Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Junaid Ahmed April 8, 2021

Can NLP Models Aid in Behavioral Economics Decision-Making?

by

Junaid Ahmed

Dr. Stephen O'Connell
Adviser

Department of Economics

Dr. Stephen O'Connell
Adviser

Dr. Mike Carr
Committee Member

Dr. Phillip Wolff
Committee Member

2021

Can NLP Models Aid in Behavioral Economics Decision-Making?

By

Junaid Ahmed

Dr. Stephen O'Connell
Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Arts with Honors

Economics / Mathematics

2021

Abstract

Can NLP Models Aid in Behavioral Economics Decision-Making?
By Junaid Ahmed

Natural Language Processing (NLP) models have seen rapid improvements in the last two years. Literature has indicated that these models are capable of reasoning, and in certain cases, reason better than humans. While behavioral economics tends to focus exclusively on human subjects, this study seeks to evaluate how NLP models fare in comparison to human subjects during cognitive bias tasks. More specifically, we evaluate how RoBERTa responds to fill-in-the-blank questions based on the conjunction fallacy. We use the conjunction fallacy due to its mathematical falsifiability and ease-of-testing. The hypothesis guiding this study is that RoBERTa outperforms human subjects. From this study, we conclude that RoBERTA does not outperform human subjects in aggregate, but shows promise for individuals prone to the conjunction fallacy, suggesting that there is value in future research. Moving forward, we recommend that other NLP models undergo similar tests across a greater range of cognitive biases to more accurately assess whether there is potential for using NLP models as external aids to decision making.

Can NLP Models Aid in Behavioral Economics Decision-Making?

By

Junaid Ahmed

Dr. Stephen O'Connell
Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Arts with Honors

Economics / Mathematics

2021

Acknowledgements

In no particular order, I would like to thank my advisors for working with me on this project:

Dr. Stephen O'Connell for helping me refine these ideas into worthwhile research relevant to economics and suggesting the use of human subjects

Dr. Mike Carr for his patience and raising questions about validity, prompting me to write the background in greater detail and input more questions into RoBERTa

Dr. Phillip Wolff for his immense tutelage, providing an introduction to RoBERTa, RoBERTa's code, and inspiring many of the ideas in this paper

1 Introduction

Behavioral economics research began as an observation of choices and preferences, and now moves towards heuristics, bias, and risk. However, whereas the bulk of behavioral economics literature originates from human subjects research, it is also possible to explore behavioral topics via a technological lens. Thus, today's advancements in research typically rely on two methodologies: agent-based modeling and machine learning (Baddeley, 2019).

This paper focuses on the latter. In particular, we take a mathematically-provable judgment error–the conjunction fallacy–and apply a machine learning model to measure its robustness against this fallacy. The machine learning model in question is RoBERTa (Liu et al., 2019), a natural language processing model. This model accepts fill-in-the-blank questions as inputs; then, it outputs the numerical probability of potential answers. We use this fill-in-the-blank mechanism to evaluate the usefulness of RoBERTa against the conjunction fallacy, comparing the model's probability rankings for fallacious answers against its rankings for correct ones. The model's accuracy is the measured against survey responses from human participants who answered the same fill-in-the-blank questions.

We judge RoBERTa's utility based on three levels. At its most useful, the model is always accurate, choosing the correct answer in every instance. This ideal is highly unlikely, due to the innate presence of error and randomness within the model. At its second best, the model performs more accurately than human respondents. This scenario is much more probable and suggests that the model detects something that human participants do not. At third best, the model performs better than random chance, suggesting that there is some solving mechanism, albeit not sophisticated enough to outperform human judgment.

If RoBERTa proves to be useful under any of these three criteria, the study findings serve as preliminary evidence that RoBERTa may be used to assist decision-making and behavioral economics problem-solving. These findings would also suggest RoBERTa may be able to compensate for failures of system 2 thinking. Based on these findings, NLP tools may eventually be used to assist economic decision-making tasks in the future, akin to how doctors rely on machine learning predictions as a second opinions (Kapoor & Mishra, 2018; Raghu et al., 2019). In an economic context, RoBERTa may be used to aid purchasing decisions in which the conjunction fallacy is involved. For example, the "less is more" error occurs when consumers evaluate product sets based on averages as opposed to the independent value of each item; this phenomenon is structured identically to the conjunction fallacy questions shared by Kahneman

(2011) and could possibly be overcome using RoBERTa, if the findings of the study prove favorable.

Based on past literature, the hypothesis guiding this study is that RoBERTa will perform equal to or better than human participants. Using an original survey of 20 conjunction fallacy questions and 53 human participants, this paper concludes that RoBERTa currently performs at the third utility level. Compared to the aggregate sample of survey participants, RoBERTa underperformed by 10%. This value exceeds random chance expectations but is still insufficient to assist with the conjunction fallacy in all cases. However, certain individuals failed up to 2.8x more questions than RoBERTa, suggesting that the model may be useful in particular cases, especially for respondents more prone to the conjunction fallacy. For these individuals, RoBERTa may offer an objective and less erroneous framework for decision-making.

2 Background

2.1 The Model

2.1.1 Machine Learning

Machine learning is a subfield of computer science which involves algorithms "learning" from large amounts of data (Hao, 2018). A machine learning model attempts to solve problems or discover patterns based on an algorithm (Hao, 2018).

For example, a machine learning model attempting to "learn" the visual characteristics of a cat would require at minimum, a few thousand images of cats. Given these images, the model records patterns such as the shape, size, color, and so on, of cats; this process is known as "training." The model may also pick up on unforeseen patterns, such as ear shape. Then, if given unlabelled images as inputs, the model becomes capable of outputting the probability that the image contains a cat. This entire process is known as image classification or image recognition.

Certain machine learning applications, such as the recognition of cats, may seem trivial, yet the same methods are revolutionizing the medical and transportation industries. For example, image classification is being used to detect COVID in X-ray images (Albahli & Albattah 2020) and develop autonomous vehicles (Fujiyoshi et al., 2019). In this paper, the subset of machine learning used will be Natural Language Processing (NLP).

2.1.2 Natural Language Processing (NLP)

NLP is a subfield of machine learning concerned with how computers understand text and spoken words (Yse, 2019). NLP has been used for clinical applications, such as predicting mental illness from patterns in language (Thorstad, R. & Wolff, P., 2019; Corcoran et al., 2019), and automated translation, as demonstrated by Google Translate (Hirschberg & Manning, 2015).

NLP models can be fed textual questions to receive textual answers. For fill-in-the-blank questions in particular, NLP models use various language-based measures, such as semantic analysis (i.e. valence) and text classification (i.e. grammar), to calculate the probability of different outputs (Wolff, R., 2020). These fill-in-the-blank questions may be adapted into a multiple-choice survey with ease, allowing us to compare the computer-generated responses versus survey participants' responses. This implementation is our focus for this paper.

2.1.3 RoBERTa

The particular NLP model used in this paper is an improved version of Google's BERT (Devlin et al., 2018). BERT was chosen as our foundation because it outperformed state-of-the-art NLP models in October 2018 (Khan, 2019) and has been researched extensively. For example, BERT has been used to improve Google search results; more specifically, BERT allows Google searches to output the intent of queries rather than simply matching keywords (Nayak, 2019). For example, searching "ticket Venezuela" will search for flights as opposed to speeding tickets or bus tickets. In addition, BERT has been used as the foundation of many new NLP models.

In 2019, Facebook AI publicly released an improved version of Google's BERT. This model is known as RoBERTa (Liu et al., 2019) and is regarded as an improvement by the GLUE (Wang et al., 2019), SQuAD (Rajpurkar et al., 2018), and RACE (Lai et al., 2017) benchmarks.

As of March 2021, DeBERTa (He et al., 2021) was built on BERT, and was ranked second by the GLUE benchmark, yet we prefer RoBERTa for three primary reasons. First, it is impractical to focus on the newest NLP models because the machine learning landscape is developing rapidly– newer models are released approximately every few months, with little time to accumulate research. RoBERTa has been released for approximately two years, allowing sizable research to be conducted and various applications to be made. For example, source code is available on GitHub for the fill-in-the-blank question style used in this paper, allowing RoBERTa to output either the likelihood of given answers or a ranked list of the most likely answers (Scheible, 2019). There is also code available that uses RoBERTa to disambiguate pronouns (Scheible,

2019), to increase typing speed on Android smartphones (Subudhi, 2020), and to detect fake news (Slovikovskaya, 2019).

Additionally, other improvements upon BERT trade prediction metrics in favor of computational speed. For example, compared to BERT, DistilBERT (Sanh et al., 2019) reduces training time by 4x and improves inference speed, but at the cost of 3% in prediction accuracy (Khan, 2019). In comparison, RoBERTa performs up to 20% better than BERT (Khan, 2019). Although RoBERTa requires more training time than both BERT and DistilBERT combined (Khan, 2019), the model comes to us pre-trained; thus, RoBERTa can be used immediately without requiring a large amount of training data or the time required to process it. Moreover, RoBERTa improves upon more than just BERT, including models such as ELMo (Peters et al., 2018), GPT (Radford et al., 2018), XLM (Lample and Conneau, 2019) and XLNet (Yang et al., 2019), detailed by Liu et al. (2019).

Lastly, RoBERTa is preferred for this study because, though more sophisticated models may exist, they are closed-source. In other words, the models are unavailable to the general public and may require large licensing costs to access. For example, GPT-3 (Brown et al., 2020) is an NLP model which gained significant popularity in Q3 of 2020. Founded by OpenAI, GPT-3 is the largest NLP model as of October 2020 (Marr, 2020). In fact, its text generation capabilities were nearly indistinguishable from human-generated text (Sagar, 2020), prompting the authors to include a section detailing dangers and risk (Brown et al., 2020). It is capable of generating creative fiction, making designs for web apps, and creating code, all via text descriptions (Marr, 2020). Unfortunately, as of September 2020, the cost to license the GPT-3 API starts at a minimum of $100 per month for 900,000 words, becoming more costly as usage increases (Bhavsar, 2020). Thus, at present, it is most cost-effective to begin with RoBERTa.

However, this analysis using RoBERTa is not intended to be comprehensive, but rather open-ended for further research and discussion. Analysis of other NLP models may yield stronger or more varied conclusions informing us of patterns in language. For example, if RoBERTa were to perform better on reasoning tests than GPT-3, it may serve as evidence that a sort of "information overload" (Yang et. al, 2003) occurs in not just humans, but NLP models as well. Or, it may suggest that information overload is a result of language-based learning, as opposed to experience-based.

2.1.4 Evidence for Logical and Inferential Reasoning Capabilities

In order to judge whether RoBERTa is capable of answering questions involving bias and judgment errors, it must first be established that RoBERTa contains knowledge and is capable of reasoning. These premises are established on three fronts: training data, pre-training, and tests of predicate logic rules.

RoBERTA's answers are based on five English-language datasets totaling at least 160GB (Liu et al., 2019). In addition to the books and Wikipedia data used to train BERT, RoBERTa includes news articles and web content from popular links. Most importantly, RoBERTa is trained on the Stories dataset, which includes multiple choice questions on reasoning tests (Trinh & Le, 2019). Trinh and Le (2019) have used this dataset to train a large number of language models. These models were subsequently tested to prove that commonsense knowledge may be embedded via training data (Trinh & Le, 2019). This is verified by Liu et al. (2019) testing RoBERTa against the RACE benchmark, a dataset of questions collected from English examinations for middle-school and high-school students in China. RACE contains 28,000 passages with nearly 100,000 multiple-choice questions.

Secondly, Wang et al. (2020) asserts that recent language models obtain knowledge automatically from this training data during pre-training. This is evidenced by improvements in various NLP tasks such as answering questions, writing poetry, and composing music. Wang et al. (2018) specifically mentions RoBERTa as capable of outperforming humans on tasks such as sentence classification, with the additional advantage that RoBERTa did not require the time and resource costs required to train a human (Wang et al., 2020). The authors have also built knowledge graphs akin to WikiData[1] based on NLP models such as BERT and GPT-3, finding that these knowledge graphs contain factual knowledge that is new to existing knowledge graphs.

We also conduct tests of logical reasoning by inputting predicate logic problems from the open source textbook *forallX: An Introduction to Formal Logic* by P.D. Magnus (2017). Fill-in-the-blank questions are used as input with blanks coded as "<mask>." RoBERTa is then asked to provide a ranked list of its top answers. These tests investigate whether RoBERTa ranks the correct answer as its first choice.

These tests are conducted on a general-purpose model such as RoBERTa as opposed to a fine-tuned model capable of solving conjunction fallacy questions because it allows us to observe

---

[1] https://www.wikidata.org/

whether bias is embedded within language models more generally. As mentioned prior, it also allows us to observe whether language models obtain bias as a result of human-like failures, such as information overload (Yang et al., 2003) or the over-reliance on language-based learning over experience-based. An ideal NLP tool would not just allow us to solve questions containing cognitive biases, but model decision-making in humans to rectify the error at its source. For example, a doctor using a predictive diagnostic tool would ideally improve their diagnoses over time. Modeling a single bias is a first step in mapping human deficiencies in decision-making.

These tests are not intended to be extensive. Rather, they are used to provide a general understanding of RoBERTa's limits by serving as falsifying tests. In other words, a successful test result is a necessary, but insufficient condition. For example, if RoBERTa is consistently incapable of processing a logical rule such as Modus Tollens, we will omit Modus Tollens from our questions. However, if another rule such as Modus Ponens yields a successful test result, it will continue to be under scrutiny. This is because, at best, successful test results should lead to more tests, and a greater variation of them. This is the same standard by which RoBERTa was tested on the RACE benchmark; no single test provides sufficient evidence, but each test provides more evidence. The tests are contained in the following section.

2.1.5 Testing for Logical and Inferential Reasoning Capabilities

Table 1: Testing RoBERTa's Ability to Understand Validity and Soundness

Type of Test	**Input**	**Output**	**Correct Result?**
Valid and Sound	Socrates is a man. All men are mortal. Therefore, Socrates is <mask>.	moral	TRUE
Valid yet Unsound	Oranges are either fruits or musical instruments. Oranges are not fruits. Therefore, oranges are <mask>.	not	FALSE
	Socrates is a man. All men are carrots. Therefore, Socrates is a <mask>.	carrot	TRUE
Invalid yet Sound	London is in England. Beijing is in China. Therefore, Paris is in <mask>.	France	TRUE

Table 2: Testing RoBERTa's Ability to Solve Abstract Arguments

Type of Test	**Input**	**Output**	**Correct Result?**
Abstract Argument (Easy)	S is M. All Ms are Cs. Therefore, S is <mask>.	C	TRUE
Abstract Argument (Medium)	A. If A, then C. Therefore, <mask>.	B	FALSE
	A is true. If A, then C. Therefore, <mask>.		
	A is true. If A is true, then C. Therefore, <mask>.		
	A is true. If A is true, then C is true. Therefore, <mask>.		
Abstract Argument (Medium & Verbose)	A is true. If A is true, then C is true. Therefore, <mask> is true.	C	TRUE
	A is true. If A is true, then C is true. Therefore, C is <mask>.	true	TRUE
Abstract Argument (Hard)	L implies (N or E). E implies B. L is true. Therefore, not B implies <mask>.	N	TRUE

Table 3: Testing RoBERTa's Ability to Understand Implicit Conditionals

Type of Test	**Input**	**Output**	**Correct Result?**
Implicit Conditional	Unless you wear a jacket, you will catch a cold. You will catch a <mask> unless you wear a jacket.	cold	TRUE
	Unless you wear a jacket, you will catch a cold. You will catch a cold unless you wear a <mask>.	jacket	
	Unless you wear a jacket, you will catch a cold. You can either wear a jacket or catch a <mask>.	cold	
	Unless you wear a jacket, you will catch a cold. You can either wear a <mask> or catch a cold.	jacket	
	Unless you wear a jacket, you will catch a cold. You can either wear a jacket or <mask> a cold.	catch	
Implicit Conditional (Same Verb for Both Options)	Unless you wear a jacket, you will catch a cold. You can either get a jacket or get a <mask>.	cold	TRUE
Implicit Conditional (Advanced Example)	If Zoog remembered to do his chores, then things are clean but not neat. If he forgot, then things are neat but not clean. Therefore, things are either neat or clean – but not <mask>.	both	TRUE

Table 4: Testing RoBERTa's Ability to Understand Quantifiers and Quantifier Negation

Type of Test	**Input**	**Output**	**Correct Result?**
Quantifier Negation	There is some x such that x is not happy. Thus, not all x are <mask>.	happy	TRUE
Quantifier Negation (Antonym)	There is some x such that x is unhappy. Thus, not all x are <mask>.	unhappy	FALSE
Quantifier Negation (Colloquial Syntax)	There exists some x that is unhappy. Thus, not all x are <mask>.	happy	TRUE
Quantifier Negation (Sentence Contains More Explicit Opposites and Uses a Masked Verb)	There is some x such that x is blind. Thus, not all x can <mask>.	see	TRUE
	There exists some x that is blind. Thus, not all x can <mask>.		
Quantifier Negation (Sentence Contains More Explicit Opposites and Uses a Masked Adjective)	There exists some x that is blind. Thus, not all x are <mask>.	blind	FALSE
Quantifier Negation (Sentence Contains More Explicit Opposites and Uses a Masked Noun)	There is some x such that x is blind. Thus, not all x have <mask>.	vision	TRUE
	There exists some x that is blind. Thus, not all x have <mask>.	sight	

Table 5: Testing RoBERTa's Ability to Understand Sets and Properties of Sets

Type of Test	**Input**	**Output**	**Correct Result?**
Set Theory	Willard is a logician. All people who are logicians wear funny hats. Therefore, it is <mask> that Willard wears a funny hat.	obvious	TRUE*
Set Theory (using "therefore")	Willard is a logician. For all x, if x is a logician, then x wears a funny hat. Therefore, it is <mask> that Willard wears a funny hat.	true	TRUE
Set Theory (using "thus")	Willard is a logician. For all x, if x is a logician, then x wears a funny hat. Thus, it is <mask> that Willard wears a funny hat.	true	TRUE
Set Theory (Advanced)	For all people, if not all people are surgeons and all people are skilled, then the hospital will not hire all people. For all people, if all people are a surgeon, then all people are greedy. All people are a surgeon and not all people are skilled. Therefore, all people are <mask> and the hospital will not hire all people.	greedy	TRUE

Table 6: Testing RoBERTa's Ability to Understand Disjunctions

Type of Test	**Input**	**Output**	**Correct Result?**
"OR" Introduction (Abstract Argument)	A is true. <mask> or B is true.	Success	FALSE
	A is true. <mask> or not B is true.	Whether	
"OR" Introduction (Explicit Argument)	The object is an apple. The object is the <mask> or it is the orange.	objective	
	It is true that the object is an apple. The object is the <mask> or the orange.	fruit	
"OR" Introduction (including the word "either")	It is true that the object is an apple. The object is either the <mask> or the orange.	apple	TRUE
"OR" Elimination	A or B is true. B is false. Therefore, <mask> is true.	A	TRUE
	R or F is true. If R is not true, then <mask>.	F	

Table 7: Testing RoBERTa's Ability to Understand Rules Involving Conditionals and Negation

Type of Test	**Input**	**Output**	**Correct Result?**
Conditional Introduction	Assume M is false. M or D is true. Thus, <mask> implies D.	M	TRUE
	Assume M is false. M or D is true. Thus, M implies <mask>.	D	
Conditional Elimination	R implies F. R is true. Therefore, <mask>.	F	TRUE
	R implies F. R is true. Therefore, <mask> is true.		
Modus Tollens	V implies C. C is not true. Therefore, <mask> is false.	V	TRUE
Dilemma	A or B. A implies C. B implies C. Therefore, <mask>.	C	TRUE
Hypothetical Syllogism	A implies B. B implies C. A implies <mask>.	B	FALSE
	A implies B. B implies C. <mask> implies C.	A	TRUE
Double Negation	Not A is false. Therefore, A is <mask>.	FALSE	FALSE
	It is false that A is false. Therefore, A is <mask>.		

2.1.6 Analysis of Tests

Magnus (2017) defines a deductively valid argument as an argument in which it is impossible for the premises to be true and the conclusion false. A sound argument is one that can be true. The tests in Table 1 evaluated whether RoBERTa has the capacity to understand valid, sound, invalid, and unsound arguments.

Sound arguments were simple to process for RoBERTa, regardless of whether they were valid or invalid. However, the valid, yet unsound argument in Table 1 did not produce a correct answer in the first instance, but did so for the second. It is well-documented that other NLP models such as GPT-3 are incapable of understanding "nonsensical" arguments. Thus, it seems likely that RoBERTa may be unable to process unsound arguments as well. We anticipate that this problem will not arise in our experiment because all of our questions are sound.

The tests in Table 2 evaluated whether RoBERTa understood abbreviated sentences of the form "if A, then B." Magnus (2017) terms these abstract arguments as sentential logic. It seems RoBERTa may be able to distinguish sentence letters and singular versus plural–which is expected. Additionally, capitalization has no bearing on its ability to interpret the sentence.

Further testing shows that RoBERTa is incapable of understanding abstract arguments. Each of the sentences which follow the first attempted to increase the context in order to "guide" RoBERTa towards the correct answer, but fail. This may be due to A, B, and C being in alphabetical order. As we saw in the previous example with S and M, this problem did not arise. However, when each letter is explicitly assigned as "true", RoBERTa answered correctly.

Subsequent tests used conditionals and disjunctions, and avoided the use of letters in alphabetical order. The last example in Table 2 demonstrated that RoBERTa may be capable of evaluating abstract arguments which include a conditional and a disjunction. The last example in particular demonstrates the use of "Modus Tollens," to infer that "Not B" logically implies "Not E." In order to solve this test, RoBERTa underwent multiple steps, detailed in Proof 1 in the Appendix. Nevertheless, we do not anticipate using abstract arguments during our experiment.

The tests in Table 3 evaluated whether RoBERTa is able to process implicit conditionals, which are conditional statements that have been translated using "and/or" syntax. Implicit conditionals do not use "If...then" structure, yet have the same meaning. For example, "If I study for the test, I will do well" can be translated to "I can either study for the test or do poorly."

This example is formalized as:

$$\neg A \rightarrow B = \neg B \rightarrow A$$
$$A \rightarrow \neg B = A \vee B$$

In the first few examples of Table 3, RoBERTa correctly identifies the implicit conditional in each of these instances. However, because the same structure "wear a jacket" and "catch a cold" was used in both sentences, the following tests used a more ambiguous verb that could be used with both "jacket" and "cold." Even so, RoBERTa passed the implicit conditional test.

The tests in Table 4 tested whether RoBERTa understood Quantifiers and Quantifier Negation. The universal quantifier ∀ denotes "for all," and the existential quantifier ∃ denotes "exists." Predicates are expressions denoted by a capital letter. For example, if I were to say all "x" are happy, I would write it ∀xHx. If I were to say there exists an "x" that is happy, I would write it ∃xHx.

In Table 4, we evaluated whether RoBERTa understands that, if a person without a condition exists, not everyone has the condition. This rule is formalized as $\exists x \neg Hx = \neg \forall x Hx$: "There exists a person that is not happy" is equivalent to "not all people are happy."

During testing, RoBERTa was able to use "quantifier negation." A universal rule can be proven incorrect by a single counterexample. RoBERTa can deduce that the existence of an object with quality Q means that not all objects have $\neg Q$. This could simply mean RoBERTa is pattern-matching, which the next tests assess.

We checked whether RoBERTa equates "not happy" to "unhappy." RoBERTa failed this test. However, this may also be an issue with the definition of happy and unhappy, and whether "not happy" entails "unhappy" by necessity. Happiness and unhappiness may not be antonyms by necessity because the lack of happiness or unhappiness may be apathy. This error could have also been due to syntax, but that hypothesis was invalidated by the subsequent test which did use more colloquial syntax.

Thus, the final tests in Table 4 evaluate whether RoBERTa can use quantifier negation with qualities that are more explicitly antonyms. When given words with clear opposites such as "blind" and "see," RoBERTa may be able to infer the *verb* used in quantifier negation (in this case, "see"). However, if the first sentence follows the form "All X are [adjective]," and the

second sentence "Thus, not all X are <mask>", RoBERTa inserts the adjective (in this case, "blind") from the preceding sentence to the following one.

RoBERTa is not always be able to infer an adjective, just as we saw in the previous examples with "happiness." Further tests were conducted, and found that RoBERTa may be able to infer nouns from verbs, i.e. first-order effects resulting from a condition: having "blindness" means I do not have "vision."

Table 5 contains one of the two most important tests: whether RoBERTa is capable of set theory. If RoBERTa is unable to recognize sets and properties of sets, it will be incapable of solving problems based on the conjunction fallacy. The first test yields "obvious," which is not the intended answer, but it is a correct answer. It seems RoBERTa may be capable of inferring that if an element exists in a set, it contains the properties of the elements in that set. These same sentences were tested with "therefore" and "thus," but there was no change.

In the final tests in Table 5, RoBERTa was capable of more advanced proofs involving quantifiers, including the use of the rules "∀ Elimination" and "∀ Introduction," while simultaneously ignoring extraneous information. Proof 2 in the Appendix shows the step-by-step proof RoBERTa was able to solve.

The second most important test is whether RoBERTa is capable of disjunction. RoBERTa was incapable of "OR introduction" (e.g. if A is true, A "or" anything is also true), but this will not be required in the experiment. Additionally, when clarifying using the word "either," RoBERTa was able to discern the missing word. RoBERTa must be capable of "OR elimination" for this experiment to be viable, which it was.

In Table 7, RoBERTa was also able to solve various problems involving conditionals, including "conditional introduction" and "conditional elimination," as well as "Modus Tollens," "dilemma," and "hypothetical syllogism." However, it was incapable of double negation.

Although more tests could be conducted on other laws such as De Morgan's laws, it seems these tests were sufficient for the scope of this experiment. In addition to the literature by Trinh and Le (2019) and Wang et al. (2020), there is enough evidence to suggest that RoBERTa may be capable of limited reasoning and inferential capabilities.

2.2 Conjunction Fallacy

2.2.1 Justification

In order to further narrow the scope of this study due to time and resource constraints, we will focus on a single judgment error: the conjunction fallacy. This formal fallacy occurs when individuals believe the union of two independent events is more probable than merely one of the two. It is mathematically impossible for the union to have greater probability.

Being mathematically falsifiable shields the conjunction fallacy from controversy, hence its selection for this study. The classification of cognitive biases as irrational or erroneous has been met with controversy, as counterarguments justifying biases exist (Dougherty et al., 1999; Gigerenzer, 2006). Because there is mathematical evidence proving the bias to be false, there is less controversy surrounding the conjunction fallacy. Mathematical falsifiability also ensures that the conjunction fallacy is independent of linguistic and cultural influences. Because the error is based on logic and not language, the conjunction fallacy remains even if questions are translated into a language other than English. The fallacy also remains consistent across different cultures, unlike non-mathematical cognitive biases such as herd behavior, which vary based on the emphasis on collectivism and conformity (Bond & Smith, 1996).

Secondly, the conjunction fallacy has wide applications to many domains, including psychology. Kahneman (2011) states that the conjunction fallacy is greatly concerning with regards to forecasting, suggesting that this research may have further reach to the natural sciences, such as meteorology. Forecasting is especially valuable in an economic context because lengthy descriptors make events seem more probable than otherwise.

The conjunction fallacy is particularly useful for this study because it has direct applications to economic reasoning. For example, Hsee (1998) and List (2002) find that consumers evaluate a set based on its average rather than the independent value of each item in the set. Given two identical dinnerware sets, adding extra broken dishes to one set lowers its average value, thus lowering the total value in the eyes of consumers. Consumers were willing to pay nearly 40% more for a dinnerware set without the extra broken dishes despite both containing the exact same number of unbroken dishes (Hsee, 1998). Kahneman (2011) states that this “less-is-more” phenomenon is structured identically to the conjunction fallacy. An NLP model capable of identifying the conjunction fallacy my be capable of arbitrage. The value of a set is often based on the average price of its elements; the set should instead be valued on the cumulative price of its elements, a more representative calculation.

2.2.2 Application

Given two events A and B, the conjunction fallacy is written as: $P(A \wedge B) \leq P(A)$ or $P(A \wedge B) \leq P(B)$. One of the more often-cited examples of this fallacy is known as the "Linda problem," originating from Amos Tversky and Daniel Kahneman (1982). A succinct version of the problem (Brogaard, 2016) is as follows:

> "Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations. Which is more probable?
>
> 1. Linda is a bank teller.
> 2. Linda is a bank teller and is active in the feminist movement."

According to Kahneman (2011), an overwhelming majority of up to 90% of undergraduates at several major universities chose the second option. Kahneman (2011) replicates his findings with other examples, both including and excluding background details in order to evaluate whether context bore an effect on judgment. He finds that, with certain scenarios, a substantial minority continue to choose the incorrect answer. However, with less context, individuals were more likely to choose the correct answer (Kahneman, 2011).

However, Tversky and Kahneman's (1982) works are merely observations–they do not explain the origin of these erroneous judgments, how to remedy them, or the best course of action given that these errors exist. At best, they explain that there is a failure in system 2 thinking by use of heuristics. Individuals may ignore lengthy descriptions to focus on the question, restate the question as a math problem, solve the problem more slowly, or become more aware of their biases. This experiment seeks to go beyond these recommendations to assess whether NLP models may be used as an external tool to assist decision-making.

Further research of NLP models may allow us to model heuristics; modeling human judgment errors may allow us the ability to correct these errors at their origin. For example, can problems be presented differently to decrease the incidence of judgment errors? Is there a certain schema involving grammar, language acquisition (sources of training data), or amount of knowledge (quantity of training data) which minimizes bias? Instead of testing judgment error tests on large human populations, these tests could be done using the model, minimizing losses on time, money, and privacy. Another application is a tool that allows us to compensate for failures in

system 2 thinking, akin to how doctors use machine learning tools as a second opinion for their diagnoses.

3 Methodology

3.1 Hypothesis

Once RoBERTa's reasoning and inferential capabilities are established, the next task is to evaluate whether it answers biased questions correctly. This assessment is completed by inputting questions containing the conjunction fallacy into RoBERTa, recording its responses, then comparing its answers to survey participants' answers. The purpose is to establish whether RoBERTa is capable of answering simple behavioral economics questions based on the conjunction fallacy.

These computer-generated responses will be graded on one of three tiers:

1. Does RoBERTa answer the questions correctly every single time?
2. Does RoBERTa answer the questions correctly more often than human participants?
3. Does RoBERTa answer the questions correctly better than a random guess?

At its best and the most ideal, RoBERTa will fall under tier #1. However, this is highly unlikely because all machine learning models have some error based on uncertainty and randomness. This is set as the first-tier standard because although perfection is unachievable, the goal is to get as close to it as possible.

Tier #2 is more likely because of prior research on judgment errors, including Kahneman's work on the conjunction fallacy (Kahneman, 2011). However, because of limited sampling, this experiment may not replicate the same results as Kahneman (2011). Sampling error is discussed in greater detail after the data has been presented.

Tier #3 is most likely because RoBERTa seems to demonstrate some level of understanding when given questions. If RoBERTa is better than random chance, we can expect the probability of success for random guesses to be low. We can verify this by using the binomial distribution formula:

$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!x!} p^x q^{n-x}$$

where

n = number of trials

x = number of successes
p = probability of success (in this case, we would compare it against a coin-flip, 0.5)
q = probability of failure (also 0.5)

3.2 Experimental Design

The questions tested on RoBERTa and the survey participants were not generated based on an official standard. The sole requirement for each question was that one of the two answer choices must be a subset of the other. This requirement was judged based on the Merriam-Webster[2] dictionary definitions of the words. For example, a car is a type of vehicle, thus the set "vehicle" contains all cars, but the set "cars" does not contain all vehicles.

Once twenty questions were generated, they were given to RoBERTa and the survey participants.

3.2.1 Questions for RoBERTa

The questions were input into RoBERTa using code available in the PyTorch/fairseq GitHub repository[3]. The program was able to either output RoBERTa's top-10 list of answers or a logit estimate for a given answer. Logits, in this case, were the non-normalized prediction values of the model.

Logit estimates were compared between the set and subset answers. If the logit values were greater for the set, it was noted as a successful test. If the logit values were greater for the subset, it was noted as a failed test. Comparing exact probability would require comparing all possible answers by applying a softmax function. Rather than computing the likelihood of all possible answers, we simply compare the likelihood between two answers.

A control group was created by stripping all details and measuring whether RoBERTa has a built-in preference for one of the answer choices. A second set of questions was based on little context to evaluate whether RoBERTa behaves the same as Kahneman's (2011) survey participants, who were able to answer correctly more often when details were removed. A third set of questions with increased context was added to test the opposite–whether the model fails more often when given more details.

---

[2] https://www.merriam-webster.com/dictionary/

[3] https://github.com/pytorch/fairseq/blob/master/examples/roberta/README.md

3.2.2 Questions for Survey Participants

The survey was hosted on Microsoft Forms. The survey did not have a time limit and participants were permitted to revise their answers. Participants were given a multiple-choice survey with two response choices: the set and the subset. Participants were only given the original question; they were not given the stripped-down control questions, the questions with decreased context, or the questions with increased context.

Survey participants were recruited at Emory University via publicly available social media channels such as GroupMe and Facebook Groups, as well as the Emory University Economics ListServ. Survey participants were encouraged to share the survey with their friends, thus the sample may contain students from nearby universities. Individuals were not offered any compensation or benefit for their participation.

In order to minimize experimenter demand effects, deceit was implemented into the title. The title shown to participants was "Sex Differences in Personal Narratives and Future Decision-Making." In addition, 10 decoy questions were scattered into the survey, asking the participant whether the frequency of males or females was more prevalent in a profession, or whether they believed one event was more likely than another, e.g. motorcycle versus car accidents.

53 data points were collected. 55 individuals participated in the survey. However, two individuals did not provide names on the consent form, thus these data points were discarded.

With 53 data points, a sample proportion of 50% (maximum possible variation), and a total population of 8079 Emory University undergraduates as of Fall 2019[4], and a 95% confidence interval, we get a margin of error of approximately ±13.5%.

[4] http://www.emory.edu/home/about/factsfigures/

4 Data

Table 8: List of All Questions

No.	Text of Original Question Given to Both RoBERTa and Survey Participants	RoBERTa Only: Questions with All Details Removed	RoBERTa Only: Questions with Some Details Removed	RoBERTa Only: Questions with Some Details Added
1	Murders are a terrible tragedy that happen far too often in the modern world. There are certain places that we know are more dangerous than others. Do more homicides occur in Michigan or Detroit? More homicides occur in <mask>.	More homicides occur in <mask>.	Do more homicides occur in Michigan or Detroit? More homicides occur in <mask>.	The United States must rework various policies to make the country safer. Murders are a terrible tragedy that happen far too often in the modern world. There are certain places that we know are more dangerous than others. Do more homicides occur in Michigan or Detroit? More homicides occur in <mask>.
2	Sarah loves working with numbers and solving complex equations. She is often picked on for being a nerd. Is Sarah more likely to study a STEM or Math subject? She is more likely to study a <mask> subject.	She is more likely to study a <mask> subject.	Is Sarah more likely to study a STEM or Math subject? She is more likely to study a <mask> subject.	Students often study a variety of subjects in school. Sarah loves working with numbers and solving complex equations. She is often picked on for being a nerd. Is Sarah more likely to study a STEM or Math subject? She is more likely to study a <mask> subject.
3	Joe loves cutting things open and understanding the inner workings of the world around him. As a young child, his mother often caught him sewing holes in furniture. Is he more likely to be a surgeon or a doctor? He is more likely to be a <mask>.	He is more likely to be a <mask>.	Is he more likely to be a surgeon or a doctor? He is more likely to be a <mask>.	Most of the time, children are unaware of their career plans. Joe loves cutting things open and understanding the inner workings of the world around him. As a young child, his mother often caught him sewing holes in furniture. Is he more likely to be a surgeon or a doctor? He is more likely to be a <mask>.
4	Sally enjoys bringing characters to life and exploring dramatic story arcs. She has read biographies since a young age and loves stories that describe unfortunate upbringings. Is she more likely to be a writer or a novelist? She is more likely to be a <mask>.	She is more likely to be a <mask>.	Is she more likely to be a writer or a novelist? She is more likely to be a <mask>.	Basic literacy is a prerequisite to understanding other subjects. Sally enjoys bringing characters to life and exploring dramatic story arcs. She has read biographies since a young age and loves stories that describe unfortunate upbringings. Is she more likely to be a writer or a novelist? She is more likely to be a <mask>.
5	Elliot is a picky eater. However, he enjoys pink seafood and will go to great lengths to find the highest quality he can. He dreams of living in Alaska and catching his own. During his next meal, is he more likely to have salmon or fish? He is more likely to have <mask>.	He is more likely to have <mask>.	During his next meal, is he more likely to have salmon or fish? He is more likely to have <mask>.	Food is necessary to survive. Elliot is a picky eater. However, he enjoys pink seafood and will go to great lengths to find the highest quality he can. He dreams of living in Alaska and catching his own. During his next meal, is he more likely to have salmon or fish? He is more likely to have <mask>.
6	Amy is an 8-year old enrolled in elementary school. Many kids at her age are technologically adept, working with tablets and computers daily at their school. Her parents recently decided she was old enough to start calling her friends on her own device. Is she more likely to have a phone or a smartphone? She is more likely to have a <mask>.	She is more likely to have a <mask>.	Is she more likely to have a phone or a smartphone? She is more likely to have a <mask>.	Paper can be used for many tasks. Amy is an 8-year old enrolled in elementary school. Many kids at her age are technologically adept, working with tablets and computers daily at their school. Her parents recently decided she was old enough to start calling her friends on her own device. Is she more likely to have a phone or a smartphone? She is more likely to have a <mask>.
7	Chad is interested in the finer things in life. He wants to become as wealthy as possible and will do whatever it takes. In college, is he more likely to study banking or business? He is more likely to study <mask>.	He is more likely to study <mask>.	In college, is he more likely to study banking or business? He is more likely to study <mask>.	Money is necessary to purchase most goods. Chad is interested in the finer things in life. He wants to become as wealthy as possible and will do whatever it takes. In college, is he more likely to study banking or business? He is more likely to study <mask>.
8	Jennifer is playing cards with a friend. She caught a glimpse of a bright color and a pointy four-sided shape. Is the card she saw more likely to be a red card or a diamond card? She is more likely to have seen a <mask> card.	She is more likely to have seen a <mask> card.	Is the card she saw more likely to be a red card or a diamond card? She is more likely to have seen a <mask> card.	Games are played to pass the time. Jennifer is playing cards with a friend. She caught a glimpse of a bright color and a pointy four-sided shape. Is the card she saw more likely to be a red card or a diamond card? She is more likely to have seen a <mask> card.
9	Patel has a significant amount of experience cooking. He is accustomed to hot food because his mother often used chili peppers in her cooking. When cooking, is he more likely to use spices or cayenne? He is more likely to use <mask>.	He is more likely to use <mask>.	When cooking, is he more likely to use spices or cayenne? He is more likely to use <mask>.	Not everyone is able to make food for themselves. Patel has a significant amount of experience cooking. He is accustomed to hot food because his mother often used chili peppers in her cooking. When cooking, is he more likely to use spices or cayenne? He is more likely to use <mask>.

Table 8: List of All Questions (continued)

10	There are certain cuisines that use a wide variety of spices. Some are more difficult than others. Are people more likely to cook Asian food or Indian food? They are more likely to cook <mask> food.	They are more likely to cook <mask> food.	They are more likely to cook <mask> food.	Salt and pepper are found in almost every pantry. There are certain cuisines that use a wide variety of spices. Some are more difficult than others. Are people more likely to cook Asian food or Indian food? They are more likely to cook <mask> food.
11	Tim has always had a sweet tooth. When he would go out for a run, he would bring a bottle of his favorite beverage every time. Is he more likely to have brought a soda or a drink? He is more likely to have brought a <mask>.	He is more likely to have brought a <mask>.	Is he more likely to have brought a soda or a drink? He is more likely to have brought a <mask>.	Dentists clean teeth. Tim has always had a sweet tooth. When he would go out for a run, he would bring a bottle of his favorite beverage every time. Is he more likely to have brought a soda or a drink? He is more likely to have brought a <mask>.
12	Sarah just finished eating, and she just opened the freezer to look for something sweet. Is she more likely to eat a popsicle or a dessert? She is more likely to eat a <mask>.	She is more likely to eat a <mask>.	Is she more likely to eat a popsicle or a dessert? She is more likely to eat a <mask>.	Without refrigeration, many foods spoil quickly. Sarah just finished eating, and she just opened the freezer to look for something sweet. Is she more likely to eat a popsicle or a dessert? She is more likely to eat a <mask>.
13	Jacqués loves baguettes and croissants, especially those made by his mother. He studied in Paris to become a baker, but wanted a profession that was more stable. Is he more likely to be European or French? He is more likely to be <mask>.	He is more likely to be <mask>.	Is he more likely to be European or French? He is more likely to be <mask>.	Bread requires wheat to make. Jacqués loves baguettes and croissants, especially those made by his mother. He studied in Paris to become a baker, but wanted a profession that was more stable. Is he more likely to be European or French? He is more likely to be <mask>.
14	Jerry is eating his meatball dinner with a fork. He doesn't have any Italian lineage, but his mother has cooked many Italian dishes for him. Is he more likely to eat the meatballs with pasta or spaghetti? He is more likely to eat the meatballs with <mask>.	He is more likely to eat the meatballs with <mask>.	Is he more likely to eat the meatballs with pasta or spaghetti? He is more likely to eat the meatballs with <mask>.	Some people decide not to eat meat and instead follow a vegan lifestyle. Jerry is eating his meatball dinner with a fork. He doesn't have any Italian lineage, but his mother has cooked many Italian dishes for him. Is he more likely to eat the meatballs with pasta or spaghetti? He is more likely to eat the meatballs with <mask>.
15	John is going to his best friend's wedding. He hasn't cared about his looks since leaving college because he learned to focus on functionality over fashion. Is he more likely to wear slacks or pants? He is more likely to wear <mask>.	He is more likely to wear <mask>.	Is he more likely to wear slacks or pants? He is more likely to wear <mask>.	Photographers can find work anywhere. John is going to his best friend's wedding. He hasn't cared about his looks since leaving college because he learned to focus on functionality over fashion. Is he more likely to wear slacks or pants? He is more likely to wear <mask>.
16	Sasha is attending senior prom and wants to look her best. She usually wears sandals day-to-day. Is she more likely to wear shoes or heels? She is more likely to wear <mask>.	She is more likely to wear <mask>.	Is she more likely to wear shoes or heels? She is more likely to wear <mask>.	Different events often have different dress codes. Sasha is attending senior prom and wants to look her best. She usually wears sandals day-to-day. Is she more likely to wear shoes or heels? She is more likely to wear <mask>.
17	Allyson hasn't been able to afford much throughout her life. The only thing she could wear besides shirts, pants, and shoes was her mother's necklace. Once she got her first job, is she more likely to wear makeup or lipstick? She is more likely to wear <mask>.	She is more likely to wear <mask>.	Once she got her first job, is she more likely to wear makeup or lipstick? She is more likely to wear <mask>.	Money can be earned at a job. Allyson hasn't been able to afford much throughout her life. The only thing she could wear besides shirts, pants, and shoes was her mother's necklace. Once she got her first job, is she more likely to wear makeup or lipstick? She is more likely to wear <mask>.
18	Different students have difficulty with different subjects. However, there are subjects that many students alike have difficulty with. Do more people struggle with calculus or math? More people struggle with <mask>.	More people struggle with <mask>.	Do more people struggle with calculus or math? More people struggle with <mask>.	People learn many things in school. Different students have difficulty with different subjects. However, there are subjects that many students alike have difficulty with. Do more people struggle with calculus or math? More people struggle with <mask>.
19	As an entrepreneur, Henry doesn't have time to eat very often. He is considering taking pills to offset his poor diet. Is he more likely to take supplements or multivitamins? He is more likely to take <mask>.	He is more likely to take <mask>.	Is he more likely to take supplements or multivitamins? He is more likely to take <mask>.	People may leave their jobs in search of better working conditions. As an entrepreneur, Henry doesn't have time to eat very often. He is considering taking pills to offset his poor diet. Is he more likely to take supplements or multivitamins? He is more likely to take <mask>.
20	Cassie has been in love with animals from a young age. She owned a tiger plush toy and always visited tigers at the zoo. She also wanted to visit her local animal shelter and play with the kittens there. Is she more likely to have a cat or a pet? She is more likely to have a <mask>.	She is more likely to have a <mask>.	Is she more likely to have a cat or a pet? She is more likely to have a <mask>.	Animals can be dangerous or useful. Cassie has been in love with animals from a young age. She owned a tiger plush toy and always visited tigers at the zoo. She also wanted to visit her local animal shelter and play with the kittens there. Is she more likely to have a cat or a pet? She is more likely to have a <mask>.

Table 9: Decoy Questions

Vehicular accidents can be fatal. Do more people die in motorcycle or car accidents? More people are likely to die in <mask> accidents.
Do more people die as a result of suicide or homicide? More people die as a result of <mask>.
Are people more likely to be killed working as a cop or a fisherman? A person is more likely to die working as a <mask>.
Are people more likely to die as a result of murder or diabetes? People are more likely to die from <mask>.
Are people more likely to die from an attack by a dog or a shark? People are more likely to die from an attack by a <mask>.
A young student grew up in a family of doctors and has always been inspired by their relatives' work. The student has never felt pressure to conform to any social ideals. Is the student more likely to be male or female? The student is more likely to be <mask>.
In the United States, nursing majors must go through rigorous training programs before being able to enter the workforce. They must also not feel queasy at the sight of blood. Are nurses more likely to be male or female? Nurses are more likely to be <mask>.
Clinical psychologists, although employing different methods, often try to understand their patients' problems through talking. Some try to form a relationship with the client before offering criticism to make it more comfortable for the individual. Are clinical psychologists more likely to be male or female? They are more likely to be <mask>.
Construction work is often physically difficult and requires long hours. Are construction workers more likely to be male or female? They are more likely to be <mask>.
Engineers are known to have strong mathematics skills. They may also need to be able to multitask. Are there more male or female engineers? They are more <mask> engineers.

Table 10: Comparison of RoBERTa versus Participants

No.	**Correct Answer** $P(A)$	**Incorrect Answer** $P(A \wedge B)$	**Was the Correct Answer Chosen by RoBERTa?**	**Was the Correct Answer Chosen by** $\geq 50\,\%$ **of Participants ?**	**Number of Participants Correct (Out of 53)**	**Proportion of Participants Correct**
1*	Michigan	Detroit	FALSE	TRUE	33	0.62
2	STEM	Math	TRUE	TRUE	34	0.64
3*	Doctor	Surgeon	FALSE	FALSE	20	0.38
4*	Writer	Novelist	FALSE	TRUE	32	0.60
5*	Fish	Salmon	FALSE	FALSE	21	0.40
6*	Phone	Smartphone	FALSE	FALSE	26	0.49
7	Business	Banking	TRUE	TRUE	42	0.79
8	Red	Diamond	TRUE	TRUE	40	0.75
9	Spices	Cayenne	TRUE	TRUE	38	0.72
10	Asian	Indian	TRUE	TRUE	47	0.89
11	Drink	Soda	TRUE	TRUE	40	0.75
12	Dessert	Popsicle	TRUE	TRUE	42	0.79
13	European	French	TRUE	TRUE	30	0.57
14	Pasta	Spaghetti	TRUE	TRUE	29	0.55
15	Pants	Slacks	TRUE	TRUE	41	0.77
16*	Shoes	Heels	TRUE	FALSE	17	0.32
17*	Makeup	Lipstick	FALSE	TRUE	35	0.66
18	Math	Calculus	TRUE	TRUE	37	0.70
19	Supplements	Multivitamins	TRUE	TRUE	36	0.68
20	Pet	Cat	TRUE	TRUE	28	0.53

** Indicates both RoBERTa and participants had either incorrect or different answers*

Table 11: Comparison of RoBERTa versus Participants with Upper and Lower Error Bounds

No.	Was the Correct Answer Chosen by RoBERTa ?	Was the Correct Answer Chosen by $\geq 50\%$ of Participants ?	Number of Participants Correct (Out of 53)	Proportion of Participants Correct	Proportion of Participants Correct (Lower Bound -13.5%)	Proportion of Participants Correct (Upper Bound +13.5%)
1	FALSE	TRUE	33	0.62	0.54	0.70
2	TRUE	TRUE	34	0.64	0.55	0.73
3	FALSE	FALSE	20	0.38	0.33	0.43
4	FALSE	TRUE	32	0.60	0.52	0.68
5	FALSE	FALSE	21	0.40	0.35	0.45
6**	FALSE	FALSE	26	0.49	0.42	0.56
7	TRUE	TRUE	42	0.79	0.68	0.90
8	TRUE	TRUE	40	0.75	0.65	0.85
9	TRUE	TRUE	38	0.72	0.62	0.82
10	TRUE	TRUE	47	0.89	0.77	1.01
11	TRUE	TRUE	40	0.75	0.65	0.85
12	TRUE	TRUE	42	0.79	0.68	0.90
13*	TRUE	TRUE	30	0.57	0.49	0.65
14	TRUE	TRUE	29	0.55	0.48	0.62
15	TRUE	TRUE	41	0.77	0.67	0.87
16	TRUE	FALSE	17	0.32	0.28	0.36
17	FALSE	TRUE	35	0.66	0.57	0.75
18	TRUE	TRUE	37	0.70	0.61	0.79
19	TRUE	TRUE	36	0.68	0.59	0.77
20*	TRUE	TRUE	28	0.53	0.46	0.60

** Indicates answer is false at the lower bound of the confidence interval*

*** Indicates answer is true at the upper bound of the confidence interval*

Table 12: Individual Survey Participant Scores

Individual No.	Raw Score (out of 20)	Proportion Correct
1	12	0.60
2	4	0.20
3	7	0.35
4	10	0.50
5	9	0.45
6	19	0.95
7	17	0.85
8	5	0.25
9	14	0.70
10	13	0.65
11	11	0.55
12	11	0.55
13	20	1.00
14	8	0.40
15	17	0.85
16	10	0.50
17	9	0.45
18	10	0.50
19	10	0.50
20	8	0.40
21	20	1.00
22	19	0.95
23	8	0.40
24	12	0.60
25	4	0.20
26	10	0.50

Table 12: Individual Survey Participant Scores (continued)		
27	19	0.95
28	10	0.50
29	20	1.00
30	20	1.00
31	9	0.45
32	17	0.85
33	20	1.00
34	12	0.60
35	6	0.30
36	12	0.60
37	20	1.00
38	8	0.40
39	19	0.95
40	13	0.65
41	20	1.00
42	20	1.00
43	10	0.50
44	8	0.40
45	19	0.95
46	10	0.50
47	7	0.35
48	19	0.95
49	18	0.90
50	3	0.15
51	7	0.35
52	10	0.50
53	15	0.75

Table 13: Change in RoBERTa's Output with Greater and Fewer Details

No.	Was the Original Question Answered Correctly by RoBERTa?	Did Removing All Details Change RoBERTa's Answer?	Did Removing Some Details Change RoBERTa's Answer?	Did Adding Some Details Change RoBERTa's Answer?
1	FALSE	FALSE	FALSE	FALSE
2*	TRUE	TRUE*	FALSE	FALSE
3*	FALSE	TRUE*	TRUE*	FALSE
4*	FALSE	TRUE*	TRUE*	TRUE*
5*	FALSE	TRUE*	TRUE*	FALSE
6*	FALSE	FALSE	FALSE	TRUE*
7	TRUE	FALSE	FALSE	FALSE
8	TRUE	FALSE	FALSE	FALSE
9	TRUE	FALSE	FALSE	FALSE
10	TRUE	FALSE	FALSE	FALSE
11	TRUE	FALSE	FALSE	FALSE
12	TRUE	FALSE	FALSE	FALSE
13*	TRUE	TRUE*	FALSE	FALSE
14	TRUE	FALSE	FALSE	FALSE
15	TRUE	FALSE	FALSE	FALSE
16*	TRUE	TRUE*	TRUE*	FALSE
17*	FALSE	TRUE*	FALSE	FALSE
18	TRUE	FALSE	FALSE	FALSE
19	TRUE	FALSE	FALSE	FALSE
20	TRUE	FALSE	FALSE	FALSE

5 Discussion

5.1 Results

The model scored 14/20 questions correctly, and survey participants, in aggregate, scored 16/20 questions correctly. In the worst case scenario (lower error bound), the participants, in aggregate, scored 14/20, the same as RoBERTa. In the best case scenario (upper error bound), participants, in aggregate, scored 17/20.

If we compare the scores individually, however, we obtain different results. Individually, the participants scored a mean $\bar{x}$ of 0.63, approximately 13/20 questions correct. Given a standard deviation s of 0.151 and a margin of error of ±10.54%, the individuals scored, in the worst case, 11/20 questions. At best, they scored 14/20, equal to RoBERTa.

Table 14: Summary of Results

Sample Type	**Margin of Error** $95\,\%\,CI$	**Lower Bound** $\mu - (\%error)$	**Mean** μ	**Upper Bound** $\mu + (\%error)$	**Number of Questions Correct**
Aggregate	±13.5%	0.67	0.80	0.94	[13, 19]
Individual	±10.54%	0.56	0.63	0.70	[11, 14]

5.1.1 Hypothesis Testing

If we perform a two-tailed hypothesis test on the individual data with parameters:

$$H_0 : \mu = 0.70$$
$$H_1 : \mu \neq 0.70$$
$$\bar{x} = 0.63$$
$$s = 0.151$$
$$n = 53$$

We obtain:

$$z = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{0.63 - 0.70}{\frac{0.151}{\sqrt{53}}} = -3.375$$

At a 95% confidence interval, $z = -3.375$ corresponds to a p-value of 0.0007. Thus, we reject the null hypothesis that the mean score of the individuals is equal to RoBERTa's.

If we perform a one-tailed hypothesis test on the individual data with parameters:

$H_0 : \mu \geq 0.70$

$H_1 : \mu < 0.70$

$\bar{x} = 0.63$

$s = 0.151$

$n = 53$

We obtain:

$$z = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{0.63 - 0.70}{\frac{0.151}{\sqrt{53}}} = -3.375$$

At the 95% confidence interval, $z = -3.375$ corresponds to a p-value of 0.0004. Thus, there is enough evidence to reject the null hypothesis that individual participants performed better than RoBERTa on average.

On average, RoBERTa did perform better. However, if we consider that these questions were generated without an external auditor, it may be worthwhile to discard outlier questions. One question in particular, shoes vs. heels, was answered correctly by RoBERTa in which an overwhelming majority of participants answered incorrectly. This question may not have been well-formed because shoes and heels may be considered two distinct objects rather than heels as a subset of shoes. If we discard this outlier question, we have less evidence that RoBERTa performs better than humans, but we obtain a more valid result.

For a one-tailed hypothesis test, consider if RoBERTa scored from 14/20 (0.7) to 13/19 (0.684) and the participants averaged a score from 12.6/20 (0.63) to 11.67/19 (0.614):

$H_0 : \mu \geq 0.684$

$H_1 : \mu < 0.684$

$\bar{x} = 0.614$

$s = 0.246$

$n = 53$

We obtain:

$$z = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{0.614 - 0.684}{\frac{0.246}{\sqrt{53}}} = -2.072$$

At the 95% confidence interval, $z = -2.072$ corresponds to a p-value of 0.02. Thus, even with the removal of the outlier question, there is enough evidence to support that RoBERTa performs, on average, better than the human participants.

5.1.2 RoBERTa's Results

If we were to insert a score of 14/20 into a binomial distribution with the following parameters:

number of trials n: 5
number of successes x: 14
probability of success p: 0.5
probability of failure q: 0.5

the likelihood of a score of 14/20 due to random chance is $P(X \geq 14) = 0.0576$, approximately 6%. Thus, we know that RoBERTa is, at minimum, better than a coin-flip guess.

There were several individuals who performed worse than RobERTa, scoring merely 4/20. It seems that, in the case that an individual were more prone to the conjunction fallacy, RoBERTa may be useful, offering up to a 3.5x improvement on 4/20. Most interestingly, it seems RoBERTa failed on the same questions as the individuals in three instances.

The control questions were intended to show that RoBERTa was not choosing its answers because it was biased to a particular answer from the start. Some answers differed when removing all context, which was ideal. If the answers were all the same, it would be evidence that RoBERTa had a bias from the start.

Removing some context was designed to test Kahneman's (2011) findings that the individuals performed better when details did not serve as distractors. His findings were somewhat replicated here because with less context, the model answered three additional questions correctly. However, the opposite was also true; adding greater context led RoBERTa to answer two additional questions correctly.

5.2 Sources of Error

There are several sources of error which must be mentioned when interpreting these results. Most obviously, the sample size is quite small, both in number of questions (20) and number of participants (53). The participants were also recruited at Emory University, so it is highly unlikely that these findings are generalizable. However, it does seem to replicate Kahneman's findings because there was a substantial number of individuals that chose the incorrect answer, as he found. Unlike Kahneman's findings, however, there were no instances where up to 90% of participants answered a question incorrectly.

Kahneman's critics mention that participants can misinterpret "possible" as "plausible" when presented with questions based on the conjunction fallacy. Although the conjunction fallacy itself

is not language-dependent, the presentation of the questions may have influenced the answer as well. As mentioned, "heels" may be viewed as distinct from "shoes," despite dictionary definitions stating that heels are a type of shoe. This experiment relies on participants believing that one of the answer choices is a subset of the other. If they believe that the two answers are different objects, the question does not accurately capture the conjunction fallacy. This pitfall is worth noting because the questions are non-standardized, unlike the questions in the RACE benchmark. The RACE questions are far larger in quantity and are quality-controlled well enough to be used for standardized testing in China. The questions used in this experiment were generated without any external auditor to verify whether the questions were well-formed.

Lastly, although there was an attempt to minimize experimenter demand effects via decoy questions, it is likely that being presented with many questions of the same type could have influenced the participants.

6 Conclusion

While the results of this study did not find that RoBERTa performed better compared to every individual, nor did it perform better compared to the consensus, RoBERTa did perform better than the individuals on average. Even if we remove an outlier question answered correctly by RoBERTa and incorrectly by the participants, we still find that RoBERTa performs better. In fact, some respondents performed 3.5x worse than RoBERTa, suggesting that these individuals prone to the conjunction fallacy may find some benefit from using RoBERTa.

There were individuals, however, who received a perfect score. If we introduce this model to each participant and it casts doubt on these perfect scorers, it will reduce their scores closer to the average, yet it will also raise the low-scorers closer to the average. This suggests that we will require a cost-benefit analysis to judge whether there is benefit in reducing the variance of answers. In short, higher risk is higher reward, but risk tolerance varies across situations.

At best, RoBERTa is able to raise scores to the upper error bound of humans. RoBERTa was incapable of a perfect score as some individuals were. Knowing RoBERTa performed better on average, but with far less variance, we must evaluate our preference for variance. Fortunately, state-of-the-art models are likely to fare better and increase this upper limit. Future research using other NLP models may choose to focus on similar tests; a greater variety of tests would more accurately gauge whether there is potential for using NLP models as external aids to decision-making.

6.1 Ethical Concerns

In the spirit of GPT-3 including a section on ethical concerns in their whitepaper, there is one concern in particular that is worth mentioning in the context of this research. If we are able to model cognitive biases in NLP models, we may be able to engineer questions that fit a particular narrative. It is plausible that questions could be reworded until the answers fit to our liking.

If we know the model is accurate enough, we then know how human participants will answer. Leading questions are already designed this way, but could become more covert if they are constructed on a sophisticated model. An example of an explicit leading question for abortion could be, "Do you support killing babies," versus "do you support a woman's right to her own body?" These are obvious enough to be spotted by some, but a model may allow an examiner to create more inconspicuous questions to obtain desired results. Using a model in this way may be disastrous for polls and surveys, particularly in this sensitive time when many people are already skeptical of the news and medical authorities.

7 Bibliography

Albahli, Saleh and Albattah, Waleed. ‘Detection of Coronavirus Disease from X-ray Images Using Deep Learning and Transfer Learning Algorithms’. 1 Jan. 2020 : 841 – 850. doi: 10.3233/XST-200720

Alec Radford, Karthik Narasimhan, Time Salimans,and Ilya Sutskever. 2018. Improving language un-derstanding with unsupervised learning. Technicalreport, OpenAI.

Baddeley, M. (2019). Behavioral economics: Past, present, and future. *MIT Technology Review*. https://www.technologyreview.com/2019/10/10/65182/behavioral-economics-past-present-and-future/

Bhavsar, Pratik. OpenAI GPT3 Profit Margins. *Pratik's Pakodas*, Substack, 23 Aug. 2020, pakodas.substack.com/p/estimating-gpt3-api-cost.

Bond, R., & Smith, P. B. (1996). Culture and conformity: A meta-analysis of studies using Asch's (1952b, 1956) line judgment task. *Psychological bulletin*, 119(1), 111.

Brogaard, Berit. 'Linda the Bank Teller' Case Revisited. *Psychology Today*, 22 Nov. 2016, www.psychologytoday.com/us/blog/the-superhuman-mind/201611/linda-the-bank-teller-case-revisited.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language Models are Few-Shot Learners. *arXiv preprint*. https://arxiv.org/abs/2005.14165

Corcoran, C. M., Mittal, V. A., Bearden, C. E., Gur, R. E., Hitczenko, K., Bilgrami, Z., ... & Wolff, P. (2020). Language as a biomarker for psychosis: A natural language processing approach. *Schizophrenia research*, *226*, 158-166.

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*.

Dougherty MR, Gettys CF, Ogden EE (1999). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review*. 106 (1): 180–209. doi:10.1037/0033-295x.106.1.180

Gigerenzer G (2006). "Bounded and Rational". In Stainton RJ (ed.). *Contemporary Debates in Cognitive Science*. Blackwell. p. 129. ISBN 978-1-4051-1304-5.

Magnus, P. D. f*orallX: an Introduction to Formal Logic*. Open SUNY Textbooks, 2017.

Fujiyoshi, H., Hirakawa, T., & Yamashita, T. (2019). Deep learning-based image recognition for autonomous driving. *IATSS research*, *43*(4), 244-252.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining.arXiv preprintarXiv:1901.07291.

Hao, Karen. *What Is Machine Learning?* MIT Technology Review, 17 Apr. 2018, www.technologyreview.com/2018/11/17/103781/what-is-machine-learning-we-drew-you-another-flowchart/.

He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, *349*(6245), 261-266.

Hsee, C.K. (1998), Less is better: when low−value options are valued more highly than high−value options. *J. Behav. Decis. Making*, 11: 107-121. https://doi.org/10.1002/(SICI)1099-0771(199806)11:2<107::AID-BDM292>3.0.CO;2-Y

Matthew Peters, Mark Neumann, Mohit Iyyer, MattGardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *North American Association for Com-putational Linguistics (NAACL).*

Kahneman, D. (2011). *Thinking, fast and slow.* Farrar, Straus and Giroux.

Kapoor, I., & Mishra, A. (2018). Automated Classification Method for Early Diagnosis of Alopecia Using Machine Learning. *Procedia Computer Science*, *132*, 437-443.

Khan, Suleiman. BERT, RoBERTa, DistilBERT, XLNet - Which One to Use? *Medium*, Towards Data Science, 4 Sept. 2019, towardsdatascience.com/bert-roberta-distilbert-xlnet-which-one-to-use-3d5ab82ba5f8.

Lai, G., Xie, Q., Liu, H., Yang, Y., & Hovy, E. (2017). Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Training Approach. *arXiv preprint*. https://arxiv.org/pdf/1907.11692.pdf

List, John, A. 2002. "Preference Reversals of a Different Kind: The "More Is Less" Phenomenon ." American Economic Review, 92 (5): 1636-1643.

Marr, Bernard. What Is GPT-3 And Why Is It Revolutionizing Artificial Intelligence? *Forbes*, Forbes Magazine, 5 Oct. 2020, www.forbes.com/sites/bernardmarr/2020/10/05/what-is-gpt-3-and-why-is-it-revolutionizing-artificial-intelligence/.

Nayak, Pandu. *Understanding Searches Better than Ever Before*. Google, 25 Oct. 2019, www.blog.google/products/search/search-language-understanding-bert/.

Raghu, M., Blumer, K., Sayres, R., Obermeyer, Z., Kleinberg, B., Mullainathan, S., & Kleinberg, J. (2019). *Direct Uncertainty Prediction for Medical Second Opinions* Proceedings of the 36th International Conference on Machine Learning, Proceedings of Machine Learning Research. http://proceedings.mlr.press/v97/raghu19a.html

Rajpurkar, P., Jia, R., & Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. *arXiv preprint arXiv:1806.03822*.

Sagar, Ram. OpenAI Releases GPT-3, The Largest Model So Far. *Analytics India Magazine*, 6 Mar. 2020, analyticsindiamag.com/open-ai-gpt-3-language-model/.

Slovikovskaya, V. (2019). Fake News Detection Powered with BERT and Friends. *Medium*. https://medium.com/@vslovik/fake-news-detection-empowered-with-bert-and-friends-20397f7e1675

Scheible, Raphael. Pytorch/Fairseq. *GitHub*, 26 July 2019, github.com/pytorch/fairseq/blob/master/examples/roberta/README.md.

Subudhi, Krishan. Type Faster Using RoBERTA. *Krishan's Tech Blog*, 10 June 2020, krishansubudhi.github.io/deeplearning/2020/06/10/smart-autocorrect-RoBERTa.html.

Thorstad, R., & Wolff, P. (2019). Predicting future mental illness from social media: A big-data approach. *Behavior research methods*, *51*(4), 1586-1600.

Trieu H Trinh and Quoc V Le. 2018. A simple method for commonsense reasoning.*arXiv preprint* arXiv:1806.02847.

Tversky, Amos; Kahneman, Daniel (October 1983). Extension versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*. 90 (4): 293–315. doi:10.1037/0033-295X.90.4.293

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). *Glue: A multi-task benchmark and analysis platform for natural language understanding*. Paper presented at 7th International Conference on Learning Representations, ICLR 2019, New Orleans, United States.

Wang, C., Liu, X., & Song, D. (2020). Language Models are Open Knowledge Graphs. *arXiv preprint*. https://arxiv.org/pdf/2010.11967v1.pdf

Wolff, Rachel. What Is Natural Language Processing & How Does It Work? *MonkeyLearn Blog*, 26 Feb. 2020, monkeylearn.com/blog/what-is-natural-language-processing/.

Yang, C. C., Chen, H., & Hong, K. (2003). Visualization of large category map for Internet browsing. *Decision support systems*, *35*(1), 89-102.

Yse, Diego Lopez. "Your Guide to Natural Language Processing (NLP)." *Medium*, Towards Data Science, 15 Jan. 2019, towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le.2019. Xlnet: Generalized autoregressive pretraining for language understanding.*arXiv preprint* arXiv:1906.08237.

8 Appendix

Proof 1: Advanced Implicit Conditional Example

$L \rightarrow (N \vee E)$	given
$E \rightarrow B$	
L	
$\neg B$	
$\neg E$	$Modus\ Tollens(2, 4)$
$(N \vee E)$	$\rightarrow Elimination(1, 3)$
N	$\vee Elimination(5, 6)$

Proof 2: Advanced Set Theory Example

$\forall x(\neg Sx \wedge Kx \rightarrow \neg Hx)$	given: Sx = *x* is a surgeon Kx = *x* is skilled Hx = *x* is hired by the hospital
$\forall x(Sx \rightarrow Gx)$	given: Gx = *x* is greedy
$(\forall xSx) \wedge (\neg \forall xKx)$	given
$\therefore (\forall xGx) \wedge (\neg \forall xHx)$	conclusion
$\forall xSx$	$\wedge Elimination(3)$
Sc	$\forall Elimination(5)$
$Sc \rightarrow Gc$	$\forall Elimination(2)$
Gc	$\rightarrow Elimination(6, 7)$
$\forall xGx$	$\forall Introduction(8)$