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

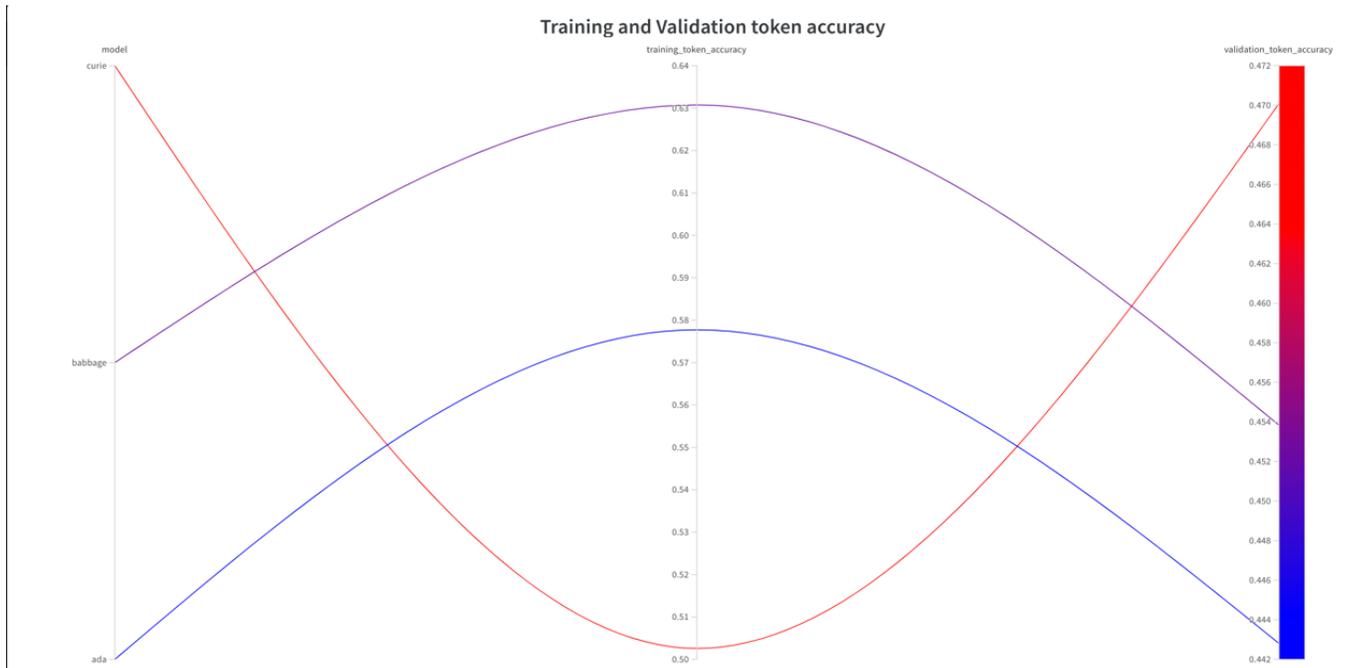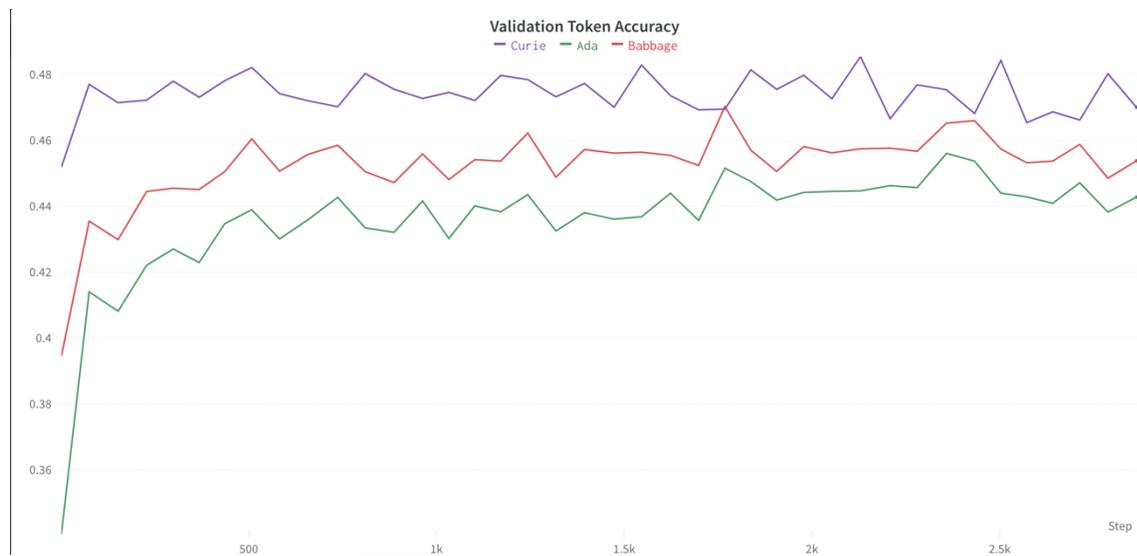*Training and Validation Token Accuracy with Ada (blue), Babbage (purple), and Curie (red) Models*

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?)	○	○	○	○	○