Machine-Assisted Social Psychology Hypothesis Generation

***Forthcoming in* American Psychologist**

Sachin Banker

University of Utah

Promothesh Chatterjee

University of Utah

Himanshu Mishra

University of Utah

Arul Mishra

University of Utah

Author Note

Address correspondence to: sachin.banker@utah.edu. This research was generously supported by the David Eccles School of Business, University of Utah. Data and code are available upon request.

© 2023, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors' permission. The final article will be available, upon publication, via its DOI: 10.1037/amp0001222

Machine-Assisted Social Psychology Hypothesis Generation

Abstract

Social psychology research projects begin with generating a testable idea that relies heavily on a researcher's ability to assimilate, recall, and accurately process available research findings. However, an exponential increase in new research findings is making the task of synthesizing ideas across the multitude of topics challenging, which could result in important overlooked research connections. In this research, we leverage the fact that social psychology research is based on verbal models and employ large natural language models to generate hypotheses that can aid social psychology researchers in developing new research hypotheses. We adopted two methodological approaches. In the first approach, we fine-tuned the GPT-3 language model on thousands of abstracts published in more than 50 social psychology journals, in the past 55 years as well as on pre-print repositories (PsyArXiv). Social psychology experts rated model-generated and human-generated hypotheses similarly on the dimensions of clarity, originality and impact. In the second approach, without fine-tuning we generated hypotheses using GPT-4 and found that social psychology experts rated these generated hypotheses as higher in quality than human-generated hypotheses on dimensions of clarity, originality, impact, plausibility, and relevance.

Keywords: generative language models, deep learning, hypothesis formation, generative network

Public Significance Statement. This work illustrates how Large Language Models (LLMs), like GPT-3 and GPT-4, can be used as an aid to generate research hypotheses for social psychology. The LLM-generated hypotheses were found to be on par with, or even better than, those written by human researchers. As research findings proliferate, these LLMs can help streamline the process of creating testable ideas and offer new avenues to accelerate psychological research.

The first step in any research project, inductive or deductive, is idea generation. A novel research idea can be developed from existing theory, may result from a flash of insight from a witnessed event, could stem from observing anomalous patterns in data, or arise from cross-connection involving inter-disciplinary findings (Jaccard & Jacoby, 2019). Idea generation occupies an important position in social psychology (Koehler, 1994; Kruglanski, 1990; McGuire, 1973) as it sets the direction for examining the different factors that affect human perceptions, attitudes, and behaviors—be they within the person or in the environment. Developing a testable hypothesis from the generated idea is the usual next step for most empirical research. The hypothesis makes the research process testable and falsifiable and the testing protocols valid, reliable, and reproducible; it also links the idea concretely to specific theories or applications.

The generation of new ideas relies heavily on existing research. Hence, the rapidly expanding volume of research findings - both published and in pre-print - is making it increasingly difficult for researchers to keep pace with, and assimilate all relevant, existing research into their idea generation process. Given that global scientific output is doubling approximately every nine years (Bornmann & Mutz, 2015; Cheadle et al., 2017), it's not surprising that synthesizing the most current understanding from the extant body of research has become increasingly challenging. The field of social psychology mirrors this trend, with the number of published articles increasing by roughly 500% over the past two decades according to the Web of Science (Li et al., 2018). The growth of pre-print repositories like PsyArXiv, which alone receives over 7000 articles annually (Condon et al., 2020), also contributes to the overwhelming amount of available research.

This growth in research findings though very welcome, since it generates new areas to study for researchers, also poses some challenges. First, there are cognitive limitations in researchers' ability to synthesize an ever-expanding literature (Bornmann & Mutz, 2015; Cheadle et al., 2017; Cowley et al., 2023). Second, the immense research output prevents researchers from seeing rich inter-connections that they may otherwise notice easily (Sybrandt et al., 2018). Third, human perceptions and behaviors are the outcome of interactive processes that rely on many factors, relevant as well as irrelevant (depending on the weight assigned to it by a person's perception). Given this, it is important that social psychology ideas and hypotheses are informed by as many of these factors documented in literature as possible. The difficulties in assimilating such vast amounts of information could inadvertently lead researchers to overlook certain aspects of multi-faceted human behavior and its interaction with their environment.

Hypothesis Generation Model

The unit of observation in social psychology is a human who is less consistent in their behavior and who interacts with their surroundings and other people in quite a varied manner. Hence, most research in social psychology relies on verbal models. Such verbal models are not amenable to the largely mathematical techniques used in other scientific domains to tackle the challenge of generating hypotheses from an ever-expanding literature (Evans & Rzhetsky, 2010; Krenn & Zeilinger, 2020; Wilson et al., 2018).

In the current work, we leverage the fact that social psychology research is based on verbal models, and use the recent advances in Natural Language Processing (NLP) in which language models not only infer meaning from text (Devlin et al., 2018; Mikolov et al., 2013) but now also have the ability to produce original text (Guo et al., 2018; Martin et al., 2016; Zhang et al., 2017; Zhu et al., 2018). The main goal of this research is to harness the power of generative language models to aid researchers in generating hypotheses in social psychology. In our work, we use two methodological approaches. One in which the GPT-3 large language model (LLM) is fine-tuned specifically on several thousands of abstracts gathered from 50 social psychology journals over more than 55 years as well as pre-prints such as PsyArXiv. Second, we use the GPT-4 large language model to generate hypotheses based on specific prompts. In order to check the quality of the hypotheses, we surveyed social psychology experts by presenting the hypotheses to them and asking them to rate both human and model-generated hypothesis on dimensions such as originality, clarity, and importance. Henceforth, we refer to the hypotheses generated by our generative language model as *model-generated hypotheses*. We would like to emphasize that the hypotheses generated from the process outlined in this research will not replace human creativity and ingenuity in developing new social psychology hypothesis. Instead, we anticipate that these models will serve as a valuable aid to researchers in synthesizing research findings but researchers must still iteratively curate and revise model-generated hypotheses when identifying promising new directions for inquiry.

Generative Language Model

Our approach uses generative language models that help in multiple text-related tasks ranging from classification of text into groups based on their meaning to generating new text (Keselj, 2009; Peters et al., 2018). They have been successful in creating human-like output in areas such as writing news articles, short stories, press releases, and lyrics (Radford et al., 2019). Specific to having the ability to make human-like judgements, recent research has demonstrated generative models' abilities in tasks of information search, causal reasoning, deliberation and decision-making (for instance, the model exhibited a conjunction fallacy in the Linda problem indicating usage of the same heuristics as a human) (Binz & Schulz, 2022).

Generative language models, can be used to generate text either through a process of fine-tuning on corpora specific to the task or through a process of providing specific prompts, without fine-tuning, that help them generate relevant text. We used both processes across two studies to highlight the diverse ways in which generative models can be used for psychological research purposes. We first describe the generative model that uses a fine-tuning process.

Generative Model with Fine-tuning

The fine-tuning process generally, follows a two-stage process. In the first stage, the model is trained to learn from large text corpora, and in the second stage the model is further trained and fine-tuned on topic-specific corpora. The first stage of training helps in learning language representations in an unsupervised manner, which does not require expensive and scarce human annotations (Mikolov et al., 2013; Pennington et al., 2014). The advantage of using an unsupervised learning in the first stage is that the corpus, from which the model learns, can be quite broad. Such a broad corpus helps the model learn meaningful relationships among words and the context in which they are used by humans—that is, simple and proper usage of words as used in human language. This helps significantly when the model is asked to generate meaningful and coherent text. The trained model obtained at the end of the first stage is generally referred to as the pre-trained model. However, if we stop here and ask the model to generate text on a specific topic, it may not do very well because its learning is general, not specific. Therefore, the pre-trained model's learning is transferred and leveraged into a specific domain (e.g., social psychology) by making the process semi-supervised to generate text in that specific domain.

In the second stage of learning, the model is provided with topic-specific text that is used to fine-tune its learning on that topic area (Radford et al., 2018, 2019). The second stage leverages and enhances the pre-trained language representations with specific topic terms to provide more accurate and matched representations of that topic (Devlin et al., 2018; Mikolov et al., 2013; Peters et al., 2018). However, if we skipped the first stage of training and trained our model only on domain-specific text corpus (social psychology research in our case) the model would not produce good-quality text since it would not have learned simple linguistic associations that are possible to learn only from large, generalized corpora (Radford et al., 2019).

Hence, in order to generate meaningful and new hypotheses, we used GPT-3 which had been trained across a wide variety of corpora in an unsupervised manner. Subsequently, we fine-tuned the learning the GPT-3 model by training it using social psychology research over 55 years.

Method

First Stage: Pre-trained Language Model

The generative language models that we use in the first stage is referred to as GPT-3 (Brown et al., 2020) where GPT stands for Generative Pre-trained Transformer. GPT-3 is a third-generation, autoregressive language model that uses deep learning to produce human-like text. GPT-3 was developed by OpenAI and has been trained on several large text corpora such as Common Crawl, Wikipedia, digitized books, WebText2 (which is based on Reddit posts), etc. The total volume of training data amounts to approximately 499 billion tokens. (Brown et al., 2020) – where tokens are pieces of words (e.g., the US Declaration of Independence has 1337 words but it has 1695 tokens).

The text dataset consisted of Wikipedia (English language text, 3 billion tokens), WebText (text of more than 45 million web pages linked to reddit posts with at least 2 upvotes, 19 billion tokens), Common Crawl (open source archived dataset from 25 billion webpages, 410 billion tokens), and digitized books (a collection of free books written by unpublished authors, scientific papers, fiction, and non-fiction published books, 67 billion tokens) (Brown et al., 2020; Thompson, 2021).

The pre-trained generative model, uses a transformer (Vaswani et al., 2017), which improves on the sequential learning process commonly used in many language models such as the Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs). Transformers use an attention mechanism to process text in parallel and learn relationships among the words. This form of processing makes the learning more efficient and accurate because it allows the model to learn long-term dependencies in text. That is, instead of learning to relate a target text to just a few words in front of and behind a target text, the model can learn relations spread out through longer text sequences (Rocktäschel et al., 2015; Vaswani et al., 2017). Such a capability is important for transferring the learning from a large corpus of pre-trained vector representations to a specific domain (Radford et al., 2018). Therefore, it is ideal for generating novel text when given a certain prompt.

By learning from billions of tokens from a wide variety of text corpora, GPT-3 can generate text, given a context, that is as human-like as possible. The generative model is designed to predict the next word, given all of the previous words used in the corpus. Unlike the earlier version, GPT-2, since GPT-3 learns from more text and has billions more trainable parameters, it is better able to capture the complexities and nuances in human language making it better at generating text that is as human like as possible. That is, when the pre-trained GPT-3 model is provided with examples specific to a domain (e.g., in our case various social psychology abstracts), it can leverage its learning from large corpora to generate text comparable to human-generated text.

GPT-3 includes four different models that can be used for generating text – Ada, Babbage, Curie, and Davinci – with each model using more parameters. Ada (with 350 million parameters) is the fastest and most cost effective but for more complicated and nuanced generative tasks it may be less accurate. For more nuanced tasks such as semantic search tasks, the Babbage model (with 1.3 billion parameters) performs better. Curie (with 6.7 billion parameters) uses more parameters and is better than Ada and Babbage for complicated tasks such as sentiment classification and question-answers (Zhou et al., 2022). Finally, Davinci can handle the most generation tasks such as determining cause and effect, producing creative content, explaining character motives and complex summarization. However, given its 175 billion parameters it was the most expensive and slowest of GPT-3 models.

Second Stage: Fine-tuning the Generative Model

In the second stage, the generative model leverages the learning from the first stage and is trained further on over 100,000 social psychology abstracts gathered from over 50 journals such as the *Journal of Personality and Social Psychology, Journal of Experimental Psychology, American Psychologist, Journal of Experimental Social Psychology, European Journal of Social Psychology, Motivation and Emotion* and many others over 55 years. Abstracts published in the journals dating back to 1965, or whenever a journal started publishing, until the present were included.

Publishing null results is challenging (Bartko, 1982; Greenwald, 1975) but if they are ignored, it results in publication bias, which in turn limits the replicability assumption of science and impedes the process of falsification of hypotheses (Ferguson & Heene, 2012; Francis, 2012). Therefore, along with including abstracts from published research, we also included abstracts from pre-prints such as PsyArXiv that are more likely to include null findings, which do not get published. Inclusion of pre-print abstracts to train our generative model helps lessen the impact of publication bias. Moreover, social psychology has seen a series of papers being retracted due to various concerns. We guard against the role of such retracted work in influencing our model's learning process by using retraction watch sites to remove such work. This way our model is not exposed to retracted findings.

During the training process, the model learned from all the abstracts, which tend to include the hypotheses of the research. As the generative model goes through all the abstracts, it learns what types of theoretical or practical constructs are more likely to be associated; for example, it might learn that "stereotype" is associated with "prejudice" or that the words "motivation" and "goals" are associated. When the model is trained on existing research it learns what hypotheses already exist in literature. This helps in two ways: first it helps the model avoid repeating an existing hypothesis and second, it helps the model learn important connections to produce novel hypotheses that are specific to social psychology. Further details on the fine-tuning procedure are presented in the Supplemental Online Material.

Hypotheses Generation and a Pre-test

We used the fine-tuned Curie model to generate hypotheses by using the prefix "hypothesize that..." – that is, the GPT-3 model on its own generated completions to this sentence prefix that specified potential social psychology research hypotheses worthy of further exploration. It is important to note that an adjustable parameter named "temperature" controls diversity of generated text in GPT-3. Low temperature leads to very predictable next word in a sequence with low variation. Higher temperature leads to diverse set of words that increase novelty but also increases chances of absurd words appearing in generated text. Given the recommendation to use temperature of 0.9 for creative applications (OpenAI, 2022), we evaluated three values of temperature: 0.8, 0.9, or 1.0 and generated 100 hypotheses at each of these temperatures.

However, there is a possibility that the generated hypotheses are not novel but reproduction of hypotheses that the model saw during training or fine-tuning. Therefore, we performed a pre-test using Turnitin software to find out whether the model-generated hypotheses were reproductions of existing hypotheses or not. Turnitin is a software that is commonly used to check for plagiarism. It leverages a vast text corpus available on the web to find similarities between submitted and already available text. Their database contains text from 99 billion web pages and from 89 million published articles across 56000 journals and 13000 open access repositories (Turnitin, 2022). We provided Turnitin with all 600 hypotheses (300 model-generated hypotheses and 300 human-generated hypotheses from previously published abstracts) and asked it to give us its plagiarism score. We predicted that since journal abstracts (and hence the hypotheses) which have been published or appeared in pre-prints tend to be part of the corpus that Turnitin uses to test for plagiarism, the score should be high for human-generated hypotheses. If our model was simply reproducing prior human hypotheses then its plagiarism score should also be high. However, if the model is generating new hypotheses then its plagiarism score should be low. The results of Turnitin indicate that the plagiarism score for human-generated hypotheses was 94% while that for the model-generated hypotheses was only 1%. Hence, we have initial support that the model-generated hypotheses are not simply copies of prior human-generated hypotheses. Examples of the model-generated hypotheses are provided in the Supplemental Online Material.

To test how model generated hypotheses compared to human-generated hypotheses, we examined three dimensions (clarity, impact, and originality) that have been used in past research to determine the quality of hypotheses (Yuan et al., 2021). We then conducted a study with social psychology experts.

Hypotheses Evaluation

We next needed to evaluate the quality of the generated hypotheses. In order to do so we approached social psychology experts to read and evaluate the model-generated hypotheses and rate them on dimensions of clarity, originality, and novelty (Yuan et al., 2021). Importantly, they needed to be shown both human-generated and model-generated hypotheses at the same time and not be told which was which. Such a within-participants design acted as a conservative test helping us find out how model-generated hypotheses performed in comparison to human-generated ones.

We aimed to recruit an inclusive sample of expert participants spanning the breadth of the social psychology research community which included PhD students, post-docs, and faculty. We recruited 50 participants from the SPSP listserv to rate the hypotheses. Three participants who did not complete the survey were dropped from the analysis. The final sample of 47 participants consisted of 19 faculty members, 2 post-docs, and 26 PhD students; they were each paid $25 for rating the hypotheses. Respondents had an average of 12 years' experience in the field of social psychology (median = 7, SD = 11). Each participant rated a total of 30 (15 human and 15 model) hypotheses on the three dimensions of clarity, impact, and originality without their being informed which hypotheses were human-generated and which were model-generated (preregistration available here: https://aspredicted.org/8J2_4RW).

For each participant, the 15 human-generated hypotheses were selected at random from a pool of 300 hypotheses, where the pool consisted of human-generated hypotheses that were scraped from previously published abstracts within the Scopus database and PsyArXiv and beginning with the phrase "hypothesize that." These hypotheses were sourced from more than 50 social psychology journals described previously in the fine-tuning section, where more than 90% of the hypotheses were from peer-reviewed publications. Similarly, the 15 model-generated hypotheses were also selected at random from a pool of 300 hypotheses generated at three different temperature levels (with temperature T set to either 0.8, 0.9, or 1.0). Each participant was randomly assigned to evaluate model-generated hypotheses at any one temperature level (i.e., a participant did not see model-generated hypothesis from two different temperatures). A power analysis using G*Power software suggested that this study design would provide over 95% power to detect an effect of at least $f = .14$.

All participants were provided with details about each of the dimensions they were to rate the hypotheses on and examples before beginning the task. That is, prior to evaluating the 30 target hypotheses, participants were provided with example hypotheses and definitions regarding the three dimensions of clarity, impact, and originality in order to ensure task comprehension. Please see the Supplemental Online Material for materials used in the study; all data and code are available upon request. We asked participants to rate all hypotheses on the dimensions of clarity, impact, and originality using a 5-point scale for each of the dimensions (anchored from "very low" to "very high"). Participants were not informed as to whether the hypothesis they were rating was model-generated or human-generated. In this study, we report all measures, manipulations and exclusions. Study protocols were approved by the University's Institutional Review Board.

Results

Overall Analysis. The overall analysis evaluated whether participants rated the model-generated versus human-generated hypotheses differently on the three dimensions of clarity, impact, and originality combining across the three temperature levels.

In a repeated measures analysis (i.e., applying participant-level random errors), we found that experts judged model-generated hypotheses to be similar to human-generated hypotheses on most dimensions. Specifically, ratings did not differ on the dimension of clarity (b = -0.032, t(1362) = .97, p = .332) and on the dimension of impact (b = 0.038, t(1362) = 1.58, p = .115), evaluating human-generated hypotheses nominally lower on clarity and higher on impact. However, ratings of human-generated hypotheses did score higher on ratings of originality overall (b = 0.051, t(1362) = 2.24, p = .025); further analysis indicated that this difference occurred only in comparisons to model-generated hypotheses produced at the lowest temperature level. When temperature was set to 0.8, human-generated hypotheses scored higher on ratings of originality (b = .092, t(492) = 2.39, p = 0.017), however this difference was not significant when temperature was set to GPT-3 suggested value of 0.9 (b = .025, t(492) = 0.65, p = .515) and at temperature set to the value of 1.0 (b = .031, t(376) = .77, p = .444). These findings show that model-generated hypotheses were perceived similar to human-generated hypotheses on the dimensions of clarity, impact, and originality, particularly with hyperparameters set to high temperature levels.

Equivalence Analysis. We also conducted equivalence tests applying the two one-sided test (TOST) method using the TOSTER package in R (Lakens, 2017) to compare human-generated to model-generated hypotheses on the three dimensions. To do so, we set the thresholds conservatively to 0.2 and -0.2 (i.e., a small effect size) such that the equivalence test would evaluate whether model-generated and human-generated hypotheses were judged to be statistically equivalent (even within the small range of $d < 0.2$). First, on the dimension of clarity we found significant evidence for equivalence between model-generated and human-generated hypotheses, $t(1407) = 1.99$, $p = .023$, $d = -.064$, 90% CI [-.177, .049]. The equivalence test analysis indicated that there was strong evidence that the difference in clarity between model-generated and human-generated hypotheses was small (i.e., within a $d < .2$ difference); even when conservatively examining relatively wide 90% confidence intervals around the effect size estimate, we observe that the confidence intervals are within the [-0.2, 0.2] range. Similarly on the dimensions of impact and originality, evidence for equivalence between model-generated and human-generated hypotheses emerged $t(1408) = 2.37$, $p = .009$, $d = -.075$, 90% CI [-.011, .162], and $t(1408) = 2.02$, $p = .022$, $d = .010$, 90% CI [.022, .181] respectively. The results indicated that differences in model-generated and human-generated hypotheses were significantly narrower than a small effect.

More granular analyses were conducted at different temperature levels. Model-generated hypotheses displayed the greatest equivalence with human-generated hypotheses at higher temperature levels, where we found significant evidence for equivalence on the dimensions of impact (at T = 1.0) and originality (at T = 1.0 and T = 0.9). Please see Table 1. However, at the lowest temperature level (T = 0.8), human-generated hypotheses were rated to be significantly more original than model-generated hypotheses, $t(508) = 2.228$, $p = .026$, $d = .064$.

In sum, our findings indicate overall that expert social psychologists evaluate model-generated hypotheses to be equivalent to human-generated hypotheses on the dimensions of clarity, impact, and originality. That is, social psychology hypotheses generated by the fine-tuned GPT-3 language model were indistinguishable in quality versus those published by human social psychologists, as judged by expert social psychologists themselves. Our findings also suggest that setting the temperature hyperparameter to higher levels improves model performance, particularly on the dimensions of impact and originality.

Table 1. Differences in expert evaluations in model-generated vs. human-generated hypotheses. Estimated 90% confidence intervals on the effect size are reported overall and at each model temperature level. Asterisks mark significance levels for equivalence tests (indicating effect size is smaller than $d = .2$). Means for model-generated (M_m) and human-generated (M_h) hypotheses are also shown, with standard deviations presented in parentheses.

	Overall	**T = 0.8**	**T = 0.9**	**T = 1.0**
Clarity	$M_m = 2.97$ (1.27) $M_h = 2.90$ (1.30) $d = [-.177, .049]$* $t(1407) = 1.99$ $p = .023$	$M_m = 3.10$ (1.28) $M_h = 2.90$ (1.37) $d = [-.397, -.010]$ $t(506) = .033$ $p = .513$	$M_m = 2.98$ (1.27) $M_h = 2.94$ (1.26) $d = [-.221, .150]$ $t(508) = 1.47$ $p = .072$	$M_m = 2.78$ (1.23) $M_h = 2.86$ (1.27) $d = [-.126, .290]$ $t(388) = .934$ $p = .175$
Impact	$M_m = 3.02$ (.982) $M_h = 3.10$ (.993) $d = [-.011, .162]$** $t(1408) = 2.37$ $p = .009$	$M_m = 3.05$ (.954) $M_h = 3.16$ (.969) $d = [-.034, .246]$ $t(508) = 1.11$ $p = .135$	$M_m = 2.91$ (1.06) $M_h = 3.01$ (1.08) $d = [-.058, .254]$ $t(508) = 1.08$ $p = .140$	$M_m = 3.12$ (.909) $M_h = 3.12$ (.905) $d = [-.146, .157]$* $t(388) = 2.12$ $p = .017$
Originality	$M_m = 2.72$ (.916) $M_h = 2.82$ (.901) $d = [.022, .182]$* $t(1408) = 2.02$ $p = .022$	$M_m = 2.73$ (.935) $M_h = 2.92$ (.933) $d = [.048, .321]$ $t(508) = .190$ $p = .425$	$M_m = 2.74$ (.967) $M_h = 2.79$ (.953) $d = [-.089, .191]$* $t(508) = 1.75$ $p = .040$	$M_m = 2.67$ (.822) $M_h = 2.73$ (.774) $d = [-.072, .195]$* $t(387) = 1.71$ $p = .044$

* $p < .05$, ** $p < .01$

Generative Model without Fine-tuning

To evaluate the quality of hypotheses generated by the recently released GPT-4 model, we conducted a second study comparing GPT-4 model-generated hypotheses to human-generated hypotheses. The GPT-4 model is different than the GPT-3 Curie model in the following ways. First, GPT-4 has been designed to work effectively with user-provided prompts directly, removing the need for fine-tuning that was often necessary with GPT-3 for particular tasks. Second, GPT-4 features significant enhancement in contextual understanding and is able to provide responses that are nuanced and complex based on the prompts given. Lastly, GPT-4 has been trained on a larger dataset, building on a more expansive knowledge base.

We used a prompt to generate as high-quality hypotheses as possible. Specifically, we used the following prompt:

> You are an expert social psychologist. Your research interests are in Social Cognition, Attitudes and Attitude Change, Violence and Aggression, Prosocial Behavior, Prejudice and Discrimination, Self and Social Identity, Group Behavior, Social Influence, and Interpersonal Relationships. Your task is to generate counterintuitive yet plausible hypotheses. They should combine different sub fields of social psychology and advance theoretical knowledge. They should not be incremental. Make sure that your hypotheses are precisely stated and incorporate a comparison group. Begin each hypothesis with "Hypothesize that" and generate 100 hypotheses.

Mimicking the design of the previous study, we generated 100 hypotheses at three different levels of the temperature parameter (temp = 0.8, 0.9, and 1.0). Both human-generated and model-generated hypotheses were evaluated simultaneously by experts who were blind to the source of each hypothesis. As in the previous study, the human-generated hypotheses were selected at random from a pool of 300 hypotheses, where the pool consisted of human-generated hypotheses that were scraped from previously published abstracts within the Scopus database and PsyArXiv and beginning with the phrase “hypothesize that.” Again, these hypotheses were sourced from more than 50 social psychology journals described previously in the fine-tuning section, in which more than 90% of the hypotheses were from peer-reviewed publications. The model-generated hypotheses were also selected at random from a pool of 300 hypotheses generated at three different temperature levels (with temperature T set to either 0.8, 0.9, or 1.0). We intentionally refrained from adding any human-supervised input to filter the hypotheses, whether generated by humans or the model. Both sets of hypotheses were selected through an automated process to ensure an unbiased and representative pool of hypotheses for empirical examination.

Adding to the previous study, we asked participants to rate each of the hypothesis on 5 dimensions: clarity, originality, impact (identical to the previous study) and also on plausibility (whether the hypothesis appeared plausible) and relevance (theoretically or practically to the field of social psychology)[1] (Ludwig & Mullainathan, 2023; Yuan et al., 2021). We aimed to recruit 50 social psychology experts from the SPSP listserv to rate the hypotheses on the 5 dimensions (preregistration available here: https://aspredicted.org/HTD_B35). A total of 56 participants completed the survey and each received $25 compensation. They included 22 faculty members, 7 post-docs, 25 current PhD students, and 2 incoming PhD students. Respondents had an average of 10 years' experience in the field of social psychology (median = 7, SD = 8.7).

Participants were provided with example hypotheses and definitions regarding the five dimensions of clarity, impact, originality, plausibility, and relevance in order to ensure task comprehension. Each participant rated a total of 30 (15 human and 15 model) hypotheses using a 5-point scale (anchored from "very low" to "very high") for each of the five dimensions.

In addition, we probed whether the respondents felt they were qualified to evaluate the hypotheses at the end of the survey ("Overall, I felt that I had sufficient social psychology subject expertise to evaluate the hypotheses presented to me in this survey," 1 = strongly disagree, 7 = strongly agree). Respondents indicated agreement with this statement, as supported by a test against the scale midpoint ($M = 5.79$, $SD = 1.16$, nonparametric Wilcoxon signed rank test $W = 754$, $p < 0.001$). Please see the Supplemental Online Material for materials used in the study. Study protocols were approved by the [university name] Institutional Review Board.

[1] We thank a reviewer for suggesting these additional dimensions.

Results

In a regression analysis applying participant-level random errors, we found that experts judged the GPT-4 model-generated hypotheses to be significantly higher in quality than human-generated hypotheses on all five dimensions. Please see Table 2. In more granular analyses examining subsets of GPT-4 model-generated hypotheses separately at each temperature level, we similarly found that experts judged the GPT-4 model-generated hypotheses to be significantly higher in quality than human-generated hypotheses on most dimensions. However, ratings of originality did not reach significance at T = 0.8 and T = 1.0, and ratings of plausibility did not reach significance at T = 0.9.

Table 2. Differences in expert evaluations in GPT-4 model-generated vs. human-generated hypotheses. Positive regression parameters indicate GPT-4 model-generated hypotheses were evaluated higher on the corresponding dimension.

Hypotheses	**Dimension**	***b***	***se***	***t***	***df***	***p***	**95% CI low**	**95% CI high**
overall	clarity	1.27	0.050	25.3	1623	<.001	1.17	1.37
overall	originality	0.19	0.042	4.5	1623	<.001	0.11	0.28
overall	impact	0.48	0.039	12.2	1623	<.001	0.40	0.55
overall	plausibility	0.31	0.045	6.8	1623	<.001	0.22	0.39
overall	relevance	0.72	0.040	17.8	1623	<.001	0.64	0.80
T = 0.8	clarity	1.42	0.086	16.6	550	<.001	1.26	1.59
T = 0.8	originality	0.08	0.076	1.1	550	0.266	-0.06	0.23
T = 0.8	impact	0.35	0.072	4.9	550	<.001	0.21	0.49
T = 0.8	plausibility	0.58	0.074	7.8	550	<.001	0.43	0.72
T = 0.8	relevance	0.67	0.071	9.5	550	<.001	0.53	0.81
T = 0.9	clarity	1.21	0.082	14.9	550	<.001	1.05	1.37
T = 0.9	originality	0.37	0.067	5.6	550	<.001	0.24	0.50
T = 0.9	impact	0.50	0.060	8.4	550	<.001	0.38	0.62
T = 0.9	plausibility	0.04	0.082	0.5	550	0.638	-0.12	0.20
T = 0.9	relevance	0.82	0.066	12.5	550	<.001	0.69	0.95
T = 1.0	clarity	1.17	0.093	12.6	521	<.001	0.99	1.36
T = 1.0	originality	0.12	0.077	1.5	521	0.127	-0.03	0.27
T = 1.0	impact	0.58	0.069	8.3	521	<.001	0.44	0.71
T = 1.0	plausibility	0.30	0.077	4.0	521	<.001	0.15	0.45

General Discussion

As the research volume in social psychology continues to grow rapidly over time, human researchers face increasing limitations in their ability to absorb findings from the scientific literature when generating new hypotheses. This problem requires rethinking how researchers process existing findings when generating new research hypotheses. Human researchers may consequently hyperspecialize and miss relevant connections to other subfields, and may use heuristics that lead to overweighting newsworthy findings and underweighting those from different countries and cultures. We next discuss potential contributions as well as limitations of using large language models as they become more and more ubiquitous.

In this work, we have used LLMs to generate social psychology hypotheses. We believe that LLMs can be leveraged in many other ways to assist the research process. It's important to note that an empirical evaluation comparing the performance of Large Language Models (LLMs) and human experts is essential before these alternative applications of LLMs can be recommended for use in research practices. First, in addition to generating hypotheses that are novel and relevant to the field of social psychology, LLMs can now also be used to provide a theoretical justification as well as practical implications of the generated hypotheses. For instance, LLMs can be prompted to provide “contextualized hypotheses” where along with the hypothesis they can provide information about the proposed relationships, relevant literature, theoretical frameworks, and potential mechanisms. This will offer a more comprehensive understanding of the generated hypotheses and their place within the broader scientific context, in our case the social psychology context. Second, LLMs can be queried to not just generate hypothesis, but when given a specific hypothesis, they can be asked what would be the experimental design or analysis method that would be appropriate. These aspects of language models can enable psychology researchers to accelerate research productivity by generating empirical tests which address new research hypotheses. Third, LLMs can potentially help in sifting through the massive amounts of published literature by providing summaries, identifying key trends, and pinpointing relevant research, thus aiding in effective literature review. Finally, as some recent research suggests, LLMs can be considered as a participant in a social psychology study (Hagendorff, 2023) since it can be said to consider the opinions and thoughts of multitude of people. Hence, LLMs have the potential to be used a vital tool by social psychology researchers.

However, along with the ways in which LLMs can be used it is also important to consider the limitations of LLMs when using them for research. Just like other language models have been demonstrated to hold several types of historical, societal biases, LLMs also generate responses based on text they have been trained on. Hence, the generated output of LLMs need to be checked for bias and debiased if possible. In our case, the bias that we needed to be cognizant of was that the LLM was trained on existing research some of which has been shown to have some flaws such as lack of replicability or the likelihood of null results being ignored. Another limitation that has been discussed in using LLMs is the fact that it can result in less diverse thinking. Since, the newer LLM models produce a summary output, its output can be less diverse because it may produce the strongest and dominant opinion as the only opinion (Park et al., 2023). Hence, researchers need to be cognizant of this fact while using the LLMs, and new strategies around prompt engineering could help to minimize this concern. Finally, it is important to underline that LLMs are fundamentally pattern-recognition tools trained on extensive textual data, which allows them to learn various patterns, interconnections, and to generate new insights based on the information they have absorbed. However, the boundaries of their creative capacity are shaped by the contours of the pre-existing knowledge they have been trained on. They are currently not capable of generating truly novel insights that often arise from deep, creative thought process that fundamentally challenge existing models and assumptions. So, while LLMs can certainly aid in generating and exploring hypotheses, their function should be perceived as an augmentation to human cognitive abilities, rather than as a replacement. Their strength lies in identifying patterns and insights from extensive literature, which can be instrumental in supporting the uniquely human task of generating truly innovative insights.

References

Bartko, J. J. (1982). The fate of published articles, submitted again. *Behavioral and Brain Sciences*, *5*(2), 199–199.

Binz, M., & Schulz, E. (2022). Using cognitive psychology to understand GPT-3. *ArXiv Preprint ArXiv:2206.14576*.

Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, *66*(11), 2215–2222.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Cheadle, C., Cao, H., Kalinin, A., & Hodgkinson, J. (2017). Advanced literature analysis in a Big Data world. *Annals of the New York Academy of Sciences*, *1387*(1), 25–33.

Condon, D. M., Arnal, J., Binion, G., Brown, B., & Corker, K. S. (2020). Not One but Many Models of Open-Access Publishing. *APS Observer*, *33*(9).

Cowley, H. P., Robinette, M. S., Matelsky, J. K., Xenes, D., Kashyap, A., Ibrahim, N. F., Robinson, M. L., Zeger, S., Garibaldi, B. T., & Gray-Roncal, W. (2023). Using machine learning on clinical data to identify unexpected patterns in groups of COVID-19 patients. *Scientific Reports*, *13*(1), 2236.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv:1810.04805*.

Evans, J., & Rzhetsky, A. (2010). Machine science. *Science*, *329*(5990), 399–400.

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, *7*(6), 555–561.

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review*, *19*(6), 975–991.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*(1), 1.

Guo, J., Lu, S., Cai, H., Zhang, W., Yu, Y., & Wang, J. (2018). Long text generation via adversarial training with leaked information. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1).

Hagendorff, T. (2023). *Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models Using Psychological Methods* (arXiv:2303.13988). arXiv. https://doi.org/10.48550/arXiv.2303.13988

Jaccard, J., & Jacoby, J. (2019). *Theory construction and model-building skills: A practical guide for social scientists*. Guilford publications.

Keselj, V. (2009). *Speech and Language Processing Daniel Jurafsky and James H. Martin (Stanford University and University of Colorado at Boulder) Pearson Prentice Hall, 2009, xxxi+ 988 pp; hardbound, ISBN 978-0-13-187321-6, $115.00*. MIT Press.

Koehler, D. J. (1994). Hypothesis generation and confidence in judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(2), 461.

Krenn, M., & Zeilinger, A. (2020). Predicting research trends with semantic and neural networks with an application in quantum physics. *Proceedings of the National Academy of Sciences*, *117*(4), 1910–1916.

Kruglanski, A. W. (1990). Lay epistemic theory in social-cognitive psychology. *Psychological Inquiry*, *1*(3), 181–197.

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355–362.

Li, K., Rollins, J., & Yan, E. (2018). Web of Science use in published research and review papers 1997–2017: A selective, dynamic, cross-domain, content-based analysis. *Scientometrics*, *115*(1), 1–20. https://doi.org/10.1007/s11192-017-2622-5

Ludwig, J., & Mullainathan, S. (2023). *Machine Learning as a Tool for Hypothesis Generation*. National Bureau of Economic Research.

Martin, L. J., Harrison, B., & Riedl, M. O. (2016). Improvisational computational storytelling in open worlds. *International Conference on Interactive Digital Storytelling*, 73–84.

McGuire, W. J. (1973). The yin and yang of progress in social psychology: Seven koan. *Journal of Personality and Social Psychology*, *26*(3), 446.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 3111–3119.

OpenAI. (2022). *OpenAI API*. https://beta.openai.com

Park, P. S., Schoenegger, P., & Zhu, C. (2023). *"Correct answers" from the psychology of artificial intelligence* (arXiv:2302.07267). arXiv. https://doi.org/10.48550/arXiv.2302.07267

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *ArXiv Preprint ArXiv:1802.05365*.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

Rocktäschel, T., Grefenstette, E., Hermann, K. M., Kočiskỳ, T., & Blunsom, P. (2015). Reasoning about entailment with neural attention. *ArXiv Preprint ArXiv:1509.06664*.

Sybrandt, J., Shtutman, M., & Safro, I. (2018). Large-scale validation of hypothesis generation systems via candidate ranking. *2018 IEEE International Conference on Big Data (Big Data)*, 1494–1503.

Thompson, A. (2021). *Integrated AI: The rising tide lifting all boats (GPT-3)*. https://lifearchitect.ai/rising-tide-lifting-all-boats/

Turnitin. (2022). *The Turnitin Difference: The largest and fastest growing database*. https://marketing-tii-statamic-assets-us-west-2.s3-us-west-2.amazonaws.com/marketing/our-content-databases_brochure_us_0322.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 5998–6008.

Wilson, S. J., Wilkins, A. D., Holt, M. V., Choi, B. K., Konecki, D., Lin, C.-H., Koire, A., Chen, Y., Kim, S.-Y., & Wang, Y. (2018). Automated literature mining and hypothesis generation through a network of Medical Subject Headings. *BioRxiv*, 403667.

Yuan, W., Liu, P., & Neubig, G. (2021). Can We Automate Scientific Reviewing? *ArXiv Preprint ArXiv:2102.00176*.

Zhang, Y., Gan, Z., Fan, K., Chen, Z., Henao, R., Shen, D., & Carin, L. (2017). Adversarial feature matching for text generation. *International Conference on Machine Learning*, 4006–4015.

Zhou, L., Martínez-Plumed, F., Hernández-Orallo, J., Ferri, C., & Schellaert, W. (2022). *Reject Before You Run: Small Assessors Anticipate Big Language Models*.

Zhu, Y., Lu, S., Zheng, L., Guo, J., Zhang, W., Wang, J., & Yu, Y. (2018). Texygen: A benchmarking platform for text generation models. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 1097–1100.

Supplemental Online Material

Fine-tuning Procedure

The fine-tuning dataset for this research consisted of 103,866 social psychology abstracts. 90% (93,480) of these abstracts were used as training data to fine tune the model and 10% (10,386) were kept aside as validation set. Given the high cost of fine-tuning DaVinci model, we considered Ada, Babbage, and Curie pre-trained models. Our goal was to compare the performance of these three pre-trained models on the validation set when they had been fine-tuned using social psychology abstracts.

For tuning hyper-parameters that control the fine-tuning process of pretrained model, we used GPT-3's default hyperparameter values (OpenAI, 2022). The hyperparameter values that work well across a range of use cases are set as default values. For example, the number of epochs that represents complete passthrough the training data were kept at 4. Batch size, the number of examples from training data used to train a single forward and backward pass, was kept at 128. The learning rate multiplier that controls learning during fine tuning was kept at 0.2.

We compared the three pretrained models on token accuracy, which represents the numbers of tokens correctly predicted by the model. For training data such accuracy is referred to as the training token accuracy and for the validation data it is referred to as validation token accuracy. We observed that during training Babbage had the highest token accuracy, followed by Ada and Curie. However, more importantly, when making prediction on the validation set, Curie shows higher validation token accuracy, followed by Babbage, and Ada. Figure 1 shows the aggregate training and validation accuracy. This indicates that Curie is less likely to overfit on the training data compared to other models that's why it performed better with the validation data and achieved a higher validation accuracy. Figures 2 and 3 shows stepwise change in

Training and Validation token accuracy respectively. Based on validation data performance we chose the fine-tuned Curie model for the subsequent study.

Figure 1

Training and Validation Token Accuracy with Ada (blue), Babbage (purple), and Curie (red) Models

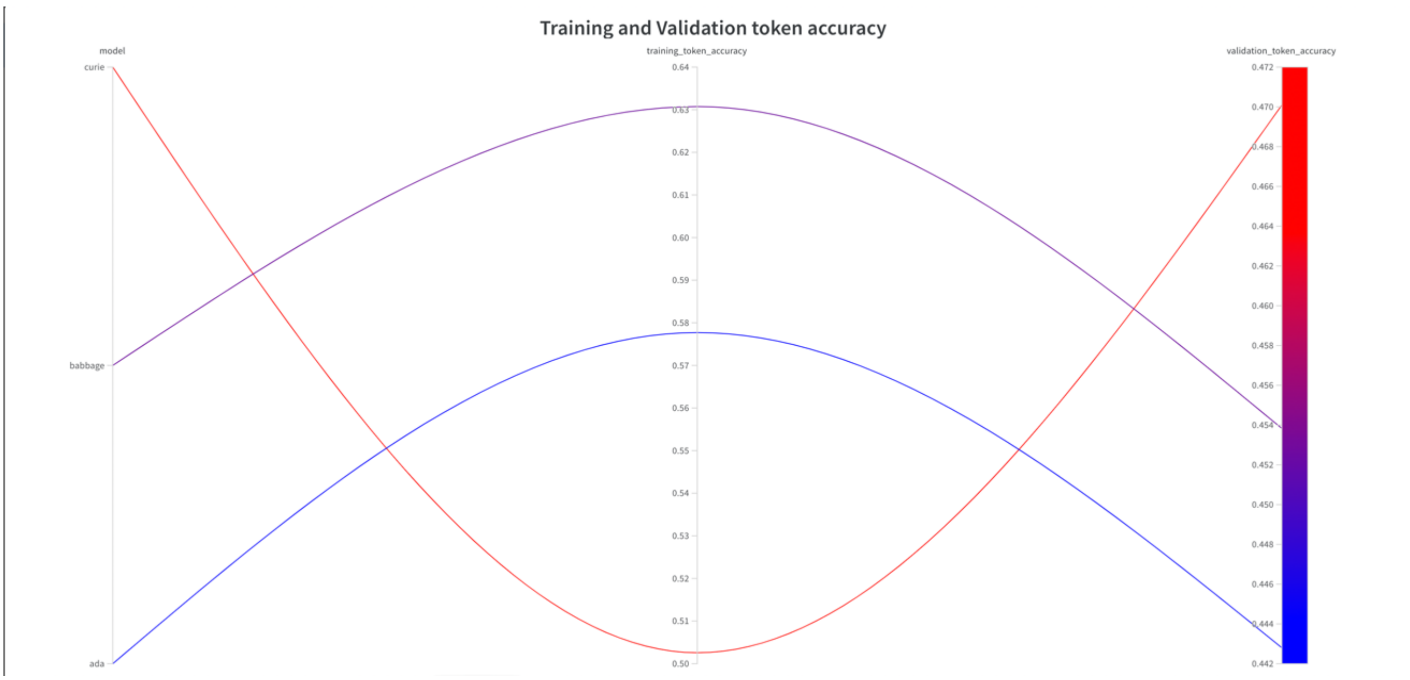
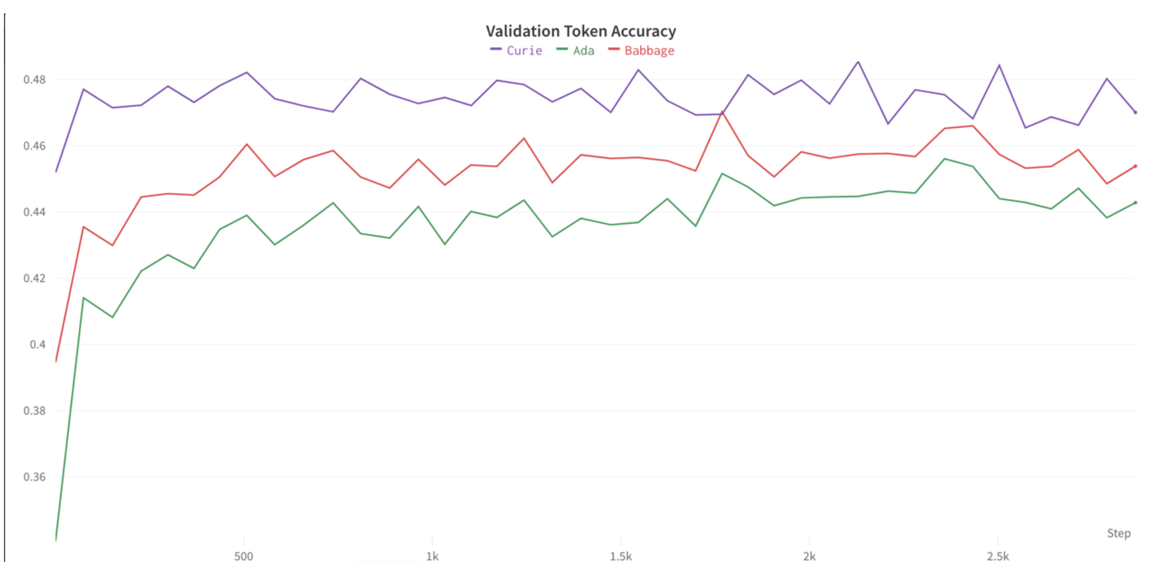


Figure 2

Training Token Accuracy

Figure 3

Validation Token Accuracy



Sample Fine Tuned GPT-3 Model Generated Hypotheses (Temperature = 0.9)

• hypothesize that the conformity phenomenon is not simply a function of a preference for the majority opinion but rather reflects a rejection of the minority opinion
• hypothesize that, given the high cost of the violation of social norms and the readiness to respond with costly revenge, the cost-reward ratio of revenge may be higher for people belonging to high-status groups
• hypothesize that the covariation between the activation of trait- and state-congruent social stereotypes (or person perceptions) would be greater among women than among men
• hypothesize that anticipated vicarious guilt, through the metaperception process of actual guilt, can mediate the relationship between actual guilt and helping behavior
• hypothesize that, as compared with people high in cooperativeness, people low in cooperativeness perceive opportunities more positively, act more cooperatively in relationships, and experience more positive outcomes as a result of their behavior in a social dilemma game
• hypothesize that individuals in collective identities who emphasize their distinctiveness from the ingroup are more likely to perceive membership in the ingroup as an end in itself, and therefore experience negative affect associated with ingroup membership
• hypothesize that peoples motives for improving ones ability vary as a function of the context of their success
• hypothesize that people who are exposed to constant negative information exaggerate their anxiety because they believe that it reflects their current level of anxiety and that others will assume the same level of anxiety in them
• hypothesize that people may perform more poorly on tasks that require taking the perspective of others
• hypothesize that self-affirmation makes people more willing to extend help to others
• hypothesize that consumers are more satisfied with a gift when it is from a stranger than from a friend, relative to gifts from the same acquaintance
• hypothesize that regulatory focus and implicit self-esteem jointly affect the quality of human relations among couples
• hypothesize that the strength of psychological-relational bonds moderates the relationship between perceived relationship commitment and satisfaction
• hypothesize that publicly accessible information is used to maintain positive images of oneself, which are important to one's future interpersonal relationships
• hypothesize that people have a fundamental need to belong and that their social identity is more important than either their national or ethnic identity
• hypothesize that the positive correlation between self-esteem and political interest is not simply due to the fact that politically interested people are more egotistic and assertive, but more likely because they also consider themselves to be good and value traits that go along with this attitude
• hypothesize that visual noise disrupts the communication of affect in three ways: it interferes with communication of feelings from the face; it interferes with face

recognition; and it interferes with the recognition of facial expression associated with particular feelings
• hypothesize that collective action to achieve a common goal generates a spontaneous polarization effect
• hypothesize that efforts to communicate information are to some extent a form of self-presentation
• hypothesize that individuals who are identified with their group and who have relatively high levels of group identification experience more intense positive affect after group-relevant positive events
• hypothesize that target-orientation self-presentation positively predicts the use of self-presentation tactics in email interpersonal negotiations

Sample GPT-4 Model Generated Hypotheses (Temperature = 0.9)

• hypothesize that individuals with higher levels of implicit prejudice towards a certain group (e.g., racial, gender, or religious) will display less violent and aggressive behavior towards that group than those with lower levels of implicit prejudice, due to a compensatory effort to suppress their prejudiced attitudes in mixed-group settings
• hypothesize that individuals belonging to a low-status group within a society will have more positive attitudes towards a high-status outgroup, in comparison to individuals in the high-status group themselves, due to a phenomenon of upward social comparison and aspirational identification
• hypothesize that acts of prosocial behavior in a shared virtual reality environment will lead to decreased real-life prosocial behavior, as individuals will perceive their virtual actions as sufficient to satisfy their moral self-image
• hypothesize that individuals primed with the concept of the self as an independent, autonomous agent will actually exhibit higher levels of conforming behavior in a group setting, in comparison to those primed with the concept of collectivism, due to a compensatory desire for social connection and validation
• hypothesize that exposure to interpersonal rejection will lead to a heightened sensitivity to social influence tactics, as rejected individuals may be more susceptible to the persuasive efforts of others in order to regain social acceptance
• hypothesize that individuals who perceive their partner's love as unconditional will engage in more aggressive behaviors within the relationship, compared to individuals who feel that their partner's love is conditional, due to a sense of security that allows them to express negative emotions more freely
• hypothesize that adopting a group-based identity (e.g., national or religious) will lead to a decrease in prejudiced attitudes towards other groups, in comparison to those who maintain an individualistic identity, as the expanded sense of social connectedness creates a more inclusive worldview
• hypothesize that individuals exposed to violent media content will be more likely to engage in prosocial behavior immediately following exposure, compared to those exposed to non-violent content, as a means of restoring a positive moral self-concept
• hypothesize that members of stigmatized groups who endorse group-based stereotypes will be more successful in making new friendships with members of other groups,

compared to those who reject stereotypes, as endorsement signals flexibility and openness to negotiation
• hypothesize that people who have had their personal beliefs attacked will be more likely to support free speech, compared to those who have not had their beliefs attacked, as a means of preserving their ability to defend and promote their own viewpoint
• hypothesize that individuals who habitually engage in self-affirmation exercises will be more resistant to external social influence, compared to those who do not engage in self-affirmation, due to enhanced self-integrity and self-esteem
• hypothesize that people who regularly participate in online activism (e.g., social media engagement, online petitions) will be less likely to engage in offline activism, compared to those who do not participate online, as a consequence of minimizing cognitive dissonance and justifying their online efforts as sufficient
• hypothesize that an increase in shared group identity among politically polarized individuals will lead to greater willingness to compromise and cooperate on divisive issues, compared to when they are primed with individualistic values and attitudes
• hypothesize that aggressive behavior towards outgroup members will be positively correlated with prosocial behavior towards ingroup members, as the act of aggression may serve to reinforce group identity and cohesion
• hypothesize that individuals who are exposed to ambiguous or mixed messages about a social issue will develop stronger attitudes on the issue, compared to those exposed to clear and consistent messages, as the ambiguity triggers a greater need for cognitive closure and certainty
• hypothesize that people who believe in a just world will display more prejudiced attitudes towards lower-status groups, compared to those who do not believe in a just world, as they may perceive the low-status groups as deserving of their position
• hypothesize that high-status individuals will be more likely to conform in group settings than low-status individuals, due to a higher need for social approval and a desire to maintain their status
• hypothesize that religious individuals will be more likely to engage in prosocial behavior when primed with secular moral concepts, compared to when primed with religious concepts, as a means of expanding their moral repertoire and demonstrating the universality of their values
• hypothesize that individuals who experience frequent self-consciousness will be more likely to engage in prosocial behavior, compared to those who do not experience self-consciousness, as a means of enhancing their self-esteem and social approval
• hypothesize that individuals who are primed with the idea of a common human identity will be less likely to conform to group norms, compared to those primed with the idea of distinct social categories, as a result of a reduced need for group differentiation and validation
• hypothesize that exposure to media promoting idealized romantic relationships will lead to increased aggression within real-life romantic relationships, as individuals may experience frustration and disappointment when reality does not meet expectations

• hypothesize that prosocial behavior will be more likely to occur in the presence of an audience, compared to when individuals are alone, due to increased social desirability and a desire for public recognition
• hypothesize that individuals who feel a strong sense of belonging to a specific social group will be more likely to endorse and perpetuate negative stereotypes about their own group, compared to those who do not feel a strong sense of belonging, as a means of affirming and maintaining their group identity
• hypothesize that consistent exposure to positive news stories about a particular racial or ethnic group will lead to increased prejudiced attitudes towards that group, as individuals may perceive the positive coverage as overcompensation and become suspicious of hidden negative attributes
• hypothesize that individuals who are highly motivated to control their prejudiced attitudes will be more susceptible to persuasion techniques aimed at increasing their prejudice, due to a heightened attention and sensitivity to relevant information

Sample Human Hypotheses

• hypothesize that social and physical pain overlap in chronic conditions as well
• hypothesize that respondents often employ an anchoring and adjusting strategy in which their response to an initial survey item provides a cognitive anchor from which they insufficiently adjust in answering the subsequent item
• hypothesize that athletes display an impact bias and, counterintuitively, that increased experience with an event increases this impact bias
• hypothesize that, by increasing competition and by reducing peoples' sense of connection to others, neoliberalism can increase loneliness and compromise our well-being.
• hypothesize that information filtering processes take place on the individual, the social, and the technological levels (triple-filter-bubble framework)
• hypothesize that rapid social change in the form of polarization results from the interplay between small group processes and perceptions of society at large. by employing a novel analytic approach that uses variances to capture non-linear societal change, we were able to study polarization processes
• hypothesize that the American flag should heighten different political beliefs depending on individuals' political ideology
• hypothesize that goals can have a broader and more dynamic impact on behaviour and, specifically, that goal desires can moderate the effect of intentions on behaviour
• hypothesize that, as a stylistic bias, sd would increase (a) the importance people attribute to values in general and (b) lead people to match own value ratings to those of importance in their social environment
• hypothesise that regulatory focus moderates the relationships between anticipated emotions of success and failure of performing an act and evaluations of the act
• hypothesise that the effectiveness of threats and encouragements is contingent on the intended recipient's level of negative affect, as evidenced by his/her negative affective display

• hypothesise that the association between mood and level of goal/action identification is impaired in depression
• hypothesise that the same emotional expression can signal different social messages and, therefore, trigger different reactions; which social message is signalled by an emotional expression should be influenced by moderating variables, such as the group membership of the expresser
• hypothesise that goal-driven modulation most strongly impacts delayed disengagement from threat
• hypothesise that rumination is a central mechanism underlying the maintenance of ative emotions
• hypothesize that dorsal hippocampal neurons, which are critical for episodic memory of personal experiences, form a memory of a meal, inhibit meal initiation during the period following that meal, and limit the amount ingested at the next meal
• hypothesize that several components of human language, including some aspects of phonology and syntax, could be embedded in the organizational properties of the motor system and that a deeper knowledge of this system could shed light on how language evolved
• hypothesize that such processing asymmetry results from greater experience with female faces than with male faces early in development
• hypothesize that the photograph helps subjects to imagine details about the event that they later confuse with reality
• hypothesize that storage is mediated by the same brain structures that process perceptual information and that rehearsal engages a network of brain areas that also controls attention to external stimuli
• hypothesize that the emotions of fear and anxiety are separable. the authors tested their hypothesis in two studies
• hypothesize that people may have multiple representations of a preference toward an object even within a single context
• hypothesize that affective processes are susceptible to similar automatic influences

List of Social Psychology Journals used for Fine-tuning the Model

The hypotheses used for fine-tuning the second stage of the generative model was obtained from the journals below. We also included abstracts from preprints on PsyArxiv in addition to publications from this list.

Advances in Experimental Social Psychology
Annual Review of Psychology
Basic and Applied Social Psychology
British Journal of Social Psychology
Cognition and Emotion
Current Directions in Psychological Science
Current Research in Social Psychology
Emotion
European Journal of Social Psychology

European Review of Social Psychology
Group Processes and Intergroup Relations
Journal of Applied Psychology
Journal of Applied Social Psychology
Journal of Behavioral Decision Making
Journal of Consumer Psychology
Journal of Consumer Research
Journal of Cross-Cultural Psychology
Journal of Economic Psychology
Journal of Environmental Psychology
Journal of Experimental Psychology: Applied
Journal of Experimental Psychology: General
Journal of Experimental Social Psychology
Journal of Language and Social Psychology
Journal of Organizational Behavior
Journal of Personality and Social Psychology
Journal of Risk and Uncertainty
Journal of Social and Personal Relationships
Journal of Social and Political Psychology
Journal of Social Issues
Journal of Social Psychology
Judgment and Decision Making
Motivation and Emotion
Organizational Behavior and Human Decision Processes
Personality and Social Psychology Bulletin
Personality and Social Psychology Review
Perspectives on Psychological Science
Psychological Bulletin
Psychological Review
Psychological Science
Psychological Science in The Public Interest
Public Opinion Quarterly
Self and Identity
Social and Personality Psychology Compass
Social Cognition
Social Influence
Social Issues and Policy Review
Social Psychological and Personality Science
Social Psychology
Social Psychology Quarterly

The American Psychologist
Theory and Decision

Hypothesis Validation Test Materials

Instructions provided to all participants prior to rating the hypotheses.

Evaluating Research Hypotheses
Instructions

- In this survey, you will be presented with **30** different social psychology research hypotheses.
- When reading the hypotheses, try to **focus on the proposed variable relationship being described**.
- To give you a sense for what to expect, here are a few **examples** of hypotheses:
    1. hypothesize that people who are high in both conscientiousness and agreeableness are sensitive to the social climate and react more positively to favorable than unfavorable social interaction
    2. hypothesize that the gender differences in aggression, both physical and verbal, are present from the youngest of children to the oldest of adults
    3. hypothesize that a person's perception of the self is less socially defined when he/she holds a complex versus a simple schema of others
    4. hypothesize that feelings of autonomy promote future self-control but that feelings of relatedness promote future self-control to a greater extent

Evaluating Research Hypotheses
Instructions

- When evaluating each hypothesis, we would like you to rate them on the following 3 dimensions:

1. Clarity
How precise is the hypothesis?
How easy is it to understand the main idea of the hypothesis?

2. Originality
How innovative or creative is the hypothesis?

3. Impact
What is the theoretical or practical importance of the hypothesis on social psychology and related fields?

Please read the rating dimensions carefully before you move on to rate the hypotheses.

Evaluating Research Hypotheses
Instruction Check

To ensure you're understanding what the rating dimensions are referring to, please answer the following questions.

Consider the following hypothesis:
hypothesize that gender prejudice occurs more in gendered rather than genderless languages

When evaluating the **clarity** of this hypothesis, I will consider:

- ○ how innovative & creative the hypothesis is
- ○ how precise & easy-to-understand it is
- ○ the theoretical & practical importance of the hypothesis to the field

When evaluating the **originality** of this hypothesis, I will consider:

- ○ the theoretical & practical importance of the hypothesis to the field
- ○ how precise & easy-to-understand it is
- ○ how innovative & creative the hypothesis is

When evaluating the **impact** of this hypothesis, I will consider:

- ○ how precise & easy-to-understand it is
- ○ the theoretical & practical importance of the hypothesis to the field
- ○ how innovative & creative the hypothesis is

Example scales used to rate each hypothesis.

Hypothesis:

hypothesize that gender prejudice occurs more in gendered rather than genderless languages

	Very Low	Low	Neutral	High	Very High
Clarity (precise/easy-to-understand?)	○	○	○	○	○
Originality (innovative/creative?)	○	○	○	○	○
Impact (theoretical/practical importance?)	○	○	○	○	○

Experimental Instructions with Additional Dimensions

Evaluating Research Hypotheses
Instructions

- When evaluating each hypothesis, we would like you to rate them on the following 5 dimensions:

1. Clarity
How precise is the hypothesis?
How easy is it to understand the main idea of the hypothesis?

2. Originality
How innovative or creative is the hypothesis?

3. Impact
What is the theoretical or practical importance of the hypothesis on social psychology and related fields?

4. Plausibility
How credible do you believe the hypothesis to be?

5. Relevance
How pertinent is the hypothesis to the field of social psychology?

Please read the rating dimensions carefully before you move on to rate the hypotheses.

hypothesize that gender prejudice occurs more in gendered rather than genderless languages

	Very Low	Low	Neutral	High	Very High
Clarity (precise/easy-to-understand?)	○	○	○	○	○
Originality (innovative/creative?)	○	○	○	○	○
Impact (theoretical/practical importance?)	○	○	○	○	○
Plausibility (credible/believable?)	○	○	○	○	○
Relevance (pertinent to the field?)	○	○	○	○	○