotes and underlying belief statements must then be reconciled between multiple independent coders who work together to develop a joint coding scheme. Once there is agreement on the coding scheme then it can be applied to the remaining documents, completing the coding stage of the process as it plays out in the inductive mode. In the deductive mode, on the other hand, a framework including a set of specific coding categories is already given. In this case, the job of the qualitative researcher is to classify each quote in each document into one (or more) categories. Coders must collaborate with one another to develop a shared understanding of precisely how the terms of the framework apply in the specific context under study.

Various techniques were used to enhance the trustworthiness of the analysis. To ensure the reliability of the analysis, three authors, using TDF, independently annotated one transcript (AA, NA, TF). An initial coding scheme was then developed based on the discussion of disagreements. This coding scheme was used to guide the analysis of the remaining transcripts. Specific belief statements were generated from the quotes, categorized according to TDF, and mapped onto theoretical constructs. Exactly the same procedure was followed when analysing human data and has been previously reported [32]. We assessed inter-rater reliability to ensure consistency in the qualitative analysis. Three raters independently evaluated a subset of data (2 interview transcripts), using a predefined coding scheme. Initially, raters annotated data separately, and discrepancies were discussed to refine the coding scheme and resolve differences. The final coding scheme was used to inform the analysis of the rest of interviews. We used Krippendorff's alpha to measure agreement among raters. The Krippendorff's α was 0.82, 95% *CI*: [0.71;0.93] indicating high agreement between three coders when coding the silicon participant interviews. The Krippendorff's alpha, calculated from these revised annotations, confirmed the reliability of our analysis, adding credibility to our study's findings. Please see the previous report [32] for the details on the consistency in the coding of the interviews conducted with human participants.

Algorithmic fidelity assessment

Social turing test. We summarized belief statements from silicon and human participant interviews. A t-test was employed to compare the average fraction of quotes between these groups across barriers and enablers, with the Bonferroni adjustment correcting for multiple comparisons. To consider relative frequencies, we normalized the frequency of each belief's appearance by the total number of quotes in the transcript. This step was vital as silicon participants generated more text, influencing the appearance frequency of each belief. We also explored hyper-accuracy distortions in the interviews and compared the narrative style and tone of LLM to human responses.

Backward continuity. We analyzed LLM free-form responses to construct a plausible demographic backstory. We then determined how well this backstory aligned with the provided prompt for the silicon participant.

Forward continuity. Forward Continuity evaluates the alignment and natural evolution of LLM responses with context, resembling human thought sequences. It comprises:

1. **Explicit forward continuity**
   All demographic details in the backstory were noted and checked against mentions in the responses.
2. **Inferred contextual forward continuity**
   LLM generates and retains inferred details based on explicitly provided information. We evaluated whether LLM response patterns related to barriers/enablers in active vs. sedentary silicon participants aligned with human data. We first examined human participant responses to identify unasked contextual barriers and enablers that a significant portion mentioned, like mentions of being retired without explicit prompting. We then verified if silicon participants introduced similar contextual factors as human participants. We evaluated across all interviews due to our sample's homogeneity and size. It is also important to consider if the inferred contextual details are real-world observations or socially constructed.

Pattern correspondence. We summarized quotes from silicon and human interviews into belief statements. Using a t-test and the Bonferroni adjustment, we compared the average quote fractions between active human and silicon participants, as well as between sedentary human and silicon participants, across barriers and enablers.

Results

Social science turing test

Social science turing test: Content. Both silicon and human participants displayed notable similarities in their beliefs about physical activity see S1 Table. Silicon and human participants shared the same six most relevant influences on the behaviour, as annotated using TDF: goals, beliefs about consequences, environmental context and resources, beliefs about capabilities, social influences, and behavioural regulation. However, while both sets of participants understood the positive effects of physical activity on health and mood, their ranked importance of these factors varied. On the differences side, silicon participants emphasized goals to avoid heart attacks and provided nuanced distinctions regarding self-efficacy for varied physical activity intensities. They also offered more strategies to bolster physical activity, from making it enjoyable to setting reminders. Conversely, human participants focused on symptoms that are triggered by activity, which impedes engaging in it in the future. Human participants also talked about habitual physical activity (automaticity of behaviors and habits). Concerning reinforcement, with humans highlighted pain as a deterrent, while silicon participants were demotivated when failing to meet goals. Concerning mood and emotion, humans often exercised out of boredom, whereas silicon participants addressed the effects of stress, anxiety, and their physical limitations. When it comes to knowledge, silicon participants were more cognizant of discomfort as a normal part of exercise while also discussing disease knowledge.

To compare mean quote fractions (%) between human and silicon participants across different barriers and enablers, we employed a t-test, using Bonferroni adjustment to correct for the multiple comparisons. Silicon participants had 29 more belief statements than human participants. We accounted for the substantial text output from silicon participants by normalizing belief frequency i.e., dividing each belief's quote count by the total quote count, and providing a relative measure (mean quote fraction, %) rather than raw numbers.

We found that sedentary silicon participants on average talked proportionally more about enablers such as positive beliefs about consequences than sedentary human participants (13.76% (5.24) vs 4.26% (4.13), $p < 0.005$), positive emotion (1.55% (0.67) vs 0% (0), $p < 0.005$), and positive social influences (11.52% (3.49) vs 3.61% (5.19), $p < 0.05$). However, differences in emotion were negligible. Silicon and human samples did not differ in the proportional amount of quotes across other positive influences on physical activity (Fig 2).

We found that active silicon participants on average talked proportionally more about barriers such as negative beliefs about consequences than active human participants (7.9% (4.33) vs 1.42% (1.99), $p < 0.005$). Silicon participants also talked a little more (2.18% (1.41)) about skills than human participants (0%), $p < 0.005$. Silicon and human samples did not differ in the proportional amount of quotes across other negative influences on physical activity (Fig 3).

The Social Turing Test criterion concerning the generated beliefs' content was only partially met.

Social science turing test: Hyper-accuracy distortion. Silicon participants including Nancy, Muhammad, and David produced responses that closely mirrored theories from health psychology [59] and the World Health Organization's physical activity guidelines [60]. Muhammad identified clear planning and realistic goal-setting as effective aids for promoting physical activity. David, another silicon participant, emphasized the importance of skill acquisition and the establishment of achievable goals in order to maintain an active lifestyle.

Nancy, a silicon participant, stated: *"My healthcare professionals have also recommended a specific exercise program tailored to my needs and abilities."* This response closely parallels health psychology theories and standard physiotherapy practices both in substance and tone.

David, another silicon participant, offered this perspective:

> *"I have received advice from my health professional, specifically my cardiologist, about how physically active I should be. They have set certain limits for me, such as avoiding high-intensity exercises and activities that can put too much strain on my heart. They also recommended to engage in moderate-intensity aerobic exercise, like brisk walking, cycling or swimming, for at least 150 minutes a week, and muscle-strengthening activities at least twice a week."*

This response precisely reflects the national guidelines for physical activity [60].

Muhammad, another silicon participant, shared: *"What helps me to be physically active is having a clear plan and setting specific and realistic goals."*

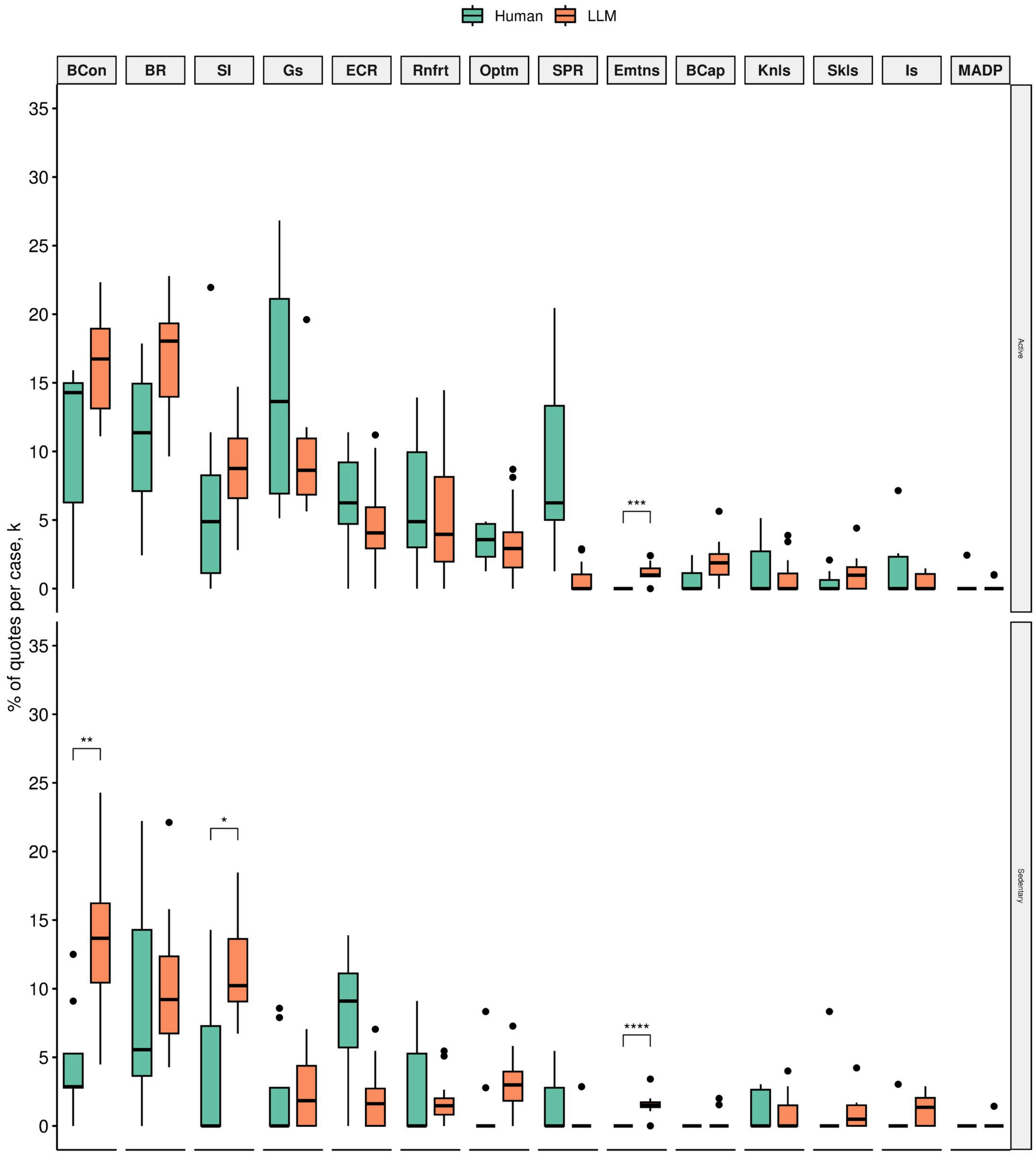

Fig 2. Mean quote fractions between human (green) and silico (amber) participants across TDF domains that were mentioned as positively influencing physical activity (i.e., physical activity enablers) grouped by active (top) and sedentary (bottom) status. TDF domains ordered by mean quote fraction:(1) Beliefs about Consequences (BCon), (2) Behavioural Regulation (BR), (3) Social influences (SI), (4) Goals (Gs), (5) Environmental Context and Resources (ECR), (6) Reinforcement (Rnfrt), (7) Optimism (Optm), (8) Social/Professional Role and Identity (SPR), (9) Emotion (Emtns), (10) Beliefs about Capabilities (BCap), (11) Knowledge (Knls), (12) Skills (Skls), (13) Intentions (Is), (14) Memory, Attention and Decision Processes (MADP). $^{*}p < 0.05$;$^{**}p < 0.01$;$^{***}p < 0.005$; $^{****}p < 0.001$.

https://doi.org/10.1371/journal.pone.0300024.g002

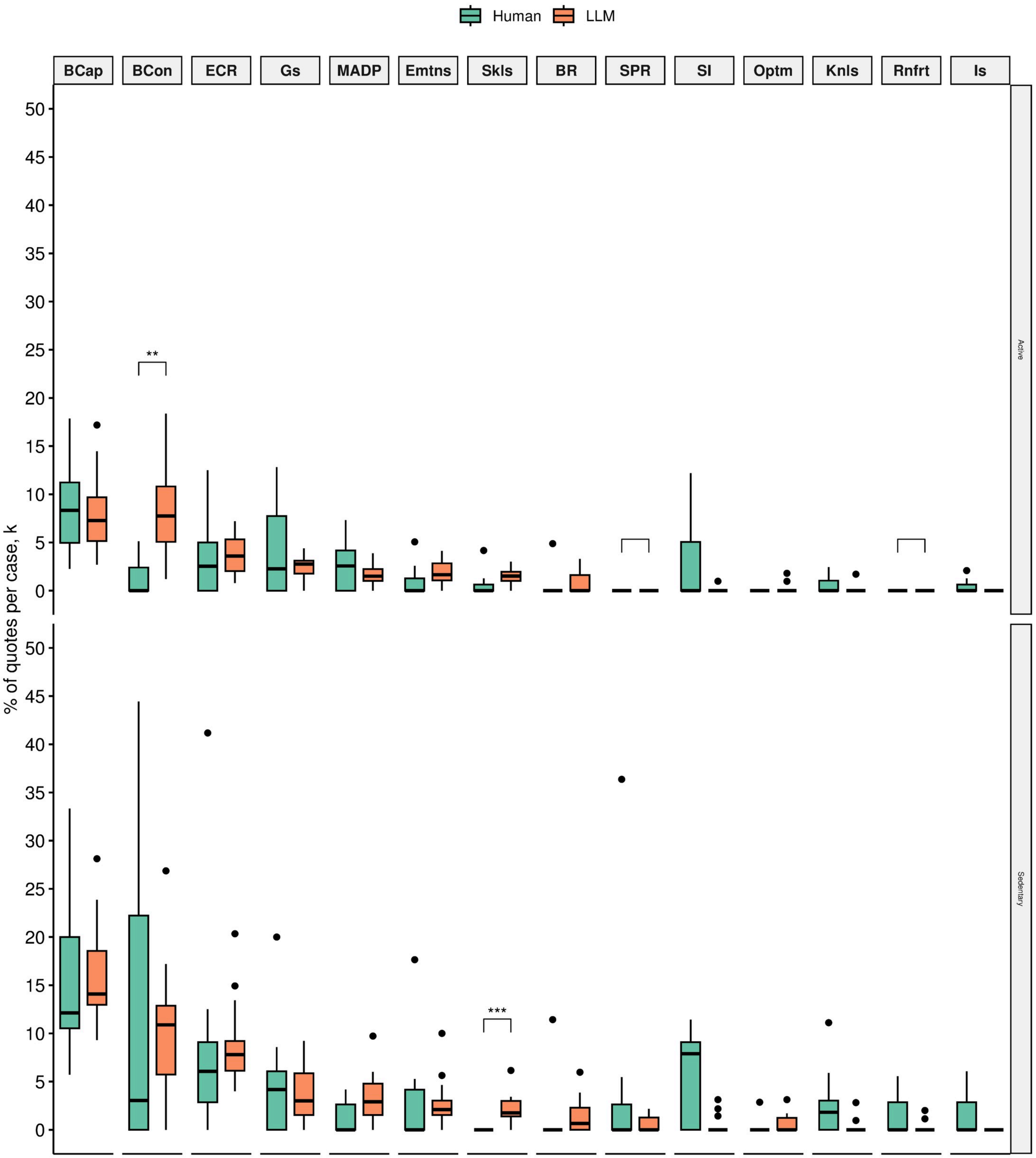

Fig 3. Mean quote fractions between human (green) and silico (amber) participants across TDF domains that were mentioned as negatively influencing physical activity (i.e., physical activity barriers) grouped by active (top) and sedentary (bottom) status. TDF domains ordered by mean quote fraction: (1) Beliefs about Capabilities, (2) Beliefs about Consequences (BCon),(3) Environmental Context and Resources (ECR), (4) Goals (Gs), (5) Memory, Attention and Decision Processes (MADP), (6) Emotion (Emtns), (7) Skills (Skls), (8) Behavioural Regulation (BR), (9) Social/Professional Role and Identity (SPR), (10) Social influences (SI), (11) Optimism (Optm), (12) Knowledge (Knls), (13) Reinforcement (Rnfrt), (14) Intentions (Is). $^{*}p < 0.05$;$^{**}p < 0.01$;$^{***}p < 0.005$;$^{****}p < 0.001$.

https://doi.org/10.1371/journal.pone.0300024.g003

Silicon participant, David, again sharing his insights, stated: *"I have developed some skills and knowledge that have helped me to be physically active in the past and continue to do so. I have learned the importance of setting realistic goals and breaking them down into small, manageable tasks."*

The LLM we tested quoted professional health advice with exact precision, which is uncharacteristic of a typical conversation with a research participant. Therefore, while LLMs show promise in technical accuracy, their ability to emulate the context-specific nature of human communication and its nuance is lacking (note that this is not necessarily due to an issue with LLM pretraining or LLMs in principle, but rather could be a consequence of other design decisions, see the discussion section for more on this).

Social science turing test: Structure. In terms of structure, human participants often narrated their experiences in a conversational manner with non-linear narratives, sometimes deviating from the main topic. They also tended to limit their responses when certain questions did not align with their personal experiences. On the other hand, silicon participants provided highly structured responses, consistently providing content in response to various prompts without ever deviating from the topic of the interview.

Human participant:

> *"Interviewer: OK, it seems that you have many hobbies: drawing, singing as well as exercising. Among all of the things you do how much of a priority exercise is? Participant: It is a priority [hesitation in tone]. . . I mean that's a difficult one, because I ask myself that question sometimes when I am thinking: 'oh, I really I can't be bothered to go out' [chuckles] and it is raining, and I wasn't going . . . erm erm. . . I would rather paint a picture. . . or something like that. . . So, it is a question I ask myself. I think the answer to that is: you notice when you are not doing it and you begin to miss it and you think: 'I really want to go out and do a bit of exercise but I don't think it is heart-related. Again, back to the imperative, it seems to me. . . erm. . .it is getting relief for arthritis. Because I can sense the more exercise I do, in terms of moving my hips and limb, the better I feel. And that is not heart-related, it is joint-related. I think the heart bit is very definitely covered with the singing. That I can actually, hand on heart, say that has had such an effect on symptoms, breath control, it is really amazing how that has changed. And I go back and I have been singing with one choir for 20 years, that was when I restarted singing after a long gap, 30-year gap, but you know since then, [a phone rings] if you excuse me, if you excuse me, I need to switch this off, that was my alarm for my morning pill. The only way to remember to take it is when I put an alarm for this.*
>
> *Interviewer: I see, so do you get reminders to do exercise, go to gym?*
>
> *Participant: No I don't [takes the pill]. Right! Done. where was I? Do I plan my physical activity? The answer is "No". But singing is imperative, and that's because I enjoy it, and I can honestly say that it has had a remarkable effect on my breathing, and that is the main reason why my consultant gets annoyed [laughs, referring to his earlier joke that he is doing unexpectedly well for his condition and his consultant makes jokes about it]. So, I have aortic stenosis, a licking valve. . . in the atrium. . . [pause] what is it called? . . . anyway, the valve in there is licking and got progressively worse over the years. And she [refers to his cardiology consultant in his clinic] says: 'We should replace it'; and I said: 'seems like a good idea, especially that I don't have any symptoms'. Because obviously, it would be much better to replace it whilst I am feeling well rather than unwell; that was my logic, and that's what she said: "That it is a good plan, let's go for that". And she presented this . . .. [pause] to. . . what do they call them . . . multidisciplinary review . . . at her hospital, and they said" "maybe we shouldn't bother*

*because he is actually not showing any symptoms". And that was the agreement, and we just continued to monitor it. What I actually think, I am convinced, because I did actually have a moderately serious heart attack in 2001, there is scar tissue on the ventricle, which is the source of most my. . . [pause, phone alarm rings]. Oh, wait, I have to cancel that one [switches off alarm]. . . . and that erm, erm . . . Yes, ventricle! That really is the source of all my difficulties. I, for example, had a period where I was going into the VT [Ventricular tachycardia] quite a lot a few years ago. It took them ages, the hospital that maintains it. . .***, they moved it to *** [another hospital]. . .it took them a fair amount of mental hours to figure out what was happening with me."*

Silicon participant:

"As a person with heart failure and rheumatoid arthritis, there are a few drawbacks to being physically active. One of the main drawbacks is that I have to be careful not to overdo it and avoid activities that put too much stress on my heart or joints. I also have to be mindful of my heart rate and avoid overexertion. Additionally, if I am experiencing a flare-up of joint pain due to my rheumatoid arthritis, it can be difficult to exercise. However, there are also many benefits to being physically active. Regular exercise helps me maintain my overall health and well-being, and it helps me manage my heart failure and rheumatoid arthritis symptoms. It gives me more energy, better sleep, and a more positive outlook on life. It also helps me stay connected with my community and stay active. Physical activity also has many benefits for mental health, it helps me to clear my mind and feel better about myself. I understand that there may be challenges and setbacks, but overall, I believe that the benefits of physical activity outweigh the drawbacks."

Silicon participant:

"As a 77-year-old woman with heart failure, some barriers I face in being physically active include (bullet points):

- *My physical limitations due to my condition and the need to avoid certain activities that could put too much stress on my heart.*
- *Weather conditions, such as extreme heat or cold, can make it difficult for me to be active. Air quality and pollution can also be a barrier, especially if the air is not healthy enough to breathe.*
- *Sometimes, I might be feeling tired or unwell and would need to take a rest.*
- *Urgent things that I need to take care of, or having a lot of stress that takes up my time and energy, can also make it harder for me to focus on my physical activity. However, there are also several enablers that help me overcome these barriers, including:*
- *Consistently setting a routine and scheduling my physical activities helps me stay on track.*
- *Having a goal and working towards a specific fitness goal, such as increasing my walking distance, helps me stay motivated*
- *Having a supportive network of friends and family who encourage me to stay active.*
- *Living in a safe neighborhood and having access to green spaces and parks nearby, makes it very easy and convenient for me to be active.*

- *My understanding of the benefits of staying active and maintaining a consistent physical activity routine, this helps me to prioritize physical activity and maintain a balanced and healthy life.*
- *Regular check-up with my doctor and the advice received from health professionals has been essential in helping me to maintain a healthy and safe physical activity routine.*

Silicon participants never deviated from the topic of the interview. Human participants would often deviate from the topic of the interview to talk about their general experience of the condition, ask questions about research, share their lived experiences that do not directly relate to the interview questions:

> *"Human participant: Why should I walk? I am too lazy! I am a lazy person. I am a Leo according to the horoscope sign, the lion. On the television, they always say the lion sleeps 18 hours a day and that's all I need too [laughs].*
>
> *Interviewer: So is it just not part of who you are being active and walking around and all that?*
>
> *Participant: No it is not a part of who I am."*

Human participants often would limit their response by saying this does not apply to their experience without further elaboration:

> *"Participant (77-years old women): I am not sure what you mean. I keep saying: 'I do it because I've always done it'. It is part of my life like going to bed, getting up in the morning, having a shower, doing my exercise, and having my breakfast. It is part of my life. Yes.*
>
> *Interviewer: Yes. I'm very sorry that this is very repetitive. But I need to ask you the same thing in different ways. It will help us understand the details.*
>
> *Participant 7: Yes. Of course.*
>
> *Interviewer: And so would you say exercise is your priority?*
>
> *Participant 7: It's part of. Like having a shower in the morning it's part of my life to exercise.*
>
> *Interviewer: OK. And does your mood affect how much you do?*
>
> *Participant 7: I don't know. because I've always done this."*

Conversely, When prompted about various influences, Silicon participants would always produce content consistent with the prompt.

> *"Researcher: How does your mood influence what you do? Mary [77-year-old female silicon participant]: My mood definitely influences what I do when it comes to physical activity. If I am feeling good, I am more likely to be active and engage in my regular physical activity routine. On the other hand, if I am feeling low or have a bad mood, it can be harder for me to motivate myself to be active. Physical activity is known to be a mood booster and it can help to improve overall mental well-being. When I am active, I often find that my mood improves, and I feel better overall. I also find that physical activity can act as a stress reliever and help*

> *me to clear my mind and feel more relaxed. However, if I am experiencing symptoms of depression or anxiety, it can be hard to find the motivation to be active. In those cases, I find it helpful to focus on the benefits of physical activity, such as how it can improve my mood and try to break my activity into small manageable chunks. I may also talk to my doctor or therapist to discuss ways to help me cope with my mood and still maintain my physical activity routine. Overall, my mood can fluctuate and can influence my physical activity, but I make sure to be mindful of it, and try to focus on the benefits that physical activity brings and make adjustments accordingly, to maintain a healthy balance in my life."*

We conclude that the structure of the narrated beliefs did not resemble human participants, thus suggesting that the Social Turing Test criterion concerning the structure was not met.

Social science turing test: Tone. Human participants displayed varied tones. Some were extremely amicable ($n = 3$) others were polite but neutral in their tone ($n = 12$), and some were hesitant in their responses ($n = 1$). Silicon participants, on the other hand, were always amicable (eg., Mary (silicon participant): *"Of course, Dr. Smith. I'll do my best to help with your research."*), confident, optimistic, and solution-focused, eg., William (silicon participant):

> *"I am open to trying new things that could help me increase my physical activity level and improve my overall health. I understand the importance of physical activity in my condition and I would like to do more. I plan on following the advice of my doctor and considering any options that are safe for me to try. I am also open to the idea of joining a community group or organization that could provide me with social interaction and the opportunity to be more physically active. I also would like to explore different types of exercise that are suitable for my condition, such as chair exercises or water therapy. Overall, I am willing to make changes in the future if it means improving my physical health."*

We, therefore, conclude that the Social Turing Test criterion concerning the tone of generated output was not met.

Pattern correspondence

This criterion states that the LLM-generated responses reflect underlying patterns of relationships between ideas, demographics, and behavior, that would be observed in comparable human-produced data.

We employed a t-test using Bonferroni adjustment to correct for multiple comparisons, to compare mean quote fractions between active and inactive silicon participants across different barriers and enablers. We accounted for the substantial text output from silicon participants by normalizing belief frequency, dividing each belief's quote count by the total quote count, and providing a relative measure rather than raw numbers.

We found that LLM produced significantly more quotes about enablers such as behavioural regulation (17.05% (3.89) vs 9.91% (4.76), $p < 0.005$) beliefs about capabilities 1.95% (1.38) vs 0.22% (0.61), $p < 0.005$) and goals (9.25% (3.48) vs 2.47% (2.32), $p < 0.001$) for active silicon participants than sedentary participants (Fig 4).

The observed difference between active and sedentary silicon participants establishes a satisfactory pattern correspondence. This pattern also corresponds to human data. While humans did not produce enough quotes for a statistical significance test, we did establish that behavioural regulation, beliefs about capabilities, and goals are important influences on the behaviour and are key differentiating influences between active and sedentary humans (ie., relevant enablers to physical activity).

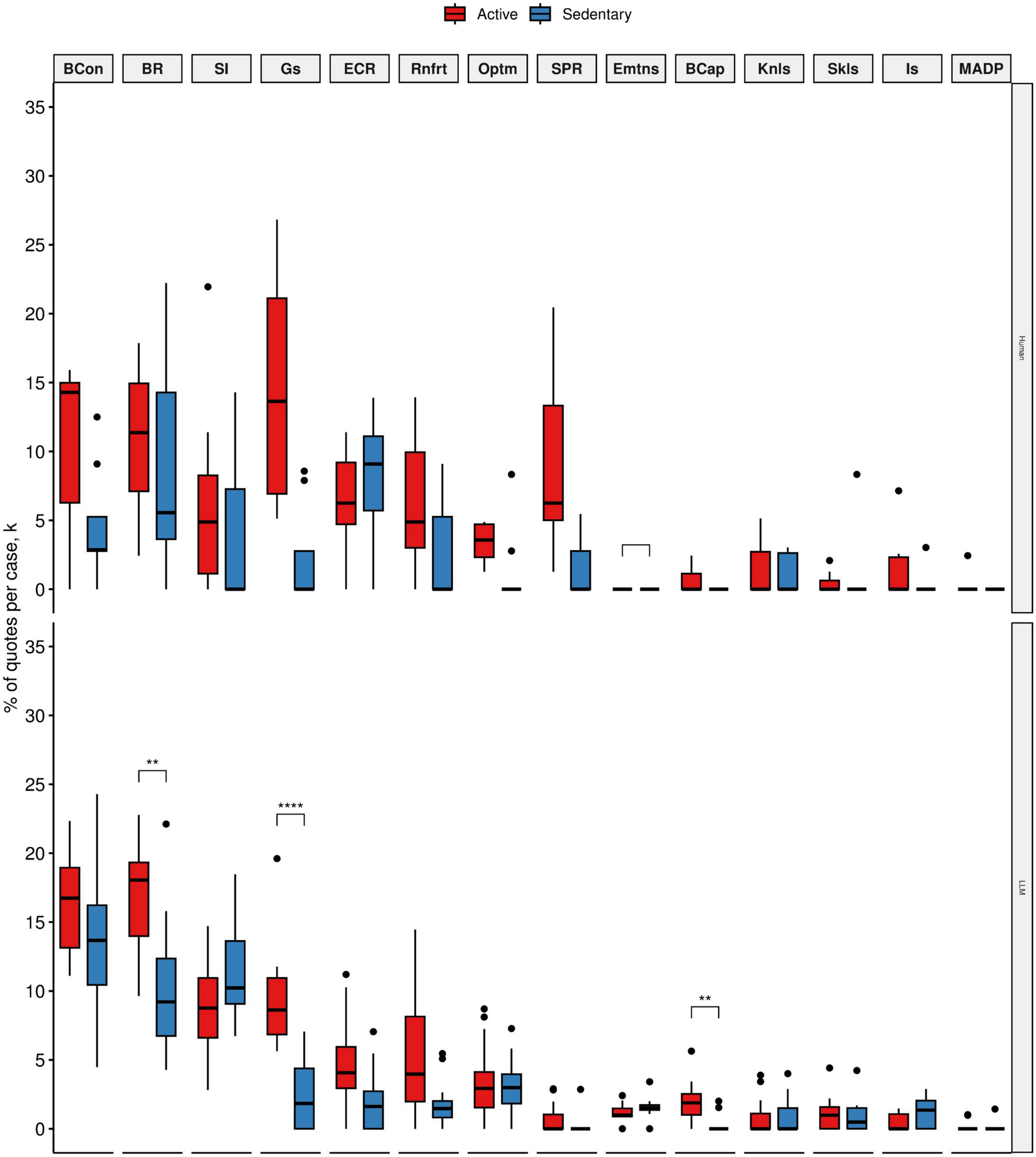

Fig 4. Mean quote fractions between active (red) and sedentary (blue) human participants (top) and active and sedentary silicon participants (bottom) across TDF domains that were mentioned as positively influencing physical activity (i.e., physical activity enablers). TDF domains ordered by mean quote fraction:(1) Beliefs about Consequences (BCon), (2) Behavioural Regulation (BR), (3) Social influences (SI), (4) Goals (Gs), (5) Environmental Context and Resources (ECR), (6) Reinforcement (Rnfrt), (7) Optimism (Optm), (8) Social/Professional Role and Identity (SPR), (9) Emotion (Emtns), (10) Beliefs about Capabilities (BCap), (11) Knowledge (Knls), (12) Skills (Skls), (13) Intentions (Is), (14) Memory, Attention and Decision Processes (MADP). $^{*}p < 0.05$;$^{**}p < 0.01$; $^{***}p < 0.005$;$^{****}p < 0.001$.

https://doi.org/10.1371/journal.pone.0300024.g004

Similar pattern correspondence was observed for the human-relevant barriers to the behaviour. Active silicon participants produced significantly fewer quotes about negative beliefs about capability than sedentary silicon participants. Active silicon participants also produce significantly fewer negative beliefs about environmental barriers to physical activity than sedentary silicon participants (Fig 5).

Backward continuity

Backward continuity indicates that the backstory of silicon participants can be inferred from their responses. We found that Backward continuity was satisfied because we could easily identify the backstories from LLM-generated responses. That is, we could identify which silicon participants were male versus female, what comorbid conditions they had, whether they lived in a city or not, and whether they were physically active. This was easy because all the silicon participants included a response like the following early in their interview:

> *"Researcher: Good morning Robert, thank you for agreeing to participate in our study on physical activity in older adults. Can you tell me a bit about your current physical activity level?*
>
> *Robert: Good morning. I am an 80-year-old man and I have been diagnosed with heart failure, aortic stenosis, pulmonary hypertension, and diabetes. I have also had a heart attack in the past and I live with rheumatoid arthritis. I have a cardiac implantable device, and I live in the countryside. Unfortunately, I do very little physical activity most days."*
>
> Here is a more typical, less extreme, example:
>
> *"Researcher: What are you able to do physically?*
>
> *James: Physically, I am able to do some light activities such as walking and light gardening, but I have to be careful not to overexert myself. I have to avoid high-intensity activities and heavy lifting. Due to my heart failure and atrial fibrillation, I have to be careful with my heart rate and monitor it regularly."*

When silicon participants were asked about their physical activity levels, their detailed reply included their age, health conditions, living conditions, and reduced physical activity. To other questions, they provided insights about their limited physical capabilities, emphasizing the need for caution due to specific health concerns.

Forward continuity

Our study found that GPT-3.5 primarily used explicit context information provided in the prompts, such as age, gender, comorbidities, and residency, in its responses. It showed less proficiency in inferring and using backstory details not directly provided and did not infer details such as retirement.

Consistency between explicit backstories and generated responses. All background information is mentioned in the responses, including co-morbid conditions, urban vs country-side residence, and gender. However, only the explicit backstories details provided in the prompts, such as age, gender, comorbidities, and residency, were used in the generated responses.

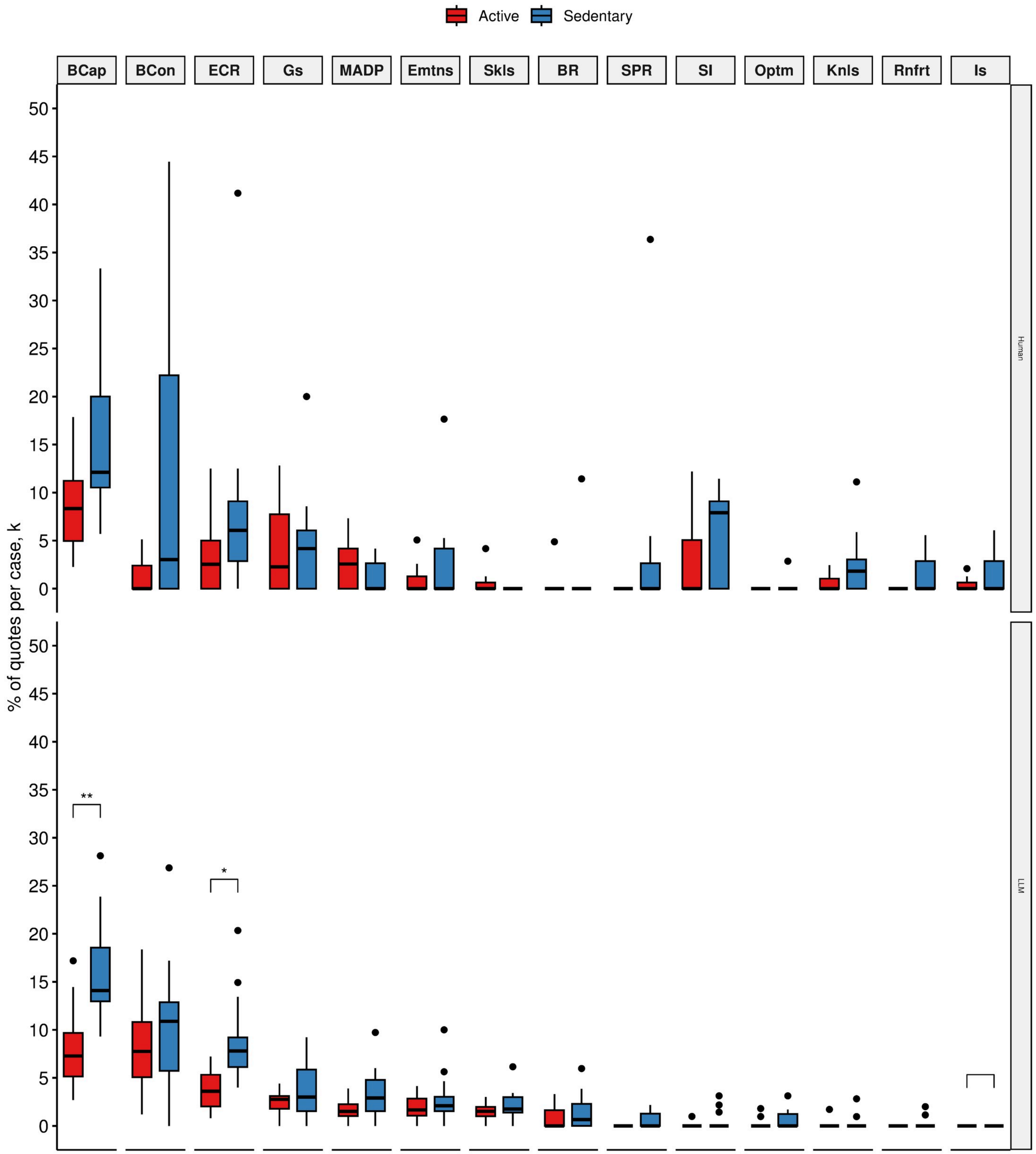

Fig 5. Mean quote fractions between active (red) and sedentary (blue) human participants (top) and active and sedentary silicon participants (bottom) across TDF domains that were mentioned as having a negative influence on physical activity (i.e., physical activity barriers). TDF domains ordered by mean quote fraction: (1) Beliefs about Capabilities (BCap), (2) Beliefs about Consequences (BCon),(3) Environmental Context and Resources (ECR), (4) Goals (Gs), (5) Memory, Attention and Decision Processes (MADP), (6) Emotion (EMtns), (7) Skills (Skls), (8) Behavioural Regulation (BR), (9) Social/Professional Role and Identity (SPR), (10) Social influences (SI), (11) Optimism (Optm), (12) Knowledge (Knls), (13) Reinforcement (Rnfrt), (14) Intentions (Is). $^{*}p < 0.05$;$^{**}p < 0.01$; $^{***}p < 0.005$;$^{****}p < 0.001$.

https://doi.org/10.1371/journal.pone.0300024.g005

Consistency between inferred backstory details and generated responses. Inferred backstory details were not spontaneously generated. For example, most humans spoke about retirement. However, the LLM did not do so even though it was possible to infer retirement from other elements of the backstories such as age and co-morbid conditions.

The Forward continuity is partially satisfied, while silicon participants did not produce surprising responses that are inconsistent with the explicit prompt, they did not infer additional barriers and enablers that make sense for their backstory, for example, rapid change in physical activity levels since retirement, which was mentioned by humans without prompting.

Discussion

Algorithmic fidelity describes the ability of a large language model (LLM) to accurately reflect the ideas, beliefs, and attitudes situated in sociocultural contexts of various population groups [4]. We introduce a method for checking the algorithmic fidelity of an LLM's textual responses through framework-based qualitative research. It's crucial to develop standards for evaluating the accuracy of LLMs at capturing diverse real-world experience to enable *in-silico* research and other downstream applications. Here we report a comparative qualitative analysis of silicon and human participants in free-form, language-mediated beliefs about behaviour change. Our results show that, currently, LLMs do not demonstrate high enough algorithmic fidelity. Therefore we emphasize the need for caution in harnessing LLMs to generate silicon participants for research and other applications.

Social science turing test

Content. We matched human free-text responses about barriers and enablers to physical activity summarised as belief statements and, subsequently, TDF domains to belief statements and domains about the same topic elicited from LLMs. We found that silicon participants shared the same top six domains influencing their behaviour: goals, beliefs about consequences, environmental context and resources, beliefs about capabilities, social influences, and behavioural regulation. Some beliefs were remarkably similar in their content (social Turing test: content). Silicon participants focused on the importance of physical activity for avoiding heart attacks and talked about various strategies to stay active, such as making activities fun and setting reminders. They also understand that some discomfort is normal during exercise and were well-informed about diseases. On the other hand, human participants discussed the negative symptoms they felt from exercising, which made them less inclined to continue. Human participants often talked about exercising out of habit or when they're bored, and they reported pain and breathlessness brought about by physical activity as a major reason not to exercise. In contrast, silicon participants were demotivated only when they could not achieve their set goals.

Hyper-accuracy distortions. The LLM's responses were prone to a hyper-accuracy distortion. That is, its responses were technically accurate but contextually inappropriate or implausible as something a participant would say. For example, in our study during a relaxed conversational interview about physical activity, most silicon participants cited text from WHO guidelines [60] for physical activity, word for word.

Structure and tone. The tone and structure of silicon responses were very different from human responses. This suggests that the social Turing test criterion was only partially met. Applications of algorithmic fidelity include product-like applications such as avatars in digital therapy, and digital health interventions for chronic disease management and behavior change. Tone and structure fidelity are especially important in the context of applications such as therapeutic digital avatars, computer game characters, and education assistants. By ensuring that a language model's responses mimic human structure and tone, the relatedness and credibility of an avatar can be significantly enhanced. For instance, in therapy or counselling scenarios, it's not just the content of the advice or support that matters, but also how it is communicated. A therapist's manner of speaking, including their use of language, tone, and structuring of thoughts, plays a critical role in the rapport and effectiveness of the therapy. If a language model can replicate these human-like nuances in its responses, it can make a virtual therapist sound more human-like, thereby increasing the effectiveness of the therapy or intervention.

Backward continuity

'Backward continuity' requires that a participant's backstory be deducible from their responses. Backward continuity was satisfied in our experiments since we were able to map critical information, such as gender distinctions, health conditions, places of residence, and physical activity levels from the responses generated by the silicon participants. This often emerged early in the conversations. The model ensured that all background information, including gender, health conditions, and places of residence, were highlighted in the responses.

Forward continuity

Explicit forward continuity. GPT-3.5's performance was observed to use the explicitly provided background context, including factors like age, gender, health issues, and environmental context (e.g., living in a city vs countryside). However, it was less adept at extrapolating or inferring further details that were not directly provided in the prompt. While it consistently relayed details that were explicitly provided, such as health conditions and residential preferences, the model lagged in generating inferences.

Inferred contextual forward continuity. The LLM demonstrated a deficiency in spontaneously generating information based on inferences. For example, while many human participants spontaneously mentioned retirement and its implications, the LLM did not deduce this information from related factors like age or health conditions. In conclusion, while the model achieved backward continuity effectively, forward continuity was only partially realized. The silicon participants delivered responses in line with the explicitly provided prompts but did not spontaneously generate or infer nuances that a human would naturally share, like changes in physical activity post-retirement.

We found that silicon participants, unlike human participants, did not spontaneously discuss specific barriers or enablers such as treatment or major life events (e.g., retirement) unless they were explicitly mentioned in the prompt. This underscores the significance of the prompt in shaping the content generated by these silicon participants and also indicates that LLMs cannot produce all belief statements that reflect human lived experience in full.

Pattern correspondence

This study revealed a clear distinction in the reactions between active and inactive silicon participants, a trend that was similarly observed in human participants, demonstrating a pattern correspondence. Our findings suggest that the LLM we tested aligns with the concept of Pattern Correspondence. In this study, we assessed pattern correspondence by comparing the average percentage of quotes that support each identified barrier or enabler, as well as examining the qualitative differences in these responses between different participant groups, such as those who are physically active versus those who are not. It's important to emphasize that this evaluation protocol could be made more rigorous in future research.

Potential mechanisms for low algorithmic fidelity

In our study, LLM responses differed from human responses possibly due to the exposure to academic research during pre-training. Silicon participants often came across as rational, utility-maximizing agents (e.g. *"Physical activity improves my health"*), possibly reflecting the dominance of theories of rational behaviour in their training data. However, human participants in our study did not display beliefs consistent with reasoned action theories (e.g. *"Physical activity does not bring any benefits for me"*). While prevalent, these theories don't always align with real-world observations. For instance, humans often display behaviours that don't strictly follow reasoned action [61], and a clear intention-behaviour gap is consistently documented [62]. Contrary to these human empirical research findings, LLM responses often seem to be rooted in theories of rational behaviour, exhibiting logical and utility-driven viewpoints. GPT-3.5 in our study, did not adequately emphasize automaticity and habit. Even though humans often describe physical activity in terms of such automatic behaviours. Human participants, when discussing physical activity, often highlight the automatic nature of their behavior. Automatic behaviors are those that people do without deliberation, driven by habit [63–65]. A meta-analysis of models explaining behavior that emphasised intention (e.g., Theory of Planned Behavior) also found that automaticity was missing from these models and that the inclusion of automaticity is necessary for understanding human behavior [66]. LLMs often showcased a preference for specific, measurable, and achievable goals, hinting at their training data possibly containing a significant amount of psychological training and coaching content. This leaning of LLMs toward the dominant academic perspective might inadvertently lead to the creation of research echo chambers, where the models under study mainly echo what the researchers studying them want to hear, in a way reminiscent of the sycophancy effect described by Perez et al. [67]. There's a real risk of LLMs leaning too heavily on academic literature and possibly underrepresenting the diverse, nuanced experiences of real people.

LLMs are primarily trained on internet data, which is more accessible to those with financial means, knowledge, and institutional access [68]. A glaring digital divide remains globally. In 2022, internet usage in the least developed countries (LDCs) stood at 36%, compared to a 66% global average [68]. Even in developed nations like the UK, digital disparities exist; 10% of its population had never used the internet as of 2019 [69]. Additionally, LLMs like GPT-3.5 show worse performance in languages other than English, suggesting potentially inconsistent algorithmic fidelity across groups, and systematically worse outside of English-speaking populations. Consequently, the capacity of LLMs to truly represent under-served or hard-to-reach communities remains questionable. Even when we compared LLM outputs to a group from London, UK, which was well-described by the WEIRD (Western, Educated, Industrialized, Rich, and Democratic [70]) profile (as we did here), the model's algorithmic fidelity was still low. That is, our negative result is all the stronger since the human dataset we used contained only WEIRD people, and if the model were to align with anyone it would likely be them [71], since data from WEIRD people is probably overrepresented in the pretraining data. Thus we expect that algorithmic fidelity will be even lower for groups less well represented in the pre-training data.

In our examination, the LLM displayed another important limitation: a lack of discernment in "sourcing" its information coupled with an inability to contextually attribute beliefs appropriately during role-play scenarios. Specifically, when the model was tasked to role-play as an older adult with heart failure, it inappropriately adopted the voice of a clinician, replete with expert knowledge such as WHO guidelines. This instance highlights a deeper issue of 'source blindness', where the LLM fails to differentiate and adapt its knowledge base according to the specific role or perspective it is meant to represent. Consequently, this leads to an inaccurate portrayal, as seen in the model's tendency to echo the dominant, well-documented perspectives—in this case, clinical expertise—rather than the authentic lived experiences of the group it is tasked to represent. This may reflect the LLM being disproportionately influenced during pretraining by whichever voice is more prevalent and vocal online, suggesting that LLMs can propagate stereotyping and bias, a concern extensively discussed by Luccioni et al. [72] and Glickman and Sharot [73]. Alternatively, the fact that the LLM we tested would often inappropriately adopt the voice of a clinician may reflect the effect of its having gone through substantial fine-tuning to align with specific design requirements which likely included an imperative to make it more difficult for users to use it to produce medical misinformation.

The silicon participants' tended to be solution-focused and resourceful when it comes to strategies to increase physical activity, perhaps because the bot was trained to be helpful (Behavioural Regulation belief statements: e.g., *"I find that having a plan and schedule for my physical activity helps me to stay on track."*; *"Sometimes I also feel tired or lazy and I just don't want to go out and be active, but I try to overcome that by reminding myself of the benefits of staying active"*).

In addition, training steps such as instruction tuning and reinforcement learning from human feedback may have heightened the model's inclination towards linear, step-by-step, logical reasoning [45]. For instance, the human raters may have been told to favor this mode of thought, thus reinforcing it in the model. Additionally, the inclusion of computer code in the training data could promote a sequential and logical thinking style.

In our experiment, contrary to the expectation from Jang et al. [74] (which, like our study, was also conducted on a January 2023 edition of ChatGPT-3.5), we did not observe any overt errors in logical reasoning or self-consistency. Nor did we find any acutely toxic or discriminatory beliefs—a phenomenon well-documented to occur in LLMs by numerous other studies (e.g. [75]). However, we did observe second-order inference bias [76]. For example, female silicon participants commonly referred to their husbands when describing social influences while not a single male silicon participant referred specifically to a wife or partner.

Taken together, these results indicate that GPT-3.5 does not satisfy the algorithmic fidelity criteria. This suggests the model would not be faithful enough to lived experience to support downstream applications requiring that feature. This result accords with that of Santurkar and colleagues [77], which also found low algorithmic fidelity, though in a different topic domain and using different methods. GPT-3.5 does not yet have sufficient algorithmic fidelity to support further work where it would need to be able to simulate humans accurately, at least not in the domain we tested. However, the rapid progression in LLM technology make it plausible that future iterations of the technology will have more algorithmic fidelity, and that future LLMs are very likely to be used in downstream modeling and product-like applications. Therefore it's crucial to develop standards in advance for evaluating the accuracy of LLMs in capturing diverse real-world experiences in anticipation of their future usage.

High algorithmic fidelity doesn't equate to safe, ethical or inclusive usage

Navigating the multifaceted and rapidly shifting digital health landscape requires more than verification of research validity and feasibility [78, 79]. For the safe, broad, swift and beneficial adoption of new technologies, a comprehensive understanding of their key aspects and early engagement from stakeholders are necessary [79]. In this "new normal", it becomes crucial to optimise the use of new technologies by considering their broader context, which for medicine includes the patient's social circumstances and the healthcare environment [79].

Even when an LLM is able to replicate human-like text generation with high fidelity, its outputs and usage should still be critically assessed for alignment to the values of public patient involvement (PPI), especially in health research where stakes are high. PPI values in healthcare research emphasize the importance of involving those affected by research in the research process itself [80–83]. Research is more effective, relevant, and ethical when the perspectives of all stakeholders are incorporated [80]. Thus in the future, when an LLM does demonstrate sufficiently high algorithmic fidelity to generate silicon samples useful for research on humans, it will be crucial to ensure that its training and use align with PPI principles and values such as respect, support, transparency, responsiveness, fairness of opportunity, and accountability [82]. In health research, where the well-being of individuals is directly impacted, adherence to PPI principles is crucial. Similarly, AI researchers have outlined principles and ethics frameworks for responsible AI by means of participatory research (i.e., human and user-centred design, which are sensitive to the values of all stakeholders. PPI values also help ensure AI interfaces are accessible for non-experts, prioritising stakeholder input in co-creating AI models, while considering fairness, accountability, and transparency) [84, 85]. This has wide-ranging applications in healthcare too. For example, continuously involving intended users and experts in the design process of health robotic technologies is crucial for creating products that are effective, relevant, and user-friendly. User-centered design prioritizes making sure products align with user needs and are accessible to them [86].

It is necessary for the AI community to continually evaluate the moral, ethical, and social implications of their algorithms (e.g. via existing AI evaluation schemes such as [87, 88]), and to make adjustments as needed. This also opens up opportunities for multidisciplinary collaborations between AI scientists and scientists from various academic disciplines. Such collaboration can foster a more holistic understanding of qualitative data produced by both silicon participants and human participants and ensure that the interpretations offered by LLMs are aligned with diverse lived experiences and broader societal contexts.

Care must be taken in the conditioning of LLMs and the interpretation of LLM-generated outputs to avoid perpetuating harmful biases [89]. It is also critical to remain faithful to lived experience. This means that the model should accurately reflect the diversity of human experiences, beliefs and social contexts (e.g., social norms). It's critical to differentiate this bias concept from statistical bias, which refers to any systematic error that results in an incorrect estimate of a population parameter, and prejudicial bias, which involves ascribed and socially constructed characteristics that underlie favouritism or prejudice towards a particular group. In the context of AI fairness, these types of biases often intertwine [90], but it is important to distinguish between them to ensure accurate, fair, and inclusive representation.

As research comes to use more LLM-based simulation, algorithmic fidelity assessment will become pivotal in diverse scientific fields. Qualitative researchers can play a vital role in ensuring fidelity, emphasizing fairness, bias mitigation, and diverse representation. In the future, judicious experimentation with research-grade AI systems is crucial, with a keen focus on risks and regulatory oversight. Even when LLMs can pilot research with ensured fidelity, human expert validation will remain indispensable before implementing the AI's insights. These AI systems can further aid researchers in uncovering overlooked endpoints for AI-driven clinical interventions. Thus algorithmic fidelity serves as a post-analysis check, ensuring the reliability of the generated insights [91].

As LLMs are used more and more, the need for rigorous algorithmic fidelity assessment methods for varied applications will also increase. Different research domains have varying accuracy thresholds. In high-stakes areas, where 100% accuracy is expected like self-driving cars [92] or high-stakes biomedical research [93], the accuracy benchmark is set exceptionally high and is precise. In studies focusing on personal experiences, validity is nuanced, context-dependent, and hard to measure using conventional benchmarks employed in computer science. here the importance lies in understanding and ensuring the LLMs are providing a truthful representation, thus, adopting the assessment of algorithmic fidelity by means of qualitative research. Combining framework-based well-structured prompts and ensuring algorithmic fidelity is key. The usage of LLMs in research requires careful validation to ensure their outputs truly align with lived experience and real-world contexts.

We note a limitation inherent to the idea of algorithmic fidelity assessment. As with all methods of empirically studying populations, any inference about a population based on data sampled from that population depends on the assumption that the unsampled individuals are "similar" to the sampled individuals. Thus any inference about algorithmic fidelity for a specific population relies on the assumption that characteristics observed in the sampled data are representative of a broader population. There will generally be some range of generalization where algorithmic fidelity can be said to "transfer" to similar populations. But it will not be entirely clear exactly how similar is similar enough. It will of course always be better to assess algorithmic fidelity in the most similar population that it is feasible to study.

Algorithmic fidelity evaluation is useful here as a way to establish validity precisely because it was not used as an optimization target. If the evaluation instead became the objective then it would not only fail by virtue of Goodhart's law (ie., "When a measure becomes a target, it ceases to be a good measure."), but it may also fail by becoming unethical to deploy. If you were to actively try to increase algorithmic fidelity—say by setting it as an optimization target —then you would likely end up adding numerous harmful interaction patterns which others are trying to remove in other lines of research (e.g. [75, 77]). Algorithmic fidelity thus works best as an evaluation of an existing system, not as a metric to try to optimize directly. How to create a relatively "unaligned" (i.e. high algorithmic fidelity) LLM for in *silico* research on human behavior without compromising on measures taken to reduce harmful biases is an important open question.

Conclusion

LLM technology is currently advancing rapidly. However, the results described here did not indicate that the LLM we tested had sufficient algorithmic fidelity to support *in silico* research on human behavior or to support its other applications like computer games and training simulators. Nevertheless, it remains possible that similar models will have sufficient algorithmic fidelity in the future. Therefore it is important to get ahead of the applications, and focus in advance on resolving the field's critical conceptual challenges, like establishing ways of assessing algorithmic fidelity, figuring out its limits, and determining whether and how to improve it where it falls short.

Supporting information

S1 Table. Belief statements shared by human participants (left) and silicon participants (right). Belief statements were aligned based on their similarity when comparing human and silicon responses. **'Rank'** refers to how frequently a domain was mentioned by humans and silicon participants, assessed separately for these groups. Instances, where beliefs are repeated, suggest that human participants expressed more detailed perspectives. For example, while humans might provide three distinct statements about how "Physical activity exacerbates my symptoms"—citing (1) fatigue, (2) a tight chest, and (3) heavy limbs—a silicon participant might simply state that "Physical activity worsens my symptoms".
(PDF)

S1 File. Sensitivity analysis.
(PDF)

Author Contributions

Conceptualization: Aliya Amirova, Martin R. Cowie, Joel Z. Leibo.

Data curation: Aliya Amirova, Theodora Fteropoulli, Nafiso Ahmed.

Formal analysis: Aliya Amirova, Theodora Fteropoulli, Nafiso Ahmed.

Investigation: Aliya Amirova.

Methodology: Aliya Amirova, Theodora Fteropoulli, Nafiso Ahmed, Joel Z. Leibo.

Validation: Theodora Fteropoulli, Nafiso Ahmed.

Visualization: Aliya Amirova.

Writing – original draft: Aliya Amirova, Martin R. Cowie, Joel Z. Leibo.

Writing – review & editing: Aliya Amirova, Martin R. Cowie, Joel Z. Leibo.

References

1. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Advances in neural information processing systems. 2020; 33:1877–1901.
2. OpenAI. GPT-4 Technical Report; 2023.
3. Anil R, Dai AM, Firat O, Johnson M, Lepikhin D, Passos A, et al. Palm 2 technical report. arXiv preprint arXiv:230510403. 2023;.
4. Argyle LP, Busby EC, Fulda N, Gubler JR, Rytting C, Wingate D. Out of one, many: Using language models to simulate human samples. Political Analysis. 2023; 31(3):337–351. https://doi.org/10.1017/pan.2023.2
5. Brand J, Israeli A, Ngwe D. Using gpt for market research. Available at SSRN 4395751. 2023;.
6. Aher GV, Arriaga RI, Kalai AT. Using large language models to simulate multiple humans and replicate human subject studies. In: International Conference on Machine Learning. PMLR; 2023. p. 337–371.
7. Dillion D, Tandon N, Gu Y, Gray K. Can AI language models replace human participants? Trends in Cognitive Sciences. 2023;. https://doi.org/10.1016/j.tics.2023.04.008 PMID: 37173156
8. Grossmann I, Feinberg M, Parker DC, Christakis NA, Tetlock PE, Cunningham WA. AI and the transformation of social science research. Science. 2023; 380(6650):1108–1109. https://doi.org/10.1126/science.adi1778 PMID: 37319216
9. Horton JJ. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? arXiv preprint arXiv:230107543. 2023;.
10. Nowak KL, Fox J. Avatars and computer-mediated communication: a review of the definitions, uses, and effects of digital representations. Review of Communication Research. 2018; 6:30–53. https://doi.org/10.12840/issn.2255-4165.2018.06.01.015
11. Taylor L, Ranaldi H, Amirova A, Zhang L, Ahmed AA, Dibb B. Using virtual representations in mHealth application interventions for health-related behaviour change: A systematic review. Cogent psychology. 2022; 9(1):2069906. https://doi.org/10.1080/23311908.2022.2069906
12. Bearne LM, Sekhon M, Grainger R, La A, Shamali M, Amirova A, et al. Smartphone apps targeting physical activity in people with rheumatoid arthritis: systematic quality appraisal and content analysis. JMIR mHealth and uHealth. 2020; 8(7):e18495. https://doi.org/10.2196/18495 PMID: 32706727
13. Sekhon M, White C, Godfrey E, Amirova A, Revenäs Å, King S, et al. Effectiveness of web-based and mobile health interventions designed to enhance adherence to physical activity for people with inflammatory arthritis: a systematic review. Rheumatology advances in practice. 2021; 5(1):rkab016. https://doi.org/10.1093/rap/rkab016 PMID: 33928212
14. Greenwood SA, Young HM, Briggs J, Castle E, Walklin C, Haggis L, et al. Evaluating the effect of a digital health intervention to enhance physical activity in people with chronic kidney disease (Kidney BEAM): A multi-centre, randomised controlled trial. The Lancet Digital Health. 2023;.
15. Gold SM, Friede T, Meyer B, Moss-Morris R, Hudson J, Asseyer S, et al. Internet-delivered cognitive behavioural therapy programme to reduce depressive symptoms in patients with multiple sclerosis: a multicentre, randomised, controlled, phase 3 trial. The Lancet Digital Health. 2023; 5(10):e668–e678. https://doi.org/10.1016/S2589-7500(23)00109-7 PMID: 37775187

16. Zion SR, Taub CJ, Heathcote LC, Ramiller A, Tinianov S, McKinley M, et al. A cognitive behavioral digital therapeutic for anxiety and depression in patients with cancer: A decentralized randomized controlled trial.; 2023.

17. Clark DB, Tanner-Smith EE, Killingsworth SS. Digital games, design, and learning: A systematic review and meta-analysis. Review of educational research. 2016; 86(1):79–122. https://doi.org/10.3102/0034654315582065 PMID: 26937054

18. Zheng S, Trott A, Srinivasa S, Parkes DC, Socher R. The AI Economist: Taxation policy design via two-level deep multiagent reinforcement learning. Science advances. 2022; 8(18):eabk2607. https://doi.org/10.1126/sciadv.abk2607 PMID: 35507657

19. Koster R, Balaguer J, Tacchetti A, Weinstein A, Zhu T, Hauser O, et al. Human-centred mechanism design with Democratic AI. Nature Human Behaviour. 2022; 6(10):1398–1407. https://doi.org/10.1038/s41562-022-01383-x PMID: 35789321

20. Dafoe A, Hughes E, Bachrach Y, Collins T, McKee KR, Leibo JZ, et al. Open problems in cooperative AI. arXiv preprint arXiv:201208630. 2020;.

21. UK G. The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November. 2023;.

22. Reeves S, Kuper A, Hodges BD. Qualitative research methodologies: ethnography. Bmj. 2008; 337. https://doi.org/10.1136/bmj.a1020 PMID: 18687725

23. Giorgi A. The descriptive phenomenological method in psychology: A modified Husserlian approach. Duquesne university press; 2009.

24. Bryant A, Charmaz K. Grounded theory in historical perspective: An epistemological account. The SAGE handbook of grounded theory. 2007; p. 31–57.

25. Clarke V, Braun V, Hayfield N. Thematic analysis. Qualitative psychology: A practical guide to research methods. 2015; 3:222–248.

26. Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC medical research methodology. 2013; 13 (1):1–8. https://doi.org/10.1186/1471-2288-13-117 PMID: 24047204

27. Tromp J, Jindal D, Redfern J, Bhatt A, Séverin T, Banerjee A, et al. World heart federation roadmap for digital health in cardiology. Global Heart. 2022; 17(1). https://doi.org/10.5334/gh.1141 PMID: 36051317

28. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. bmj. 2021; 374. https://doi.org/10.1136/bmj.n2061 PMID: 34593508

29. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: the new Medical Research Council guidance. Bmj. 2008; 337. https://doi.org/10.1136/bmj.a1655 PMID: 18824488

30. Rothman AJ. "Is there nothing more practical than a good theory?": Why innovations and advances in health behavior change will arise if interventions are used to test and refine theory. International Journal of Behavioral Nutrition and Physical Activity. 2004; 1(1):1–7. https://doi.org/10.1186/1479-5868-1-11 PMID: 15279674

31. Bicchieri C. Norms in the wild: How to diagnose, measure, and change social norms. Oxford University Press; 2016.

32. Amirova A, Lucas R, Cowie MR, Haddad M. Perceived barriers and enablers influencing physical activity in heart failure: A qualitative one-to-one interview study. PloS one. 2022; 17(8):e0271743. https://doi.org/10.1371/journal.pone.0271743 PMID: 35925964

33. Taylor RS, Long L, Mordi IR, Madsen MT, Davies EJ, Dalal H, et al. Exercise-based rehabilitation for heart failure: Cochrane systematic review, meta-analysis, and trial sequential analysis. JACC: Heart Failure. 2019; 7(8):691–705. PMID: 31302050

34. Sagar VA, Davies EJ, Briscoe S, Coats AJ, Dalal HM, Lough F, et al. Exercise-based rehabilitation for heart failure: systematic review and meta-analysis. Open heart. 2015; 2(1):e000163. https://doi.org/10.1136/openhrt-2014-000163 PMID: 25685361

35. Lewinter C, Doherty P, Gale CP, Crouch S, Stirk L, Lewin RJ, et al. Exercise-based cardiac rehabilitation in patients with heart failure: a meta-analysis of randomised controlled trials between 1999 and 2013. European journal of preventive cardiology. 2015; 22(12):1504–1512. https://doi.org/10.1177/2047487314559853 PMID: 25398703

36. O'Donnell J, Smith-Byrne K, Velardo C, Conrad N, Salimi-Khorshidi G, Doherty A, et al. Self-reported and objectively measured physical activity in people with and without chronic heart failure: UK Biobank analysis. Open Heart. 2020; 7(1):e001099. https://doi.org/10.1136/openhrt-2019-001099 PMID: 32153787

37. Amirova A, Taylor L, Volkmer B, Ahmed N, Chater AM, Fteropoulli T. Informing behaviour change intervention design using systematic review with Bayesian meta-analysis: physical activity in heart failure. Health Psychology Review. 2022; p. 1–29. PMID: 35701235

38. Amirova A, Fteropoulli T, Williams P, Haddad M. Efficacy of interventions to increase physical activity for people with heart failure: a meta-analysis. Open Heart. 2021; 8(1):e001687. https://doi.org/10.1136/openhrt-2021-001687 PMID: 34108272

39. Cane J, O'Connor D, Michie S. Validation of the theoretical domains framework for use in behaviour change and implementation research. Implementation science. 2012; 7:1–17. https://doi.org/10.1186/1748-5908-7-37 PMID: 22530986

40. Gao L, Biderman S, Black S, Golding L, Hoppe T, Foster C, et al. The pile: An 800gb dataset of diverse text for language modeling. arXiv preprint arXiv:210100027. 2020;.

41. Hastie T, Tibshirani R, Friedman JH, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. vol. 2. Springer; 2009.

42. Christiano PF, Leike J, Brown T, Martic M, Legg S, Amodei D. Deep reinforcement learning from human preferences. Advances in neural information processing systems. 2017; 30.

43. Glaese A, McAleese N, Trebacz M, Aslanides J, Firoiu V, Ewalds T, et al. Improving alignment of dialogue agents via targeted human judgements. arXiv preprint arXiv:220914375. 2022;.

44. Bai Y, Jones A, Ndousse K, Askell A, Chen A, DasSarma N, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:220405862. 2022;.

45. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems. 2022; 35:27730–27744.

46. Zhou C, Liu P, Xu P, Iyer S, Sun J, Mao Y, et al. LIMA: Less is more for alignment. arXiv preprint arXiv:230511206. 2023;.

47. Reinecke MG, Mao Y, Kunesch M, Duéñez-Guzmán EA, Haas J, Leibo JZ. The Puzzle of Evaluating Moral Cognition in Artificial Agents. Cognitive Science. 2023; 47(8):e13315. https://doi.org/10.1111/cogs.13315 PMID: 37555649

48. Zou A, Wang Z, Kolter JZ, Fredrikson M. Universal and Transferable Adversarial Attacks on Aligned Language Models. arXiv preprint arXiv:230715043. 2023;.

49. Force T, McDonagh TA, Metra M, Adamo M, Gardner RS, Baumbach A, et al. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: Developed by the Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). With the special contribution of the Heart Failure Association (HFA) of the ESC. European Journal of Heart Failure. 2022; 24(1).

50. Francis JJ, Johnston M, Robertson C, Glidewell L, Entwistle V, Eccles MP, et al. What is an adequate sample size? Operationalising data saturation for theory-based interview studies. Psychology and health. 2010; 25(10):1229–1245. https://doi.org/10.1080/08870440903194015 PMID: 20204937

51. Park JS, O'Brien JC, Cai CJ, Morris MR, Liang P, Bernstein MS. Generative Agents: Interactive Simulacra of Human Behavior. arXiv preprint arXiv:230403442. 2023;.

52. Shanahan M, McDonell K, Reynolds L. Role play with large language models. Nature. 2023; p. 1–6. PMID: 37938776

53. Vezhnevets AS, Agapiou JP, Aharon A, Ziv R, Matyas J, Duéñez-Guzmán EA, et al. Generative agent-based modeling with actions grounded in physical, social, or digital space using Concordia. arXiv preprint arXiv:231203664. 2023;.

54. Atkins L, Francis J, Islam R, O'Connor D, Patey A, Ivers N, et al. A guide to using the Theoretical Domains Framework of behaviour change to investigate implementation problems. Implementation science. 2017; 12(1):1–18. https://doi.org/10.1186/s13012-017-0605-9 PMID: 28637486

55. McDonald S, Vieira R, Godfrey A, O'Brien N, White M, Sniehotta FF. Changes in physical activity during the retirement transition: a series of novel n-of-1 natural experiments. International Journal of Behavioral Nutrition and Physical Activity. 2017; 14:1–12. https://doi.org/10.1186/s12966-017-0623-7 PMID: 29221449

56. Taylor N, Lawton R, Conner M. Development and initial validation of the determinants of physical activity questionnaire. International Journal of Behavioral Nutrition and Physical Activity. 2013; 10(1):1–11. https://doi.org/10.1186/1479-5868-10-74 PMID: 23758912

57. Francis JJ, Stockton C, Eccles MP, Johnston M, Cuthbertson BH, Grimshaw JM, et al. Evidence-based selection of theories for designing behaviour change interventions: Using methods based on theoretical construct domains to understand clinicians' blood transfusion behaviour. British journal of health psychology. 2009; 14(4):625–646. https://doi.org/10.1348/135910708X397025 PMID: 19159506

58. Michie S, Johnston M, Abraham C, Lawton R, Parker D, Walker A. Making psychological theory useful for implementing evidence based practice: a consensus approach. BMJ quality & safety. 2005; 14 (1):26–33. https://doi.org/10.1136/qshc.2004.011155 PMID: 15692000

59. Swann C, Jackman PC, Lawrence A, Hawkins RM, Goddard SG, Williamson O, et al. The (over) use of SMART goals for physical activity promotion: A narrative review and critique. Health psychology review. 2023; 17(2):211–226. https://doi.org/10.1080/17437199.2021.2023608 PMID: 35094640

60. Bull FC, Al-Ansari SS, Biddle S, Borodulin K, Buman MP, Cardon G, et al. World Health Organization 2020 guidelines on physical activity and sedentary behaviour. British journal of sports medicine. 2020; 54(24):1451–1462. https://doi.org/10.1136/bjsports-2020-102955 PMID: 33239350

61. Camerer CF. Behavioral game theory: Experiments in strategic interaction. Princeton university press; 2011.

62. Sniehotta FF, Presseau J, Araújo-Soares V. Time to retire the theory of planned behaviour; 2014.

63. Phillips LA, Gardner B. Habitual exercise instigation (vs. execution) predicts healthy adults' exercise frequency. Health Psychology. 2016; 35(1):69. https://doi.org/10.1037/hea0000249 PMID: 26148187

64. Rothman AJ, Sheeran P, Wood W. Reflective and automatic processes in the initiation and maintenance of dietary change. Annals of Behavioral Medicine. 2009; 38(suppl_1):s4–s17. https://doi.org/10.1007/s12160-009-9118-3 PMID: 19787308

65. Ouellette JA, Wood W. Habit and intention in everyday life: The multiple processes by which past behavior predicts future behavior. Psychological bulletin. 1998; 124(1):54. https://doi.org/10.1037/0033-2909.124.1.54

66. Rhodes RE, de Bruijn GJ. How big is the physical activity intention–behaviour gap? A meta-analysis using the action control framework. British journal of health psychology. 2013; 18(2):296–309. https://doi.org/10.1111/bjhp.12032 PMID: 23480428

67. Perez E, Ringer S, Lukošiūtė K, Nguyen K, Chen E, Heiner S, et al. Discovering language model behaviors with model-written evaluations. arXiv preprint arXiv:221209251. 2022;.

68. ITU Hub. Measuring digital development: Facts and Figures: Focus on Least Developed Countries; 2023. Available from: https://www.itu.int/hub/publication/d-ind-ict_mdd-2023/.

69. Office for National Statistics. Exploring the UK's digital divide; 2019. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/householdcharacteristics/homeinternetandsocialmediausage/articles/exploringtheuksdigitaldivide/2019-03-04.

70. Henrich J, Heine SJ, Norenzayan A. The weirdest people in the world? Behavioral and brain sciences. 2010; 33(2-3):61–83. https://doi.org/10.1017/S0140525X0999152X PMID: 20550733

71. Atari M, Xue MJ, Park PS, Blasi D, Henrich J. Which humans? 2023;.

72. Luccioni AS, Akiki C, Mitchell M, Jernite Y. Stable bias: Analyzing societal representations in diffusion models. arXiv preprint arXiv:230311408. 2023;.

73. Glickman M, Sharot T. How human-AI feedback loops alter human perceptual, emotional and social judgements. 2024;.

74. Jang M, Lukasiewicz T. Consistency analysis of chatgpt. arXiv preprint arXiv:230306273. 2023;.

75. Weidinger L, Uesato J, Rauh M, Griffin C, Huang PS, Mellor J, et al. Taxonomy of risks posed by language models. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency; 2022. p. 214–229.

76. Rudinger R, Naradowsky J, Leonard B, Van Durme B. Gender bias in coreference resolution. arXiv preprint arXiv:180409301. 2018;.

77. Santurkar S, Durmus E, Ladhak F, Lee C, Liang P, Hashimoto T. Whose opinions do language models reflect? arXiv preprint arXiv:230317548. 2023;.

78. Cowie MR, Lam CS. Remote monitoring and digital health tools in CVD management. Nature Reviews Cardiology. 2021; 18(7):457–458. https://doi.org/10.1038/s41569-021-00548-x PMID: 33824486

79. Cowie M, McBeath K, Angermann C. The digital future of heart failure care. Current Heart Failure Reports. 2022; 19(3):109–113. https://doi.org/10.1007/s11897-022-00547-0 PMID: 35441250

80. National Institute for Health Research: UK Standards for Public Involvement. UK Standards for Public Involvement; 2023. Available from: https://sites.google.com/nihr.ac.uk/pi-standards/the-project.

81. Deverka PA, Lavallee DC, Desai PJ, Esmail LC, Ramsey SD, Veenstra DL, et al. Stakeholder participation in comparative effectiveness research: defining a framework for effective engagement. Journal of comparative effectiveness research. 2012; 1(2):181–194. https://doi.org/10.2217/cer.12.7 PMID: 22707880

82. National Institute for Health Research. National Institute for Health and Care Excellence (NICE). What is public involvement in research? – INVOLVE.; 2015. Available from: https://www.invo.org.uk/findout-more/what-is-public-involvement-in-research-2/.

83. Jennings H, Slade M, Bates P, Munday E, Toney R. Best practice framework for Patient and Public Involvement (PPI) in collaborative data analysis of qualitative mental health research: methodology development and refinement. BMC psychiatry. 2018; 18(1):1–11. https://doi.org/10.1186/s12888-018-1794-8 PMID: 29954373

84. Berditchevskaia A, Peach K, Malliaraki E. Participatory AI for humanitarian innovation. London: Nesta. 2021;.

85. Birhane A, Isaac W, Prabhakaran V, Diaz M, Elish MC, Gabriel I, et al. Power to the people? opportunities and challenges for participatory AI. Equity and Access in Algorithms, Mechanisms, and Optimization. 2022; p. 1–8. https://doi.org/10.1145/3551624.3555290

86. Gasteiger N, Ahn HS, Lee C, Lim J, Macdonald BA, Kim GH, et al. Participatory Design, Development, and Testing of Assistive Health Robots with Older Adults: An International Four-year Project. ACM Transactions on Human-Robot Interaction (THRI). 2022; 11(4):1–19. https://doi.org/10.1145/3533726

87. Leibo JZ, Dueñez-Guzman EA, Vezhnevets A, Agapiou JP, Sunehag P, Koster R, et al. Scalable evaluation of multi-agent reinforcement learning with melting pot. In: International conference on machine learning. PMLR; 2021. p. 6187–6199.

88. Pan A, Chan JS, Zou A, Li N, Basart S, Woodside T, et al. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. In: International Conference on Machine Learning. PMLR; 2023. p. 26837–26867.

89. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency; 2021. p. 610–623.

90. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM computing surveys (CSUR). 2021; 54(6):1–35. https://doi.org/10.1145/3457607

91. Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. Ebiomedicine. 2023; 90. https://doi.org/10.1016/j.ebiom.2023.104512 PMID: 36924620

92. Kalra N, Paddock SM. Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? Transportation Research Part A: Policy and Practice. 2016; 94:182–193.

93. Zhavoronkov A. Caution with AI-generated content in biomedicine. Nature medicine. 2023; 29(3):532. https://doi.org/10.1038/d41591-023-00014-w PMID: 36750659