A THEORETICAL & EMPIRICAL ANALYSIS OF TRANSFORMER LANGUAGE MODEL BEHAVIOR

By

Jesse Taylor Noah Roberts

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Computer Science

August 9, 2024

Nashville, Tennessee

Approved:

Dr. Douglas H. Fisher, Ph.D.

Dr. Gautam Biswas, Ph.D.

Dr. Jesse Spencer-Smith, Ph.D.

Dr. Jules White, Ph.D.

This Dissertation is dedicated to Lindsey. She encouraged me to take a leap. If not for her, I would never have started this journey, to say nothing of finishing it.
Throughout, she listened to me drone on about research, she came up with ideas, and, together, we dreamed.
It's been fun. Let's never do this again.

ACKNOWLEDGMENTS

I would like to thank:

My advisor, Dr. Doug Fisher for being patient and allowing me to explore topics far and wide, and encouraging me to corral my focus. Additionally, he has been an insightful and helpful co-author. He supported my work through the National Science Foundation Grant No. 1521672.

Kyle Moore and Drew Wilenzick as excellent co-authors. Their help has been invaluable.

My committee members: Dr. Biswas Gautam, Dr. Jesse Spencer-Smith, Dr. Janos Stipanovitz, and Dr. Jules White for their helpful comments and guidance.

A number of people who were instrumental in my starting this Phd. Dr. Doug Talbert advised me and connected me with Dr. Fisher, Dr. Maithilee Kunda advised and supported my admission, and Dr. Akos Ledeczi found a place for me.

Dr. Pat Langley, who's comments have helped me to develop a more systematic understanding of research. Two that have been of untold value are: "First, choose your claims and then develop experiments to support those claims." and "Anything worth doing is worth doing badly." Technically the latter is from GK Chesterton, but Dr. Langley repurposed it to convey, done is better than perfect.

Dr. Ray and Katrina who have been consistently supportive for as long as I can remember, and who proved that it was indeed possible to finish a Phd.

My parents for always being willing to take care of Asa when we had to travel.

Alice Reed (Lindsey's Gran) for being a great example of success.

Wade and Emily for being ever present and loyal best friends and adventurers.

Colonel Jim Kennedy and Dr. Jim Peace for providing an endless source of conversation topics and challenging me (by their own example) to live an interesting life.

Asa for being a wonderfully behaved and patient little-son (so long as he's not hungry).

Poet's Coffee, Soul Craft, Mookies on Mugford, Plenty Books, Jamie's Sweets and Eats, the Dartmoor Moorland Hotel at Haytor, Helen's Tower in County Down, the Tennessee Tech Lab Science Center, the Vanderbilt Divinity Library, the Boston Public Library, the Harvard Science Center Cafe, the Yale Harkness clock tower, and the Dartmouth Green for providing beautiful surroundings fit for writing and inspiration.

Plutarch, Ben Franklin, Arthur Conan Doyle, Voltaire, Patrick O'Donnell, Cade Metz, Adam Smith, Charles Dickens, Carlene O'Connor, Mary Shelley, Stephen King, Washington Irving, Henry Wadsworth Longfellow, Oliver Wendell Holmes, William Cullen Bryant, James Russell Lowell, Marilynne K. Roach, William Butler Yeats, James Joyce, John Sedgwick, Elias Boudinot, Brian Christian, Andrzej Sapkowski, David French, A. W. Jantha, Washington Irving, Oscar Wilde, Adriana Mather, Delia Owens, Agatha Christie, Earl Hamner, Walter Isaacson, Robert Louis Stevenson, Ernest Cline, Shirley Jackson, H. P. Lovecraft, Frank W. Abagnale, Kevin Mitnick, Michael Pollan, Ayn Rand, Mark Twain, Malcolm Gladwell, Richard Powers, Bram Stoker, Nathaniel Hawthorne, C. S. Lewis, J. R. R. Tolkien, Alan Hirshfeld, Archimedes, James Gleick, Ablert Einstein, and Isaac Newton for providing literary and intellectual inspiration (beyond explicit citation) during my journey through the program.

The sovereign God, from whom all good things come.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction and Scope

Transformers are a powerful class of neural networks well suited to applications like language processing which involve context-dependent sequence processing (Vaswani et al., 2017). These models are relatively new, having been proposed in 2017, and much is still unknown about their fundamental and behavioral capabilities. The aims of this dissertation are (1) to better understand the fundamental nature of this class of models and (2) to better understand the complex behaviors that large language models (LLMs) learn to exhibit from training on human language data. The former is accomplished by applying methods of theoretical analysis intrinsic to computer science, while the latter is accomplished by adapting and applying social science techniques, uncertainty estimation, and data science. In Figure 1.1, this relationship is visually summarized.

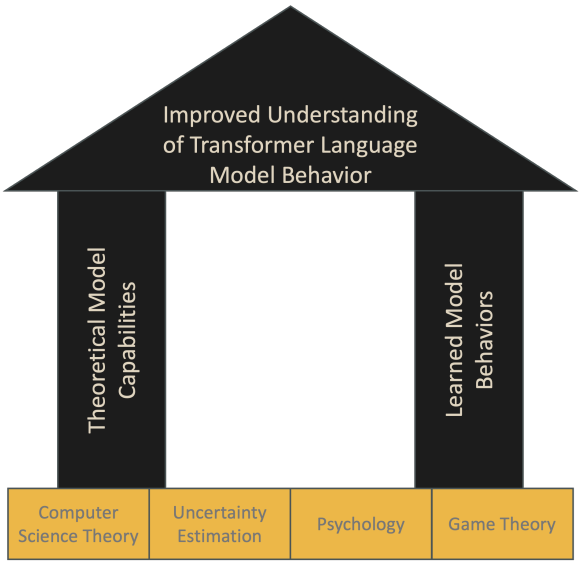

Figure 1.1: The understanding of transformer-based LLMs is advanced by studying both their theoretical and empirical behaviors, leveraging both computer and social science methods.

1.1 Why Employ Both a Theoretical and Empirical Approach?

It's reasonable to question why, to understand the behavior of LLMs, it is useful and important to approach them from both theoretical and empirical frames.

Understanding the theoretical capabilities of algorithms is fundamental to the field of computer science. The work of Gödel (Gödel, 1962), Turing (Turing, 1937), and Church (Church, 1940) founded the field of computer science on theoretic mathematical proofs of the computing capabilities of machines. Following in

their footsteps not only preserves that foundation, but is essential to understanding what is *possible* as opposed to what is presently *practicable*. Theoretical work of this variety has yielded important insight into limitations of language model self-attention (Hahn, 2020) and base transformer model computational expressivity (Pérez et al., 2019).

On the other hand, neural network models are famously opaque (Zhang et al., 2021). So, while theoretical study is indispensable, understanding top level neural model behavior from first principles is a major goal of current research and has thus far been largely unyielding. Due to this, social science methods have increasingly been used to make sense of the high-level behaviors exhibited by neural models (Hagendorff et al., 2023). This dissertation adds to the behavioral characterization of transformer models by adapting uncertainty estimation methods so that they may be used to create artificial populations, facilitating more robust LLM behavioral study.

Table 1.1: Major dissertation theoretical, scholarly, and empirical contributions. This dissertation:

Type	Chapter	Identifier	Contribution
Empirical	2	E.1	Establishes fuzzy-weak domination as a **novel model** of **empirical human strategic behavior** in the Traveler's Dilemma and provides a **hypothesis regarding its provenance**
	3	E.2	Develops a **methodology** for more **robust investigation** of neural **model behavior** based on generating **perturbed populations of base models**
	3	E.3	Applies the developed methodology to the **replication of cognitive studies of LLMs** (Typicality and Structural Priming)
	4	E.4	Applies the developed methodology to **understand LLM strategic behavior** as compared to human-like strategic behavior in the **Prisoner's Dilemma** and the **Traveler's Dilemma**
Theoretical & Scholarly	5	TS.1	Shows that **decoder-only transformer** models are **Turing complete** under reasonable **assumptions**
	5	TS.2	Establishes the vanilla and **decoder-only transformer** models as **causal B machines**
	6	TS.3	Suggests that large **pre-training** regimes may be **consistent with human-like learning** and compares **in-context learning** in LLMs to **human-like learning** with specific attention to identify **open questions for future research**
	7	TS.4	Shows that for, **causal B machines**, passing a **Turing test precludes Turing complete computation**
	7	TS.5	Identifies architectural changes which may permit LLMs to potentially **pass a Turing test without precluding Turing complete computation**
	7	TS.6	Establishes recommendations regarding closed and open source LLMs to **support reproducible LLM research**

1.2 Summary of Contributions

This dissertation contributes to the body of knowledge regarding theoretical and empirical transformer-based LLM behavior, with the major contributions listed in Table 1.1. A brief overview of the dissertation as a unified rhetorical document follows.

The empirical work improves understanding of what is currently *practicable* using human cognitive behavior as a comparison. To characterize LLM cognitive behaviors, a method was developed to apply systematic variations to the model and characterize the robustness of the behaviors under variations (E.2). The method is used to replicate two previous studies in the LLM cognitive studies literature finding that typicality is present among the tested model populations while structural priming is not (E.3). The developed method is then leveraged to study the strategic behavior of LLMs, finding that populations of sufficiently large models trained with sliding window attention are robustly able to evaluate strategic preference based on value and engage in the prisoner's dilemma in a human-like manner (E.4).

To establish human-like behavior in the traveler's dilemma, empirical literature is used and expanded by the introduction of a model of human behavior. As is the hope with any model, it yields novel predictions regarding, in this case, the source of an interesting strategic behavior (E.1). LLM behavior in the traveler's dilemma is found to not only be consistent with empirically established human behavior, but also shows that the mentioned hypothesis regarding human behavior holds among LLMs that robustly prefer strategies based on value (E.4).

This dissertation augments the empirical work by providing a better understanding of what is *possible* for transformer-based LLMs by first establishing that while decoder-only transformers are Turing complete (TS.1), when strong limitations are placed on the content of their output, they are no longer Turing complete (TS.2, TS.4). Further, large data quantity pre-training is argued to not be inconsistent with human-like intelligent behavior (TS.3). Therefore, in the pursuit of artificial general intelligence (AGI), this dissertation advocates for exploration of alternate architectures that do not conflate the interaction and computation spaces (TS.5).

Finally, the theoretical work leads to an important discussion on reproducible future LLM research. From the survey of empirical studies of language model cognitive behaviors in Ch 3 and the work to establish a method for creating systematically perturbed populations of models, it is apparent that models which are released as closed-source artifacts diminish the ability to reproduce research. This dissertation argues that, like privately held fossils, closed-source models are not appropriate targets for scientific research because the results produced are commonly not reproducible as the research target is likely to change and be unavailable long term (TS.6).

1.3 About the Dissertation

This dissertation is separated into primarily empirical research (chapters 2, 3, and 4) and primarily theoretical & scholarly work (chapters 5, 6, and 7) with the major contributions of each detailed in Table 1.1.

Each of the chapters is self-contained, having been first written as a published or in press paper with a targeted contribution to the body of knowledge. Each of these possesses nuanced contributions beyond those listed that serve to illuminate and refine the listed primary findings. In Ch 8, they are assembled into a collective narrative to further the field in a meaningful, if modest, way and offer insight into important LLM architectures, applications, and behaviors for future research.

1.3.1 First Person Plural

In this dissertation the first person plural is typically used to refer to the author. At times, this also refers to co-authors (enumerated in the acknowledgments) of the associated papers which have become chapters herein. For all constituent papers, the author of this dissertation is the principal author.

References

Church, A. (1940). A formulation of the simple theory of types. *The Journal of symbolic logic*, 5(2):56–68.

Gödel, K. (1962). *On formally undecidable propositions of Principia Mathematica and related systems*. Courier Corporation.

Hagendorff, T., Fabi, S., and Kosinski, M. (2023). Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt. *Nature Computational Science*, pages 1–6.

Hahn, M. (2020). Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171.

Pérez, J., Marinković, J., and Barceló, P. (2019). On the turing completeness of modern neural network architectures. *arXiv preprint arXiv:1901.03429*.

Turing, A. M. (1937). On computable numbers, with an application to the entscheidungsproblem. *Proceedings of the London Mathematical Society*, s2-42(1):230–265.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Zhang, Y., Tiňo, P., Leonardis, A., and Tang, K. (2021). A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742.

CHAPTER 2

Finding an Equilibrium in the Traveler's Dilemma with Fuzzy Weak Domination

2.1 Introduction

At the heart of game theory lies a pair of fundamental assumptions. First, it is assumed that each player is rational. That is, they will each make the best decision that they can. The second assumption prescribes that a decision is considered best if it maximizes the player's payoff. The rules of a given game and the rationale available to each player are considered common knowledge. So, given a space of possible strategies, a player will choose the strategy that is a best response to the strategy that they believe their opponent will pick.

If each player believes that their opponent is perfectly rational then an infinite sequence in which player A knows that player B knows that player A knows ... that player B knows some fact, is possible. However, a sequence of this type is so obviously intractable that it exists as a comedic trope (Goldman, 2001). A more recent sub-field called epistemic game theory attempts to formally reason about games given some set of beliefs held by the players. In this way, the players can be assumed to be rational and have common knowledge without requiring that they believe their opponent is rational.

In nature, groups of rational agents tend toward certain strategies in a game. These strategies are referred to as equilibria. While it is guaranteed that every game possesses at least one Nash equilibrium (Nash et al., 1950), a given game may have various other equilibria as well. Mathematically studying the rational analysis that underpins a given equilibria is important as analysis methods often generalize to other games, leading to the explanation or expectation of equilibria behaviour in these games as well.

In this paper we present early work in which we look at a set of results from an experiment involving a one-shot traveler's dilemma game. From an emergent equilibrium we argue that the human participants hold some level of uncertainty regarding their opponent's rationality. We then show that iterated elimination of weakly dominated strategies, which is used to find the Nash equilibrium in the traveler's dilemma, does not converge to the Nash equilibrium if players have non-zero uncertainty regarding opponent rationality. Finally, we present the first formulation of an extension to the idea of weak domination, referred to as fuzzy weak domination, which facilitates equilibrium analysis in the face of uncertainty regarding opponent rationality.

2.1.1 Traveler's Dilemma Paradox

We provide a short introduction to the traveler's dilemma (TD). For a more thorough discussion, see (Basu, 2007).

Suppose there are two people traveling back from vacation. Both of the travelers have purchased the same

antique and have checked the antiques as luggage on the flight home. The airline breaks both antiques. The baggage claim team informs the two people of the broken antique and informs them that another passenger on the plane also had their identical antique broken.

The travelers are each told to give a value for the antique on the range $[2, 100]$, but they are warned that quoting a higher price than the other passenger will result in a penalty. Thus the airline has engaged the passengers in a 2-player game. If the two players provide the same quote, then they will each receive the amount quoted. However, given the quote from player A is Q_A and the quote from player B is Q_B, if $Q_A > Q_B$ then the payoff for player A will be $Q_B - 2$ and the payoff for player B will be $Q_B + 2$. The reciprocal statement is true if $Q_A < Q_B$.

Definition 1. *A strategy is a Nash equilibrium iff no change in strategy can achieve a higher payoff assuming the opponent(s) does not change strategy.*

Partial and Total Ordering by Weak Domination

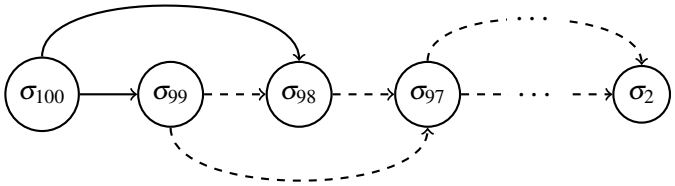

Figure 2.1: An edge directed from vertex *a* to vertex *b* signifies a relative preference for vertex *b*. The solid edges are implied by $\sigma_{99} >_{wd} \sigma_{100}$ (Partial Ordering). While the dashed edges are implied by $\sigma_{98} >_{wd} \sigma_{99}$ etc. after weakly dominated strategies are eliminated due to zero uncertainty regarding opponent rationality (Extending to Total Ordering).

Definition 2. *For two strategies, σ_α is said to weakly dominate σ_β, that is $\sigma_\alpha \underset{wd}{>} \sigma_\beta$, iff for every opponent strategy, σ_α provides a payoff no worse than σ_β and there exists one or more scenarios in which σ_α provides a better payoff (Osborne and Rubinstein, 1994).*

The only Nash equilibrium strategy given in the format (Q_A, Q_B) for the TD is (σ_2, σ_2). It is obvious from definition 1 that (σ_2, σ_2) is a Nash equilibrium. However, the analysis that leads us to it is not obvious.

Consider that each player will likely begin to analyze strategies at a focal point (Schelling, 1958). In this case, we would expect the focal point to be either 2 or 100 with 100 being the most likely as it can provide a higher payoff. Each player considers σ_{100} and realizes that it is weakly dominated (definition 2) by σ_{99}, that is $\sigma_{99} \underset{wd}{>} \sigma_{100}$. In Figure 2.1, an edge directed from vertex *a* to vertex *b* signifies a relative preference for vertex *b*. So, there is an edge from σ_{100} to σ_{99} due to weak domination.

Each player then decides that it is impossible for their opponent to choose σ_{100} because they are **completely certain their opponent is rational** and in every situation, σ_{99} is as good or better than σ_{100}. So, the

resulting set of available quotes after eliminating σ_{100} is $[2,99]$. Each player then performs the same analysis and realizes that in the resulting set of available quotes, $\sigma_{98} \underset{wd}{>} \sigma_{99}$. Thus, σ_{99} is eliminated. This process is iterated until the only remaining option is σ_2. The preferences that are generated by iteratively eliminating the weakly dominated strategies are shown in Figure 2.1 as dashed lines.

2.2 Experimental Evidence

In (Becker et al., 2004), the authors ran a one-shot TD competition in which the competitors were drawn from the Game Theory Society. Each of the competitors submitted a strategy and the competitors were matched pairwise to every other competitor. The most successful strategy was σ_{97}. The mode of the strategies was σ_{100} with $N = 10$. The second most common was σ_{98} with $N = 9$. There were a total of 25 participants that used a strategy on the interval $[94,99]$, 10 used $[100]$, 3 used $[2]$, and 7 on the interval $[4,93]$. The authors of that paper theorize that there were 3 types of players involved. One type was an irrational player that played σ_{100}. The second was a rational player that played a best strategy given a belief about what others would play. The last was a type that either defaulted to the Nash equilibrium or started from a focal point of 2.

2.2.1 Non-zero Uncertainty

We focus on the rational type which settled into an equilibrium at σ_{98}. This study is interesting because the participants are people who have knowledge of game theory. Assuming that each player was attempting to maximize their payoff, it must be true that the players of this type were not certain that other players would not choose σ_{100} as their strategy. If the players were certain that opponents would not choose σ_{100}, they would have eliminated this as a strategy. In turn, this would have eliminated σ_{99}. And, if no player will choose σ_{99}, then σ_{98} will not maximize the payoff. In general we formalize this into proposition 1.

Proposition 1. *Let p be some rational player in the traveler's dilemma attempting to maximize their payoff. If p does not choose the Nash equilibrium strategy, σ_N, then it must be true that their belief regarding whether other players will choose σ_{100} has non-zero uncertainty. This is equivalent to having non-zero uncertainty regarding opponent rationality in general.*

Proof. If a player believes that their opponent will not choose σ_{100} with zero uncertainty due to weak domination, then it may be eliminated as a possibility. In the resulting space of possible strategies, $[2,99]$, σ_{99} possesses all the same properties that led to the elimination of σ_{100}. Thus it is proved by induction. □

2.2.2 Ramifications for Analysis

From proposition 1, in order to account for the evidence in (Becker et al., 2004) we must consider that players may have a non-zero uncertainty regarding the rationality of opponents. However, if we allow uncertainty, the

iterated elimination of weakly dominated strategies that we used to rationally deduce the Nash equilibrium fails. Consider, that if we can't eliminate σ_{100} then there is one possible opponent strategy on which σ_{98} does not provide a payoff equal to or better than σ_{99}. Therefore, σ_{98} does not weakly dominate σ_{99}. Visually, the effect is only the solid edges in Figure 2.1 can be deduced if there is any uncertainty regarding opponent rationality.

Even in the face of uncertainty, it seems intuitive that σ_{98} should be preferred over σ_{99} since the only possible scenario in which σ_{99} provides a better payoff is weakly dominated and therefore *unlikely*. To address this shortcoming, we will define a more general notion of weak domination, called fuzzy weak domination, that doesn't require certainty.

2.3 Fuzzy Logic

Here we will briefly introduce the core concepts of fuzzy logic.

Fuzzy logic was originally formulated in (Zadeh, 1965) to provide a method of reasoning in non-boolean contexts. As an example, consider a situation in which a recipe prescribes 2 minutes of boiling for large eggs and 1 minute for small eggs. By boolean logic, if an egg belongs to the set of small eggs it should be boiled for precisely 1 minute and 2 if an egg belongs to the set of large eggs. But what is the precise definition of large and small? Any value given to precisely identify the weight of a small egg and large egg will fail to cook the eggs properly unless the egg is the precisely prescribed weight.

In reality an egg may be somewhat large and somewhat small. That is, an egg can be considered to belong to both the set of large eggs and the set of small eggs with varying degrees of **membership** or **certainty**. Therefore, we can define a function with range $[0,1]$ that fuzzifies the weight of the egg into the eggs membership in the fuzzy set of large eggs. We can likewise define a function that fuzzifies the weight of the egg into the eggs membership in the fuzzy set of small eggs. Then based on the certainty that an egg is in the fuzzy set of large eggs and the fuzzy set of small eggs an appropriate combination of the associated boiling times can be found.

Fuzzy logic necessarily redefines the logical operators common in boolean logic to work in an infinitely valued logic context. The result is that the boolean operator exists as a special case of the fuzzy operator. We quickly give the fuzzy operator definition for the NOT, AND, and OR operations.

The logical boolean NOT operator converts 1 to 0 and 0 to 1. The fuzzy **NOT** operation is defined as $1-\mu$. Boolean AND operations return 1 if every operand in the operation is equal to 1. Fuzzy **AND** is equivalent to the min of the list of operands. Finally, boolean OR returns 1 if any of the operands are 1. The fuzzy **OR** operation returns the max of the list of operands.

2.4 Fuzzy Weak Domination

In definition 3 we give a formal definition for fuzzy weak domination (FWD). To develop an intuitive understanding we will demonstrate how FWD allows us to reason about the ordering of strategies in the TD in the face of uncertainty.

Definition 3. *For two strategies, σ_α is said to fuzzy weak dominate σ_β with certainty $\zeta = \mu(s_\beta)$, that is $\sigma_\alpha \underset{fwd}{>} \sigma_\beta$, iff there exists one or more opponent strategies s.t. σ_α provides a better payoff than σ_β and the set of opponent strategies which provide a better payoff to σ_β must be a subset of the fuzzy set of all weakly dominated strategies with certainty $\zeta' > 0$. That is $s_\beta \in S_{fwd}$ with certainty $\zeta' > 0$.*

Let $\mu(s_\beta)$ be a membership function that transfers the membership in S_{fwd} from s_β to σ_β, where $s_\beta \in S_{fwd}$ with certainty ζ'. Further, if s_β contains more than one strategy, then the fuzzy membership of each element in the set is combined through an AND operation with a result equal to the minimum membership of any element in $s_\beta \in S_{FWD}$. Finally, for $y = \mu(x)$ it must be true that (1) $y \in [0,1]$ $\forall x \in [0,1]$, (2) $\exists x > 0$ s.t. $\mu(x) = 0$, and (3) if $x = 0$ then $y = 0$.

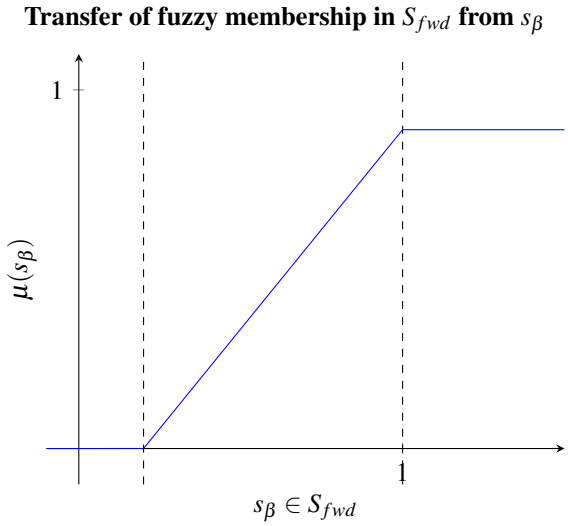

Figure 2.2: A simple piecewise membership function for transference of fuzzy weak domination that satisfies the requirements in definition 3.

2.4.1 Example Execution of Partial Ordering by Iterated FWD

Initially the set of all FWD strategies contains only an empty set, $S_{fwd} = \{(\{\}, 1)\}$. Notice that S_{fwd} is a fuzzy set, so an entry in the set constitutes the value and membership pair. First, we find the set of opponent strategies for which σ_{99} provides a better payoff than σ_{100}. The resulting set is $\{\sigma_{100}, \sigma_{99}\}$. So, the first requirement in the definition evaluates to true. Next, we find the set of opponent strategies for which σ_{99} provides a worse payoff, s_β.

By the definition of FWD, we need to compute $\mu(s_\beta)$ and in this case $s_\beta = \{\}$. We can easily retrieve the membership associated with each element in s_β from the S_{fwd} and then apply the membership transference function, μ, to the minimum. A convenient μ is a simple piecewise linear function as shown in Figure 2.2 that satisfies the requirements in definition 3. The minimum membership of any element in $s_\beta = \{\}$ in S_{fwd} is 1 and $\mu(1) = 0.89$. We now update the set of FWD strategies to be $S_{fwd} = \{(\{\}, 1), (\sigma_{100}, 0.89)\}$. Now, rather than having absolute certainty that an opponent will not choose σ_{100} we say that σ_{100} is a member of the fuzzy weak dominated fuzzy set by with membership certainty equal to 0.89.

We iteratively apply this process and find the set of opponent strategies for which σ_{98} provides a better payoff than σ_{99}. This set is $\{\sigma_{99}, \sigma_{98}\}$. Next, we find the set of opponent strategies for which σ_{99} provides a worse payoff. This set is $s_\beta = \{\sigma_{100}\}$. The minimum membership of any element in the resulting s_β in S_{fwd} is 0.89 and $\mu(0.89) = 0.77$.

Continuing to apply iterative FWD yields a partial ordering of strategies that tends to uncertainty. After a small number of iterative steps the certainty goes completely to zero. In Figure 2.3 we show that μ in Figure 2.2 allows the strategies from σ_{100} to σ_{94} to be ordered based solely on fuzzy weak domination.

Partial and Total Ordering by Fuzzy Weak Domination

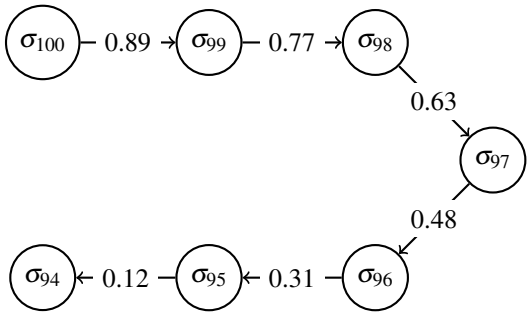

Figure 2.3: An edge directed from vertex *a* to vertex *b* signifies a relative preference for vertex *b*.

2.4.2 Finding the Equilibrium in FWD

It is not initially obvious that the ordering in Figure 2.3 provides a specific prediction regarding the game's equilibrium. To see how the equilibrium emerges we consider the logical complement of each of the propositions.

For each member in the fuzzy set of fuzzy weak dominated strategies we can calculate our certainty that the member is not in the set. As an example, we can say that if $\sigma_{99} \underset{fwd}{>} \sigma_{100}$ with certainty 0.89 then it is also true that σ_{99} does not fuzzy weak dominate σ_{100} with certainty 0.11.

So, for a fuzzy partial ordering like that shown in Figure 2.3 there exists a natural equilibrium at the point of certainty inflexion. Notice that with certainty 0.63 $\sigma_{97} \underset{fwd}{>} \sigma_{98}$. However, we can say that $\sigma_{96} \underset{fwd}{\not>} \sigma_{97}$ with

certainty 0.52. So, we are more certain that σ_{97} is not weakly dominated by σ_{96} than we are certain that it is weakly dominated. Therefore, a player attempting to maximize their payoff would not prefer σ_{96}.

It is not obvious if every partial ordering which results from iterated fuzzy weak domination possesses an inflexion point equilibrium.

2.4.3 Intuitive Understanding

The intuitive rationale that results from this analysis would be as follows. It is highly unlikely that the opponent will choose σ_{100} since σ_{99} is always as good or better. Since, σ_{100} is highly unlikely and σ_{98} is always as good or better than σ_{99} on every other strategy, it is unlikely that an opponent will choose σ_{99}. It still seems likely that σ_{97} is preferred because it provides a better payoff than σ_{98} on every strategy that is not highly unlikely or unlikely. At this point, certainty has fallen low enough that the player's belief has shifted such that the player has a higher certainty that σ_{96} does not fuzzy weak dominate σ_{97}. Thus the player chooses σ_{97}.

2.4.4 Comparing FWD to the Experimental Results

The results reported in (Becker et al., 2004) provided inspiration. Specifically, their work led us to consider that players may hold their opponents rationality as uncertain. That being said, it is notable that the equilibrium predicted by FWD with μ in Figure 2.2 is very close to the experimental result equilibrium (σ_{98}) as this was not engineered. We consider this to be affirmation (though not quite evidence) that FWD may accurately capture the rationale involved when humans engage in the TD.

It is more noteworthy that σ_{98} is the equilibrium in Figure 2.4 that corresponds to the largest interval when the x intercept of the general piecewise linear μ is swept on the domain $[0, 0.5]$. Therefore, if we assume that μ for an individual player may have an x intercept drawn randomly from the possibility space, then in general the most probable equilibrium is predicted to be σ_{98} by FWD. We consider this to be compelling evidence that FWD may accurately capture a specific type of rationale employed by the highest performing humans when engaged in the TD.

2.5 Related Work

In (Basu, 2007) the author and original creator of the TD posits that it may be necessary to relax the rationality assumption in order to resolve the paradox. In this paper we do not relax the assumption that each player is rational. We instead relax the players' beliefs regarding the other's rationality.

The same author previously pointed out in (Basu, 1994) that fuzzy logic could be used to arrive at a better equilibrium in the TD. However, that discussion centered around treating the quote as a fuzzy value. Here we

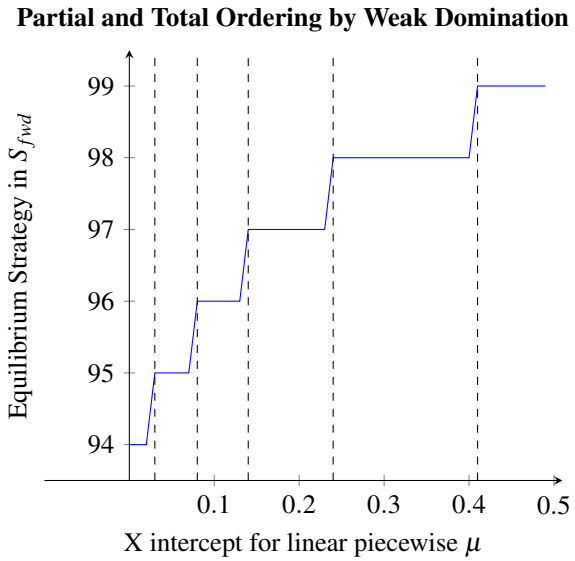

Figure 2.4: Effect on the equilibrium predicted by FWD as a function of x intercept

use fuzzy logic to deal with uncertainty regarding opponent rationality.

The experimental analysis in (Becker et al., 2004) cast the TD into a Bayesian game. They showed that if $P(\sigma_{100}) \geq 2\%$ then the Nash equilibrium of the Bayesian game is no longer located at σ_2. They continue to calculate the probability required to explain the results in the experiment given the Bayesian formulation. Probability and uncertainty are related but distinct ideas. Further, the intention of their analysis was to fit the data whereas FWD is formulated as a general analysis tool to find an expected equilibrium and natural extension of weak domination.

This is far from the first work to address the TD paradox. Neither is this the first paper to apply fuzzy logic to game theory. However, from our literature review this does seem to be the first paper to use fuzzy logic to address the inapplicability of iterated elimination of weakly dominated strategies when opponent rationality is not held certain in the TD.

2.6 Discussion and Future Work

As already stated, the goal in defining FWD was not to fit the data in the experimental results of previous work. Rather, FWD was formulated as a fuzzy logic extension of weak domination to enable strategy ordering in the face of uncertainty regarding opponent rationality. Weak domination then exists as a special case of FWD in which μ is a unit step function written as $u(x-\tau)$ with $\tau > 0$. So, the fact that the experimental equilibrium from (Becker et al., 2004) emerges as the most probable FWD equilibrium suggests that FWD captures an important facet of rational thinking in the TD.

With that said, more testing in other games is needed to be able to evaluate if this rationale is applicable

in a more general sense. It may also be found that FWD in the current formulation is incomplete. We intend to test this by changing the penalty involved in the TD game and comparing the effect on the predicted equilibrium against experimental results.

Another important point is that the choice of μ is non-trivial. More work is needed to evaluate the effect of other membership functions. An interesting and potentially promising extension would be the application of type 2 fuzzy logic so that one could formally reason when both the rationality of the opponent and the transference membership function are considered uncertain.

2.6.1 Application to Computational Sustainability

A method like FWD could potentially be applied to a wide range of games to identify likely equilibria. However, we specificially consider that FWD may prove useful in predicting the behaviour of opponents with uncertain rationality in games similar to the green security games defined in (Fang et al., 2015). These green security games are important to the field of computational sustainability as they help to anticipate the actions of poachers and direct conservation efforts. FWD may be potentially well suited to this as green security games are derivative of Stackelberg security games, which possess open problems regarding scalability when faced with uncertainty (Sinha et al., 2018).

2.7 Conclusion

We have shown that by allowing a player to consider the opponent's rationality to be less than certain, iterated elimination of weakly dominated strategies does not provide a total ordering. In this scenario weak domination does not facilitate the deduction of a Nash equilibrium in the TD. We formulated an infinitely valued logic (fuzzy logic) extension of weak domination referred to as fuzzy weak domination. By iterated application of fuzzy weak domination we can generate a partial ordering. In the case of the TD, this partial ordering possesses an uncertainty inflexion point at which the complement of some partially ordered strategy's membership in the fuzzy set of fuzzy weak dominated strategies is greater than the membership itself. This inflexion point seems to be an equilibrium in the TD based on similarity to experimental results in (Becker et al., 2004).

References

Basu, K. (1994). The traveler's dilemma: Paradoxes of rationality in game theory. *The American Economic Review*, 84(2):391–395.

Basu, K. (2007). The traveler's dilemma. *Scientific American*, 296(6):90–95.

Becker, T., Carter, M., and Naeve, J. (2004). Experts playing the traveler's dilemma. Technical report, Department of Economics, University of Hohenheim, Germany.

Fang, F., Stone, P., and Tambe, M. (2015). When security games go green: Designing defender strategies to prevent poaching and illegal fishing. In *IJCAI*, pages 2589–2595.

Goldman, W. (2001). *The Princess Bride*. MGM Home Entertainment.

Nash, J. F. et al. (1950). Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, 36(1):48–49.

Osborne, M. J. and Rubinstein, A. (1994). *A Course in Game Theory*. The MIT Press.

Schelling, T. C. (1958). The strategy of conflict: Prospectus for a reorientation of game theory. *Journal of Conflict Resolution*, 2(3):203–264.

Sinha, A., Fang, F., An, B., Kiekintveld, C., and Tambe, M. (2018). Stackelberg security games: Looking beyond a decade of success. In *27th International Joint Conference on Artificial Intelligence, IJCAI 2018*, pages 5494–5501. International Joint Conferences on Artificial Intelligence.

Zadeh, L. A. (1965). Information and control. *Fuzzy sets*, 8(3):338–353.

CHAPTER 3

Using Artificial Populations to Study Psychological Phenomena in Neural Models

3.1 Introduction

In the wake of success following the introduction of transformers in (Vaswani et al., 2017) and the public deployment of powerful variants of GPT, many have started to question if these models exhibit behavior similar to human cognition.

Work analyzing the cognition of these powerful, Turing-complete (Roberts, 2023) models is important not only to the explanation and interpretation of the models themselves but may offer insight into cognition more generally, a synergism best embodied by the interplay of reinforcement learning and neuroscience (Subramanian et al., 2022). We believe this emerging study of cognitive behavior in neural models can be improved by adopting methods from branches of science more typically associated with statistical testing. Without appropriate experimental methodology conclusions may not be robust in the face of variations, a symptom associated with the greater replicability crisis (Goodman et al., 2016). Research attempting replication and extension of ToM results in GPT-4 found that relatively small experimental alterations caused the effect to disappear (Ullman, 2023). This suggests the experimental design in the original study was insufficient to support the drawn conclusions.

This is precisely the motivation for the present paper. Claims that fail to be reproducible regarding powerful AI models may ultimately result in erosion of the public's trust and attention. Any study of neural model cognitive behavior should characterize not only the presence but the size of the effect and the significance. Doing so necessitates rigor which may decrease erroneous conclusions, and will permit better explanation of neural model cognitive behavior through meta-analytic study. To this end, we present and demonstrate an artificial population generation method for the study of cognitive phenomena in neural models with the hope that it will aid in the reproducibility of research regarding the behavior of neural models.

This paper contributes by drawing connections between social and behavioral experimental design and neural model uncertainty estimation resulting in a (1) tool called PopulationLM for the creation of populations of neural models via stratified MC dropout. We harvest novel metrics and explore population best practices by applying artificial populations to the (2) replication and extension of (Misra et al., 2021) (correlation analysis) and (3) (Sinclair et al., 2022) (difference analysis). We present novel results regarding the presence of typicality and structural priming effects in language models.

Phenomena	Study by	Measure(s)	Statistic	Significance	Experimental Var
Theory of Mind	Bubeck et al.	qualitative	—	—	not isolated
	Kosinski	frequency	—	—	not isolated
	Sap et al.	frequency	—	—	not isolated
	Ullman	frequency	—	—	isolated*
	Trott et al.	token probs	$\chi^2 + \beta$	reported	not isolated
Logical Reasoning	Binz and Schulz	token probs	$\chi^2 + t + \beta$	reported	isolated*
	McCoy et al.	frequency	—	—	isolated
	Lamprinidis	frequency	—	—	not isolated
Framing & Anchoring	Binz and Schulz	token probs	$\chi^2 + t + \beta$	reported	isolated*
	Jones and Steinhardt	frequency	—	—	isolated
	Suri et al.	frequency	—	reported	isolated*
Decision-making Heuristics	Binz and Schulz	token probs	$\chi^2 + t + \beta$	reported	isolated*
	Jones and Steinhardt	frequency	—	—	isolated
Typicality	Misra et al.	token probs	$r + \rho$	reported	isolated
Priming	Sinclair et al.	token probs	—	—	isolated
Emotion Induction	Coda-Forno et al.	frequency	$r + t$ + probit β	reported	not isolated

Table 3.1: Review summary of large language model behavioral studies. r = Pearson, ρ = Spearman, β = Berksons, t = t-test.

3.2 Behavioral Phenomena in Neural Models

In this section we review the current work related to the study of cognitive behavior in neural language models paying specific attention to the measures reported and methodology. Table 3.1 summarizes the works that have been identified, organized by the behavioral phenomenon that they investigate. This review and meta-commentary does not invalidate any of the findings in the associated papers. Rather, it serves as a compendium of work so far in this field and helps to illuminate the problem we wish to address.

The measures reported refers to the measure applied to the model output. Statistic refers to the statistical analysis applied to the measures. We find that most papers used atypical measures of effect like frequency of occurrence or qualitative analysis and tend to not use statistical testing. Those employing t-tests don't typically specify the particular test. Analogously, we find that less than a third of the papers report significance levels for their results. In contrast, most authors did isolate the experimental, independent variable. Rows marked with an * indicate works that did so only in a subset of reported experiments.

No study in our review utilizes uncertainty estimation to systematically perturb the model or the input. Therefore, no work has been done to study neural behavior in a population. In the latter half of this paper, we study two behavioral phenomena from table 3.1 in artificial populations: typicality and structural priming (SP). Typicality refers to a high degree of agreement across subjects in humans when ranking items as more or less typical of a given category and is known to be related to rate of retrieval of an item given the category (Rosch, 1975). Structural priming refers to the predilection for a sentence structure similar to the most recently observed syntactical structure (Pickering and Ferreira, 2008).

3.3 Populations of Neural Models

In all social and behavioral science, conclusions drawn from a single subject face severe limitations. Without a population of subjects it is impossible to know if the individual is typical along the dependent variable in the population or an outlier.

Studying the cognitive behavior of neural models, either as an ontology or in relation to human psychology, suffers from a similar limitation. There always exists a possibility that an expression of a behavior is anomalous or that the behavior is tenuously supported in the network.

In this paper we refer to models and their derivatives as different species. i.e. BERT and DistilBERT are individual species, while they both belong to the same family. Genus is reserved for fine-tuned variants.

Forming inter-species populations is an intuitive but flawed approach since we wish to facilitate the study behaviors that may emerge in specific species, as is known to occur as a function of model size (Wei et al., 2022). Inter-species populations don't permit this type of myopic study.

Instead we form populations using work from neural model uncertainty estimation. In that context, the goal is not a population but an estimation of model uncertainty. However, this is precisely the characteristic typically extracted from a population study, the degree to which a result is consistent across individuals. We refer to this as the population uncertainty. Several uncertainty estimation methods have been proposed in literature and can be placed into 4 broad groups (Gawlikowski et al., 2023), single network deterministic, test time augmentation, ensemble techniques, and Bayesian approximations.

Single network deterministic methods attempt to estimate the uncertainty of a network without multiple predictions being made. However, they trade accuracy for speed. Test time augmentation methods perform perturbations of the input data and estimate uncertainty across a single model's outputs (Lyzhov et al., 2020). Though this is a promising solution for closed source models, there exists a bound on the perturbation resolution possible in transformers with test time augmentation due to Hahn's lemma (Hahn, 2020). Ensemble techniques, generally outperform Bayesian methods (Lakshminarayanan et al., 2017) but require multiple models trained independently. The price associated with from scratch training makes this a poor solution (Sharir et al., 2020). Therefore, Bayesian approximation is the most applicable uncertainty estimation technique for the creation of populations of open source models.

3.4 PopulationLM

We use Monte Carlo (MC) dropout (Gal and Ghahramani, 2016) to form populations from base models. A neuron mask is assembled from the instances of random variables and placed on the network. The resulting masked network is then used to perform inference. Each network mask is typically applied once and discarded. However, in the context of behavioral studies, it is desirable to apply a set of stimuli to the static

Model Species	Paper	Typicality KS test	SP KS test	Type (parameters)	Training Data
DistilBERT	Sanh et al.	0.056 (p≈0.055) ✗	0.04 (p<0.05)	MLM (66M)	BookCorpus, Wiki
BERT Base	Devlin et al.	0.051 (p≈0.108) ✗	0.03 (p≈0.06) ✗	MLM (110M)	
BERT Large		0.072 (p<0.01)	0.05 (p<0.01)	MLM (340M)	
GPT	Radford et al.	0.069 (p<0.01)	0.08 (p<0.01)	MLM (120M)	BookCorpus
DistilGPT-2	Sanh et al.	-0.072 (p<0.01)	0.45 (p<0.01)	CLM (82M)	BookCorpus, WebText
GPT-2	Radford et al.	-0.03 (p≈0.685) ✗	0.29 (p≈0.1) ✗	CLM (117M)	
GPT-2 Medium		0.075 (p<0.01)	**0.51 (p<0.01)**	CLM (345M)	
RoBERTa Base	Liu et al.	0.065 (p<0.02)	0.08 (p<0.01)	MLM (125M)	BERT train data, Stories, CC, OpenWebText, News
RoBERTa Large		**0.15 (p<0.1)**	0.19 (p<0.1)	MLM (355M)	

Table 3.2: Kolmogorov-Smirnov test for each population and each experiment compared to the base model. Null hypothesis H_0 is population probabilities and base model probabilities are drawn from the same underlying distribution per species. Populations very similar to the base model have an ✗.

population for within-group, paired-sample tests. We contribute stratified MC dropout, a variation that generates and maintains a user defined number of masks for any PyTorch compatible network. While the provided library is implemented only for PyTorch, the method is, in principle, applicable to any neural network library that supports inference-time dropout.

While it is true that dropout populations approximate the distribution of a deep Gaussian process (Gal and Ghahramani, 2016), the degree to which this will approximate a group of humans is not known. Therefore, we don't claim that this method approximates results typical of human studies. We claim that evaluating the dropout population outputs as a group will help the results to be more robust in the face of variation due to decreased presence of poorly supported behaviors as a direct consequence of their tendency to converge to a Gaussian process.

We do not apply any aggregation to the population outputs. Instead, we adopt methodologies from psychological and pharmaceutical domains to treat the model responses as populations of individuals and directly apply statistical analysis. This approach provides a more robust view of expected model behavior under variation with improved insights regarding population certainty and statistical significance.

3.4.1 Analysis of the Populations

We evaluated the efficacy of the populations to generate outputs which are statistically distinct from the base models for each species via the non-parametric Kolmogorov–Smirnov (KS) test. It compares the shape and location of two distributions but makes no assumptions about the nature of the underlying distributions. The null hypothesis is that the sample distributions will have similar shape and location.

In table 3.2, we find that the underlying distributions for the species' base models and their populations are not representative of the same distribution with the exception of GPT-2, BERT base, and DistilBERT as judged by the significance of the p value. We inspected these results by observing plots of each and include RoBERTa in figure 3.1 juxtaposed with its associated dropout population. An obvious benefit of

the population is the narrowing of the confidence bounds on the regression due to augmented elimination of alternative regressions. As suggested by the KS test, the relationship between typicality and population probability in experiment 1 is shown to be quite distinct from that of the base model.

The KS test is a useful method to characterize the likelihood with which population results will vary from the base model on a given task, with a high effect size indicating high likelihood. However, the test is meaningful for the target behavior or context only. This is evinced by the large change in effect among the GPT-2 family from the typicality experiment to the structural priming in table 3.2.

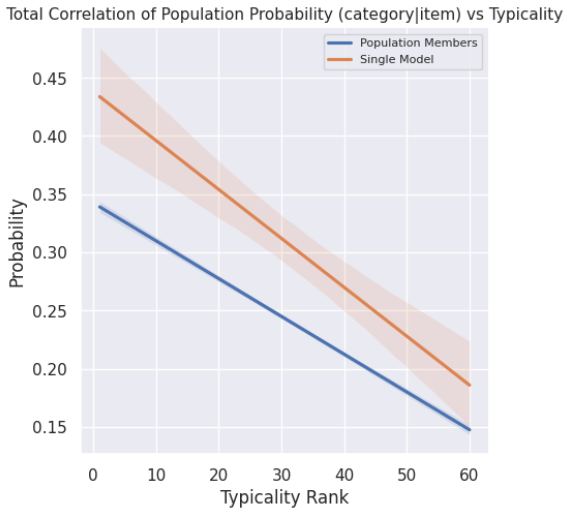

Figure 3.1: Single model regression vs population of models probability-typicality regression for RoBERTa Large. Rank is inversely related to typicality. 95% confidence intervals shown for both with very narrow bounds on the population. All rights reserved.

3.4.1.1 How Much Dropout Is Necessary?

MC dropout has been applied to transformers previously in (Shelmanov et al., 2021) and (Vazhentsev et al., 2022). In both of these papers the authors experimentally found 0.1 to be the most effective dropout rate for discouraging incorrect, poorly supported outputs.

We experimented with dropout rates from 0.1 to 0.8. We found no advantage in using larger rates of dropout for experiments, as increased rates caused signal erosion with all behavioral correlations being dissolved beyond rates of 0.5 Therefore, we recommend that statistical studies adopt a 0.1 nominal dropout rate.

3.4.1.2 How Big Should the Artificial Population Be?

Population size for a study is related to two important statistical measures, significance and power. The significance of a *result* is a measure of the probability of the null hypothesis. The power of a *test* is a measure

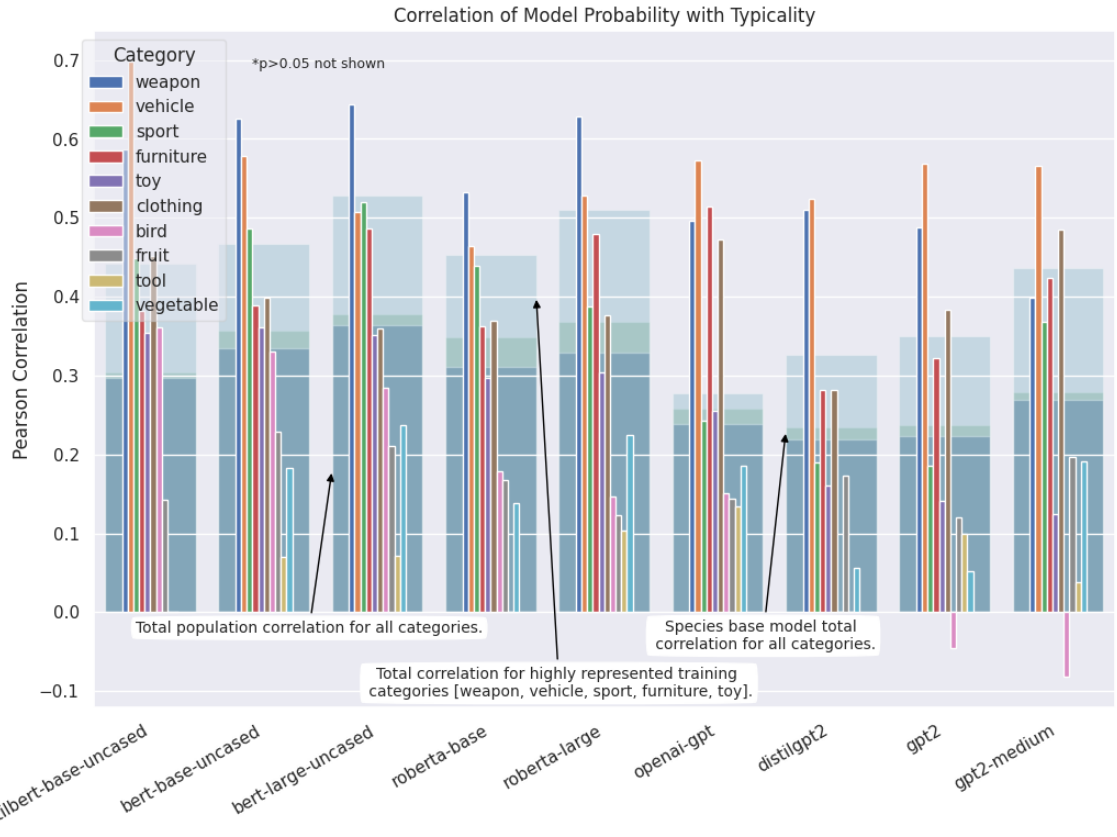

Figure 3.2: For each model the colored bars show within category Pearson correlation ($p<0.03$). For each model the total Pearson correlation ($p<0.03$) is shown as the gray background bar. The total Pearson correlation ($p<0.01$) for well understood categories (categories with an average item frequency > 60000 in training data for Bert) is shown as the light blue bar. In well understood categories, typicality of the item may explain up to $r^2 \approx 20\%$ of the category probability volatility. All rights reserved.

of the probability that the test will correctly reject the null hypothesis and avoid a false negative.

For evaluating cognitive behavior in neural models, the power of a test is less important as the effect of a false negative is not likely to cause damage. However, the significance of a result is of the utmost importance as this permits meta-analytical extension and can act as a mitigator of sensationalism when applied properly.

We empirically find that a population of 50 is an acceptable compromise, providing sufficient statistically significant deviation from the base model in table 3.2 without dramatic computational costs. Interestingly, it seems the models tend to have correlated relationships with associated dropout populations. The KS tests for the two experiments in table 3.2 show that BERT, GPT-2, and RoBERTa models tended to have KS effect sizes which were rank correlated across experimental populations within model families. However, the model correlations don't extend outside the family. This suggests that 50 member populations may tend to be sufficient for the approximated deep Gaussian process to emerge.

3.5 Experiment 1: Typicality Effects

We reproduce and extend the experiment conducted in (Misra et al., 2021) which assessed the base model total correlation between probability and typicality. Our base model probabilities agree with past results, and we contribute novel tests using dropout populations and within category analysis which shed light on the factors that support the emergence of typicality effects in language models.

3.5.1 Experimental Setup

We use typicality data from (Rosch, 1975) which gives a typicality rank, r_i, for each item, i, in category C. As in the original experiment, we construct prompts, π_i, for each $i \in C$ and measure the probability assigned to the category given the prompt, $P(C|\pi_i)$. So, for each category, only the item in the prompt (independent variable) will change across queries while the effect on the category probability is measured (dependent variable). After each prompting, the model input is flushed, guaranteeing that only the independent variable is manipulated for each trial for each category. This necessitates that the results be evaluated within category, since cross category results are not controlled. However, we also evaluate the test results across all categories for each model as a direct comparison to the results reported in the original paper.

3.5.2 Individual Probability Correlation Test

For each population, we test for behavior consistent with typicality by evaluating the Pearson correlation between $P(C|\pi_i)$ and r_i for all $i \in C$ and for all categories in the dataset. We hypothesize that, consistent with previous results, the probabilities output by the models will be positively correlated with typicality.

As predicted, all models show significant ($p<0.05$) probability/typicality correlation within nearly all categories consistent with typicality in humans in figure 3.2. DistilBERT shows insignificant correlation with the categories tool and vegetable, while DistilGPT2 and RoBERTa base both have insignificant correlation with tool. More generally, the correlation between probability and typicality is strongly conditioned upon category for all models. The behavior shows strong differentiation between causal (CLM) and masked language models (MLMs). Among all MLMs the total correlation is markedly higher than for CLMs. Further, the categories for which each model most exhibits typicality behavior differs across MLMs and CLMs.

The green bars in figure 3.2 represent the total correlation (across categories) obtained by evaluating only the base models and agree with past results (Misra et al., 2021). However, the total population correlation, shown in dark blue, suggests that the base model total correlation is an over estimation of the true total correlation.

3.5.3 Population Uncertainty Correlation Test

We hypothesize that population uncertainty will be positively correlated with diminishing typicality. That is, as stimuli become less typical, the population will have decreasing agreement. Therefore, we test for correlation between normalized population standard deviation (as a measure of group uncertainty), $\frac{\sigma(P(C|\pi_i))}{\mu(P(C|\pi_i))}$ and typicality rankings r_i.

In figure 3.3, for masked language models, mean normalized population uncertainty has a significant positive correlation as typicality diminishes mediated by category. The models tend to become more uncertain as the items become less typical. Therefore, we believe that masked language models, like humans (Rosch, 1975), are more certain when inferencing about typical items. The categories which are most positively correlated with population certainty tend to be consistent with those which were most correlated with probability.

Interestingly, the standard deviation of model probabilities was found to scale with the mean of the probability, giving the appearance of increased population agreement as probability declined. Therefore, mean normalization is used. Mean normalized uncertainty may be more meaningful than standard deviation alone for models which learn to output probabilities.

In sharp distinction, all causal language models exhibit negative certainty/typicality correlation. We speculate that this may be due to differences in training data and modeling objective. i.e. it is not typical for humans to say extremely obvious things like "A sparrow is a bird." Therefore, a dropout trained conversational model may have high uncertainty regarding highly typical item/category pairings in completions. However, this hypothesis is not readily testable due to GPT-2 training data unavailability. Further, the categories among the CLMs which are most negatively correlated with population certainty do not seem to be the same categories as those which were most positively correlated with probability in figure 3.2. This suggests that CLMs represent something all together different than MLMs in their population uncertainty.

3.5.4 Confound Test

We considered that frequency of an item within the training data could act as a third variable and confound the results. To address this we evaluate the Pearson correlation of item frequency in the training data with typicality ranking. We hypothesized that item frequency would act as a confound at some level. We used the BERT family training data frequencies from (Zhou et al., 2022) to assess training data frequency correlations.

We found no correlation between item typicality and frequency in the training corpus. Nor did we find a correlation between the normalized certainty and item frequency. There was a slight correlation (Spearman's r=-0.08 $p<0.01$) between probabilities output by BERT and item frequency. However, the effect size suggests that this is insignificant.

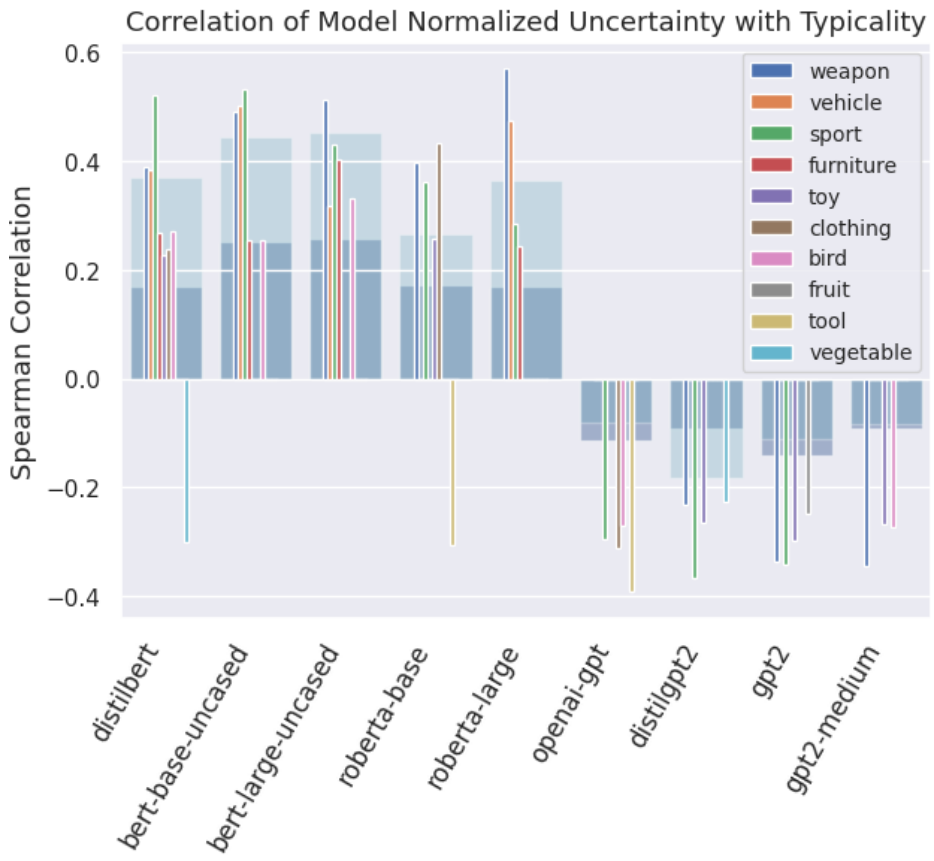

Figure 3.3: CLMs exhibit negative Uncertainty/Typicality correlation while MLMs exhibit positive correlation. Within category, total Spearman correlation ($p<0.08$), and total well represented category (item frequency > 60000) Spearman correlation ($p<0.01$) are shown. All rights reserved.

We hypothesized that category "understanding" may be important for the emergence of typicality behavior and that mean within category item frequency in training data may predict category understanding. To test this, we performed a regression between within category Pearson correlation and mean within category item frequency in the training data for the BERT population.

In figure 3.4 we find that average item frequency within a category is highly correlated with the strength of typicality effects exhibited by the model within that category. The exception being the categories tool and toy. Further research may be necessary to determine why these categories do not fit the otherwise established trend. We suspect that this is the result of conflicts from the basic-level effect, that humans have a preferred level of categorization, which has a known relationship with typicality (Rosch et al., 1976). Tool and toy may be outliers because they are not at the basic categorization level for many of the items listed in those categories.

If the anomalous categories are removed, the correlation between within category probability/typicality correlation and within category mean item frequency in training data is Pearson r=0.98 ($p<0.01$) and with

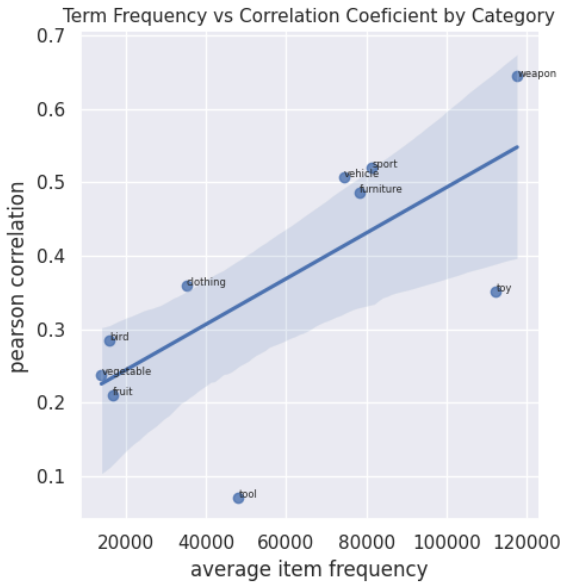

Figure 3.4: Emergence of within category behavior consistent with typicality in BERT is strongly predicted by within category item frequency in training data. All rights reserved.

tool and toy included Pearson r=0.7 ($p<0.03$).

Another measure of concept "understanding" is the persistence of a concept through growing rates of dropout. As the dropout rate increases, more and more neurons are masked in the population members causing concepts with fewer constituent neurons to be ablated. So, we swept the dropout rate for the population from 0.1 to 0.8 and found that the categories which were highly represented in the training data tended to persist as dropout increased, while the categories with less training data representation tended to decorrelate at lower dropout rates.

We interpret these complimentary results to suggest that model "understanding" of a category may be driven by overall category representation in the training data and that, within categories which are well understood, models are likely to exhibit typicality effects. We find that this is the case for all tested model species as restricting the total Pearson correlation to the categories which are well represented in the BERT training data, a partial constituent of all other model's training, leads to significant increases in all model probability-typicality correlations in figure 3.2.

3.5.5 Comments

We find that language models strongly exhibit typicality effects both in individual model probabilities and in population uncertainty mediated by model "understanding" of category. The square of probability/typicality correlation in figure 3.2 shows that $10\% < r^2 < 25\%$ of the well represented category probability variances

for all model populations, excluding GPT-1 which was trained on substantially less data, are accounted for by typicality effects. Strong typicality effects tend to emerge in categories with at least 80000 training examples per within category item.

Model Species	Wilcoxon $P(S_x\|\pi_x) > P(S_x)$	$\frac{\mu(P(S_x\|\pi_x))-\mu(P(S_x))}{\mu(P(S_x\|\pi_y))-\mu(P(S_x))}$	Pearson(AT-CT, PT-CT) r	Structural Priming
DistilBERT	0.42 (p≈1)	—	—	None
BERT Base	0.27 (p≈1)	—	—	None
BERT Large	0.47 (p≈1)	—	—	None
GPT	0.14 (p≈1)	—	—	None
DistilGPT-2	0.82 (p≈0)	0.93 ±0.001	0.89 (p<0.01)	None
GPT-2	0.96 (p≈0)	0.96 ±0.001	0.93 (p<0.01)	None
GPT-2 Medium	0.99 (p≈0)	0.98 ±0.001	0.94 (p<0.01)	None
RoBERTa Base	0.99 (p≈0)	0.98 ±0.001	0.70 (p<0.01)	Marginal
RoBERTa Large	0.99 (p≈0)	0.96 ±0.001	0.67 (p<0.01)	Marginal

Table 3.3: Test results used to detect structural priming. From left to right: the first relates preference for priming; the second finds the percent of the preference magnitude not attributable to SP; the third measures the correlation between SP and an alternative.

3.6 Experiment 2: Structural Priming Effects

In (Sinclair et al., 2022) the authors investigated whether language models exhibit behavior consistent with the structural priming effect. We run a similar experiment using sentence data from their work. However, we use a dropout population, modify the experimental setup to control for unaddressed confounds, and perform a split-group cross validation.

3.6.1 Experimental Setup

To test for SP in language models we adopt 3 treatment conditions: the control (CT) is the probability of a sentence, S_x, without any priming $P(S_x)$; the primed treatment (PT) is the probability of that sentence when the language model is first prompted with a sentence, π_x, of similar structure $P(S_x|\pi_x)$; and the alternative treatment (AT) is the probability of S_x when prompted with a sentence, π_y, of differing structure $P(S_x|\pi_y)$ but identical semantic meaning. Any effect AT has will not be analogous to SP. However, it is not a placebo as it may not be inert. Therefore, both AT and PT must be compared to CT for contextualization.

We split 3000 examples into two groups and conduct all 3 treatments on all 50 population members per species. The results for the first group of 1500 are reported and the results for the second set of 1500 are used for cross validation. The cross validation showed all results repeated within ± 0.02 ($p<0.01$) of our reported results.

3.6.2 Individual Probability Difference Test

For behavior consistent with SP to be present, the relationship PT>CT must tend to hold. To test this, we employ the Wilcoxon signed rank test, a non-parametric test appropriate for testing relative ranking of paired samples.

In table 3.3 the results show that only GPT-2 and RoBERTa exhibit a preference for PT over the control. These models require subsequent testing as SP is one possible explanation for their preference, but an alternative hypothesis is that the models prefer being primed with anything at all.

Preference for priming could be induced by the presence of WebText and OpenWebText in the training data of GPT-2 and RoBERTa families as these possess conversational data in which SP is more likely to be observed.

3.6.3 Elimination of Alternative Hypotheses

To eliminate a preference for priming regardless of structure as an alternative hypothesis, we find the 95% confidence interval of $\frac{\mu(AT)-\mu(CT)}{\mu(PT)-\mu(CT)}$. This is the fraction of the probability change induced by PT which is not attributable to SP. Wilcoxon is not used in this case as it would result in the cancellation of the control group due to internal subtraction. It is possible that the effect magnitude will be similar but the individual samples not be correlated. Therefore, we also find the Pearson correlation between PT-CT and AT-CT.

For all models the alternative treatment produced an average effect which was 96% as large as the mean change due to SP. Further, the GPT-2 family showed strong correlation between AT-CT and PT-CT, suggesting these models do not prefer priming with a similar structure. However, the results in table 3.3 show that the RoBERTa family has a response to PT distinct from the AT response based on Pearson's r.

3.6.4 Comments

In contrast to previous work we find little evidence for the presence of structural priming effects. The RoBERTa family of models exhibits a response that is distinct when primed with a sentence of similar structure to the target sentence. However, the preference magnitude is not differentiable from an alternative structure priming. No other models exhibit significant, distinct effects.

3.7 Conclusions

This paper addresses a current need in the study of cognitive behavior in neural models by introducing PopulationLM[1], a system built on MC dropout for the creation of efficient populations of neural models. This permits population based analysis of model behavior which may decrease the presence of atypical behaviors.

[1] https://github.com/JesseTNRoberts/PopulationLM

In both experiments our population studies, when compared to the original experiments of other authors, show that conclusions drawn from single models tend to over estimate the presence of cognitive behaviors. Beyond robustness, populations permit the study of divergence or decorrelation as a function of dropout and characterization of population uncertainty or disagreement.

We have conducted novel experiments using PopulationLM regarding the presence of typicality and structural priming in language models, being careful to isolate and analyze along independent variables and report effect sizes and significance. We find that typicality is consistently present while structural priming seems to not be, with both having predictable ties to behavior representation in training data.

PopulationLM may have further reaching applications beyond the study of cognitive behavior. Many papers have begun to systematically study prompt pattern effects (White et al., 2023). These possess similar issues of robustness to cognitive studies and could benefit from study among a population. Further, it's possible that populations of models may serve as proxies for initial human behavior studies in the future. This could augment the ethical and financial efficacy of psycholinguistic research (Brysbaert, 2019).

Test time augmentation (Gawlikowski et al., 2023) may be used to create local variations that perform similarly to dropout populations. However, the effects will decay with the length of the decoder context (Hahn, 2020). The longer the priming, the less effect each individual token, including the experimental prompt, will have. We intend to investigate the use of test time augmentation for closed source language model systematic population studies in future work.

Finally, future work should investigate the (1) surprising increase in CLM certainty with decreased typicality, (2) the use of mean normalization to characterize probabilistic model certainty, and (3) the presence and impact of other cognitive phenomena like basic level effects.

3.8 Ethical Statement

Some work in the area of large language model cognitive behavior has produced conclusions which are not replicable when small variations are applied to the experiments. This coupled with the wide attention being given to large language models can lead to sensationalism and potentially contribute to the erosion of the public's trust in the scientific community. The hoped effect of this paper is to bring awareness and partially address the current situation by providing a more systematic method for improving the robustness of results.

Alternatively, if data were to be improperly handled or inappropriate statistical testing performed, the impact could be negative. PopulationLM has the ability to augment the number of datapoints on which testing may be performed. If not analyzed appropriately, the increased data may potentially be used to support fallacious conclusions.

Published Version

Roberts, J., Moore, K., Wilenzick, D., & Fisher, D. (2024). Using Artificial Populations to Study Psychological Phenomena in Neural Models. Proceedings of the AAAI Conference on Artificial Intelligence, 38(17), 18906-18914. https://doi.org/10.1609/aaai.v38i17.29856

3.9 Appendix

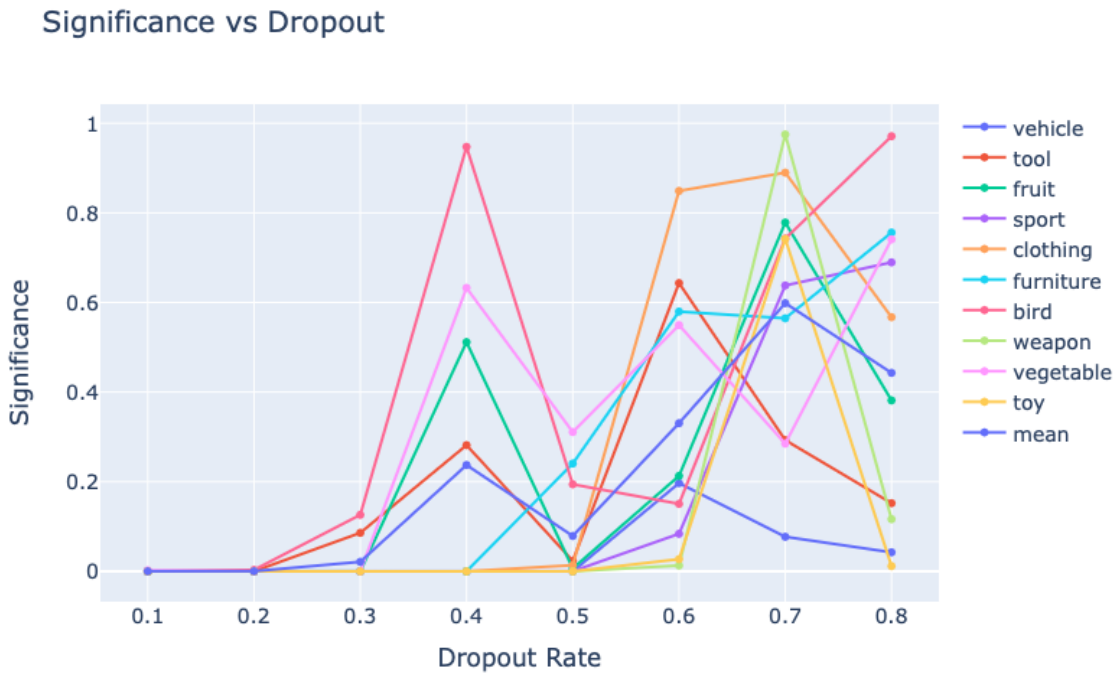

Figure 3.5: Pearson correlation significance between GPT-2 medium model probabilities and typicality as a function of dropout rate. Weapon, toy, sport, vehicle, clothing, and furniture are persistent as dropout increases.

In Figure 3.5, the order of category erosion when dropout is increased is strongly predicted by the magnitude of correlation of the category. This suggests that model understanding of a category is related to the dropout rate necessary to erode the cognitive effects associated with that category. Category understanding is also strongly predicted by the number of within category training tokens. This supports the hypothesis that PopulationLM tends to erode poorly supported behaviors. Those that are more well supported require a larger dropout rate to be eroded.

References

Binz, M. and Schulz, E. (2023). Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

Brysbaert, M. (2019). How Many Participants Do We Have to Include in Properly Powered Experiments? A Tutorial of Power Analysis with Reference Tables. *Journal of Cognition*, 2(1):16.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Coda-Forno, J., Witte, K., Jagadish, A. K., Binz, M., Akata, Z., and Schulz, E. (2023). Inducing anxiety in large language models increases exploration and bias. *arXiv preprint arXiv:2304.11111*.

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

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., and Zhu, X. X. (2023). A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*.

Goodman, S. N., Fanelli, D., and Ioannidis, J. P. (2016). What does research reproducibility mean? *Science translational medicine*, 8(341):341ps12–341ps12.

Hahn, M. (2020). Theoretical Limitations of Self-Attention in Neural Sequence Models. *Transactions of the Association for Computational Linguistics*, 8:156–171.

Jones, E. and Steinhardt, J. (2022). Capturing failures of large language models via human cognitive biases. *Advances in Neural Information Processing Systems*, 35:11785–11799.

Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Lamprinidis, S. (2023). Llm cognitive judgements differ from human. *arXiv preprint arXiv:2307.11787*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lyzhov, A., Molchanova, Y., Ashukha, A., Molchanov, D., and Vetrov, D. (2020). Greedy policy search: A simple baseline for learnable test-time augmentation. In *Conference on Uncertainty in Artificial Intelligence*, pages 1308–1317. PMLR.

McCoy, R. T., Pavlick, E., and Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*.

Misra, K., Ettinger, A., and Rayz, J. T. (2021). Do language models learn typicality judgments from text? *arXiv preprint arXiv:2105.02987*.

Pickering, M. J. and Ferreira, V. S. (2008). Structural priming: a critical review. *Psychological bulletin*, 134(3):427.

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

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

Roberts, J. (2023). On the computational power of decoder-only transformer language models. *arXiv preprint arXiv:2305.17026*.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104(3):192–233.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., and Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8(3):382–439.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sap, M., LeBras, R., Fried, D., and Choi, Y. (2022). Neural theory-of-mind? on the limits of social intelligence in large lms. *arXiv preprint arXiv:2210.13312*.

Sharir, O., Peleg, B., and Shoham, Y. (2020). The cost of training nlp models: A concise overview. *arXiv preprint arXiv:2004.08900*.

Shelmanov, A., Tsymbalov, E., Puzyrev, D., Fedyanin, K., Panchenko, A., and Panov, M. (2021). How Certain is Your Transformer? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1833–1840, Online. Association for Computational Linguistics.

Sinclair, A., Jumelet, J., Zuidema, W., and Fernández, R. (2022). Structural persistence in language models: Priming as a window into abstract language representations. *Transactions of the Association for Computational Linguistics*, 10:1031–1050.

Subramanian, A., Chitlangia, S., and Baths, V. (2022). Reinforcement learning and its connections with neuroscience and psychology. *Neural Networks*, 145:271–287.

Suri, G., Slater, L. R., Ziaee, A., and Nguyen, M. (2023). Do large language models show decision heuristics similar to humans? a case study using gpt-3.5. *arXiv preprint arXiv:2305.04400*.

Trott, S., Jones, C., Chang, T., Michaelov, J., and Bergen, B. (2023). Do large language models know what humans know? *Cognitive Science*, 47(7):e13309.

Ullman, T. (2023). Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Vazhentsev, A., Kuzmin, G., Shelmanov, A., Tsvigun, A., Tsymbalov, E., Fedyanin, K., Panov, M., Panchenko, A., Gusev, G., Burtsev, M., Avetisian, M., and Zhukov, L. (2022). Uncertainty Estimation of Transformer Predictions for Misclassification Detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8237–8252, Dublin, Ireland. Association for Computational Linguistics.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. (2022). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., and Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Zhou, K., Ethayarajh, K., Card, D., and Jurafsky, D. (2022). Problems with cosine as a measure of embedding similarity for high frequency words. *arXiv preprint arXiv:2205.05092*.

CHAPTER 4

Do Large Language Models Learn Human-Like Strategic Preferences?

4.1 Introduction

Transformer-based large language models (LLMs) have famously achieved state of the art performance on many tasks since their introduction by Vaswani et al. (2017). The analysis of these models is typically focused on benchmark tasks like Big-Bench (Srivastava et al., 2022), MMLU (Hendrycks et al., 2020), and Agieval (Zhong et al., 2023). Theoretical analysis of their computational abilities like (Roberts, 2024; Bhattamishra et al., 2020; Pérez et al., 2019) and empirical investigations of LLM cognitive/psychological behaviors are proliferating (Misra et al., 2021; Trott et al., 2023; Roberts et al., 2024; Binz and Schulz, 2023; Ullman, 2023; Suri et al., 2023), but remain far from common. However, in many human-adjacent applications, this latter type of investigation is paramount to successful LLM integration.

As both an illustration of this point and a motivational example relevant to the present paper, consider, a human carrying a heavy box who asks a human collaborator for help. Based on this input alone, the collaborator can quickly choose and apply their most preferred strategic mixture of vertical and horizontal force. In fact, the individual asking for help has implicitly relied upon the collaborator's possession of a compatible set of preferences over the possible strategies. Otherwise, the originator of the request would have found it necessary to provide a more detailed and precise request to ensure the collaborator acted appropriately.

If asked to help with a box, a robot is currently incapable of selecting from the possible strategies unless it has previously been imbued with a precise value function over the strategies or has suffered a regiment of reinforcement learning. In future work we hope to apply LLMs in support of human-robot interaction (HRI). However, for this to be possible, the supporting LLM must likewise have strategic preferences sufficiently similar to that of a human to permit effectual communication.

Further, applications like HRI require LLM behavior be stable under variations to avoid potentially dangerous strategic variations due to slight prompt variations. This point is timely as recent evaluations of some language model cognitive behaviors have been shown to fail to repeat under small variations (Ullman, 2023). To address this, we use PopulationLM (Roberts et al., 2024) to create systematically varied populations of each model for experimentation.

4.1.1 Aims of This Paper

In service of human-adjacent LLM integrations, this paper aims to understand if any current open-source language models exhibit stable, human-like preferences. To do this, we construct populations from a wide

variety of LLM species and evaluate their strategic preferences in a number of scenarios. Open-source models are studied exclusively in support of reproducibility. Closed source models are not static and may change without warning, resulting in the loss of previously studied behaviors as occurred in (Suri et al., 2023).

We first consider whether language models in the presence of simple and explicit values associated with strategies tend to have value-based preferences (VBP). From this task, we identify interesting models appropriate for additional testing. We then engage these models in two masked versions of the prisoner's dilemma, one with high and one with low stakes. Finally, we engage these models in experiments based on a low penalty and high penalty traveler's dilemma.

We find that, although LLMs are not explicitly trained to replicate the strategic preferences of humans, (1) some acquire stable human-like strategic preferences. Specifically, Solar (Kim et al., 2023) and Mistral (Jiang et al., 2023) are shown to have human-like self-consistent, non-brittle VBP. On the other hand, we find that (2) small models tend to prefer strategies based on superficial heuristics, (3) larger models tend to have decreased reliance on superficial information, and (4) some large models that exhibit VBP prove to be highly brittle under variations which may be related to the attention architecture. Finally, from the *in silico* experimentation we provide (5) evidence for the origin of human deviation from the Nash equilibrium in the traveler's dilemma.

In the course of this work, we contribute novel datasets for each scenario and (6) a novel method for constructing preference relations from a population of LLMs. The dataset and code for reproduction of these studies is made available and open source[1].

4.2 Related Work

In (Akata et al., 2023) the authors engaged GPT-3.5 and GPT-4 (OpenAI, 2023) in a number of iterated games including an iterated prisoner's dilemma. The authors find that both models tended to be punishing in response to betrayal though, prior to betrayal, they tended to cooperate. No matter how many times an opponent cooperated after the a betrayal, the models would not again reciprocate cooperation.

In (Fan et al., 2024), the authors engaged GPT-3.5 and GPT-4 in a number of games to evaluate their ability to act consistently with a prompted preference, refine belief, and take optimal actions. Their work is aimed at evaluating the potential integration of GPT-4 into games for research in social science. Their results suggest that GPT-4 fails to appropriately update and maintain beliefs necessary to choose optimal strategies and is therefore yet unsuited to integration into social science experiments.

In (Wang et al., 2023), the authors engage GPT-4 and Claude in a social game involving misinformation and provide a prompt pattern, related to chain of thought (Wei et al., 2022), to help the models reason in that

[1] https://github.com/JesseTNRoberts/Do-Large-Language-Models-Learn-Human-Like-Strategic-Preferences

environment.

A number of authors have explored LLM behavior in games. Their work provides a confidence from which it is reasonable to believe that some LLMs may learn strategic preferences from human language data. However, the focus of the existing work has been distinctly different from our aims. Our work is specifically differentiated by the fact that none of the existing work considers the stability of model preference, the effect of stake/penalty size (human-like or otherwise), or games like the traveler's dilemma in which human behavior tends to sharply contrast with game theoretic predictions. Further, all existing peer-reviewed related work is based on closed-source models, something we specifically avoid. Consequently, the measurement of model preference used by all existing work is akin to a cloze type task based on the generated token. In contrast, we use a method called counterfactual prompting to measure model evaluation probability. Finally, the number of models which previous work has considered is relatively small in comparison to the number we consider.

4.3 Do LLMs Prefer Strategies Based on Value?

Although past work has shown that GPT-3.5 and GPT-4 have preferences for higher valued strategies in a dictator game (Fan et al., 2024), it is not clear whether other model species possess similar preferences. Further, if a model has value-based preferences (VBPs), it is unclear how these preferences will fair under systematic perturbation. Poorly supported preferences in the network may fail to be sufficiently reliable to support human-adjacent NLP tasks. We therefore formulate RQ 4.3.1.

Research Question 4.3.1. *Given a set of strategies each with a clearly specified value, do LLMs tend to have value-based-preferences?*

4.3.1 Experimental Method

To evaluate RQ 4.3.1, we create a prompt that defines 3 strategies labeled A1, A2, and A3. Each strategy is ascribed a value 5, 10, or 20 points with each value being assigned once in the prompt context c. The model then provides the probability for all in-vocabulary completions. However, we consider only the probability of a constant evaluation word. This is repeated for each strategy option, changing only the strategy s. This measures the probability of the evaluation word given the strategy, $p(e_{word}|c,s)\ \forall s \in \mathbf{S}$. We refer to this as *counterfactual prompting*. The following is an example of such a prompt with *A1* as the evaluated strategy.

Option A1 gives 5 points. Option A2 gives 10 points. Option A3 gives 20 points. A1 is ____

Our hypothesis is that the preference, as measured by the probability of the evaluation word, will tend to be correlated with the assigned value. If the correlation is 0.3 or higher, based on *Applied Statistics for the Behavioral Sciences* (Hinkle et al., 2003), then a significant correlation is present and the LLM is considered

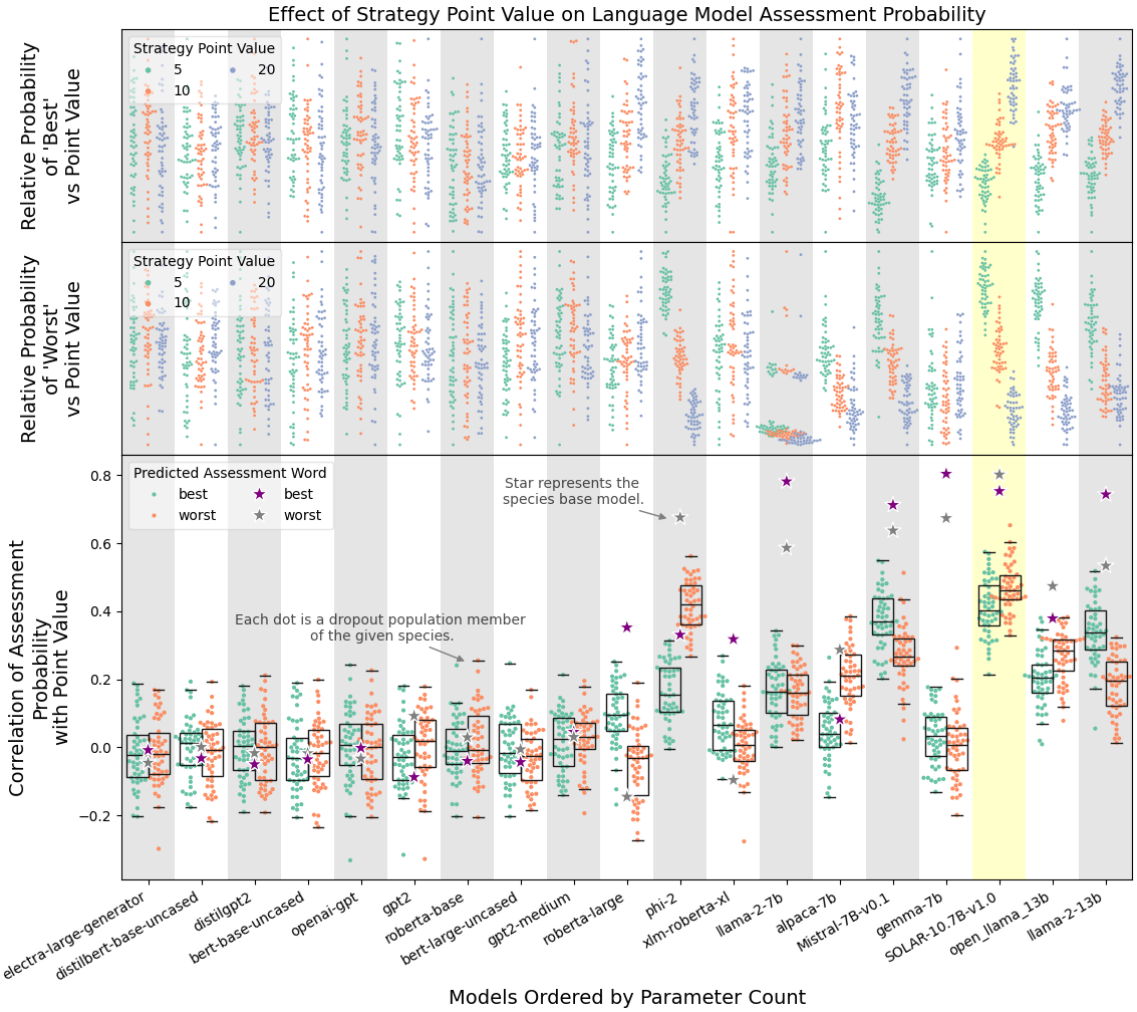

Figure 4.1: Top: Individual population member probabilities for "Best" evaluation of strategies. Middle: Individual population member probabilities for "Worst" evaluation of strategies. Bottom: Spearman's ρ for value-preference correlation and negated anti-correlation.

capable of VBP. There are 2 alternative hypotheses that this experiment must control for: preference based on ordering of the labels and preference for a label absent of value. So, we generate a prompt for every permutation of the order of labels and the assigned value. This results in 36 unique prompts. We then test the LLM preference for each strategy for each prompt permutation. This yields 108 individual calls to each model.

Additionally, we investigate if models with VBP are self-consistent across evaluation words of differing sentiment. We perform the described experiment first with a positive sentiment evaluation word (best), and then with a negative sentiment evaluation word (worst). We say that a model with positive sentiment probability that is correlated with strategy value and negative sentiment probability that is anti-correlated with strategy value have VBP and are *self-consistent*.

As mentioned previously, the effect of variation on model preference is important given the targeted HRI application domain. We use PopulationLM (Roberts et al., 2024) to construct populations for each model species tested. Models that differ on architecture, size, training data, or training task are considered different species. This approach uses Monte Carlo dropout to generate perturbed versions of the base model. Such a population is known to approximate a gaussian random process (Gal and Ghahramani, 2016). Intuitively, this means that model behaviors which are constituted in a small number of neurons, referred to as poorly supported, are likely to be ablated in the perturbed population. So, if the base model of a given species has VBP but the derived population does not, we say the model is *brittle* since variation tends to erode the behavior of interest.

Finally, to understand how model size relates to VBP and the tendency to prefer strategies based on more superficial criteria, we conduct the described set of experiments on 19 different model species with sizes varying from $< 10^8$ to $> 10^{10}$ parameters. For each base model species the generated population has 50 members.

4.3.2 Results: Value-Based Preference

In answer to RQ 4.3.1, we find that a surprisingly small number of models have VBP. In figure 4.1, the correlation of the evaluation probability and strategy point value for each of the population members (dots) as well as the species base model (stars) are shown in the bottom row. Those that do have VBP are those that have high base model correlation like Solar (Kim et al., 2023), Mistral (Jiang et al., 2023), Gemma (Team et al., 2024), Llama-2 (Touvron et al., 2023), and Phi-2 (Javaheripi et al., 2023). Among these, only Phi-2 fails to be self-consistent.

We find that, among self-consistent models with VBP, the populations are likewise self-consistent. However, the Gemma and LLama-2 populations don't exhibit VBP. Therefore, Gemma and Llama-2 are considered brittle.

4.3.3 Effects of Model Size

We investigate the effect of model size on the presence of VBP. In Figure 4.2 the model size is tellingly correlated with the model's preference for higher value strategies. From the figure it seems that model size is predictive of VBP. More precisely we conclude that sufficient model size may a necessary, though insufficient, condition for a model to learn VBP from human language data.

We further consider the effect of superficial information, like the label, on model preference. In Figure 4.3 the non-parametric Kruskal-Wallis test is used to evaluate if the probabilities assigned to a strategy are independent of the label. The null hypothesis for this test expects the medians of the groups to be equal.

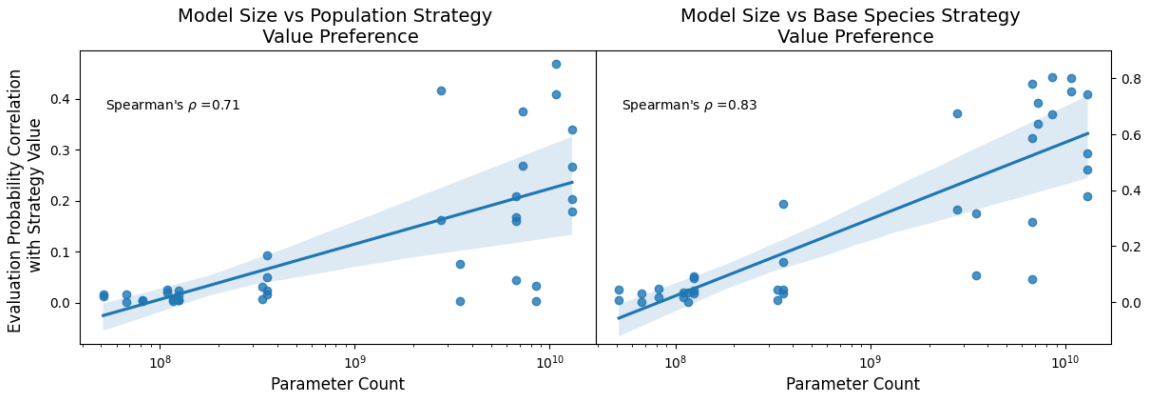

Figure 4.2: As models get larger they tend to have value-based strategy preferences. Left: Model size vs populationLM strategy VBP. Right: Model size vs base species strategy VBP.

The figure shows that, for the smaller species base models, preferences tend to be sensitive to superficial information like label. On the other hand, as the model size increases, sensitivity to the label tends to decrease.

Interestingly, it seems that preference sensitivity to label is much more correlated with model size in the base models ($\rho = 0.39$), shown on the right of the figure, as compared to the populations ($\rho = 0.06$) on the left. This shows that intra-species populations of language models may tend to be less sensitive to superficial information. This suggests reliance on superficial information is not a well supported behavior in many of the base models.

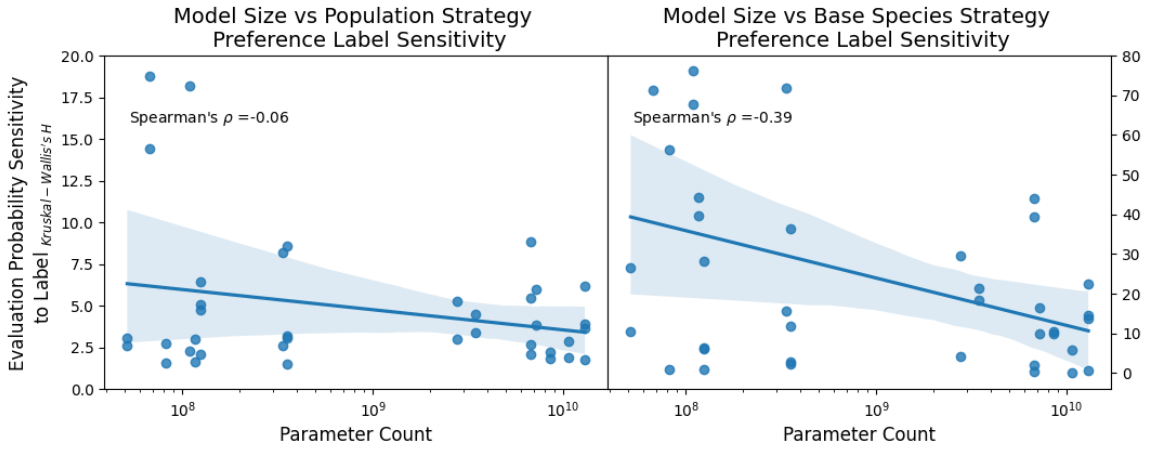

Figure 4.3: Left: Model size vs populationLM strategy preference sensitivity to label. Right: Model size vs base species strategy preference sensitivity to label.

4.3.4 Why are Solar and Mistral Not Brittle?

From these experiments we find that Solar is the current best model among those tested when required to make stable, VBP judgments, with Mistral being a close second. Though Gemma and Llama-2 base models

exhibit comparable VBP than Solar and Mistral, they are brittle. The process of dropout and proportion of affected neurons in each of the models is constant. We therefore, consider why Gemma and Llama-2 brittle while Mistral and Solar are not.

It's first important to understand the provenance of Mistral and Solar. Llama-2 was trained on 2 Trillion tokens and the authors of the technical report (Touvron et al., 2023) point out that the model had not reached training saturation. This message to the community was clear, further training of the Llama-2 model may increase the performance.

The creators of Mistral adopted the Llama-2 architecture, converted the attention to sliding window attention (SWA) (Beltagy et al., 2020) with a window of 3 tokens, and then retrained the model. Importantly, SWA does not force a model to only consider the last 3 tokens during generation. Rather, the path to access information in tokens prior to the window requires adjacent representations to package and pass the information. The resultant model outperforms Llama-2 7B and 13B substantially (Jiang et al., 2023). The number of training tokens is unknown. However, in light of the shared architecture, the comment regarding Llama-2 training saturation, and the established supremacy of Mistral, it is reasonable to assume that it has been trained on a number of tokens greater than 2 trillion.

Solar was subsequently constructed by adopting the architecture of Llama-2 and increasing the number of layers through depth upscaling (Kim et al., 2023). The initial layers of the architecture are initialized with the Mistral weights and then all layers were received additional training. So, Solar must be considered to have been trained on more tokens than Mistral. Solar does not use SWA, however, it inherited weights that were learned through SWA.

Finally, Gemma exhibits VBP that is similar to Solar, however it is more brittle than Llama-2 though it was trained on 4 times the number of tokens. We therefore surmise that the total number of training tokens and model size may tend to improve VBP. However, we hypothesize that SWA may tend to encourage distributed representations, in Mistral and Solar, that are less brittle.

4.4 Do LLMs Have Human-Like Preference in the Prisoner's Dilemma?

Having shown that some LLMs exhibit VBP, we ask whether their preferences tend to be human-like in established, empirically considered game scenarios. The prisoner's dilemma (PD) is a well known game in which two players have two strategy options. If player 1 chooses to betray player 2, then they may either receive 0 or 3 months in jail. If player 1 instead decides to be silent, they will receive either 2 or 5 months in jail. For various scenarios, the payoff matrices are shown in Table 4.1 with Player 1 being the first number in each pair.

The Nash equilibrium strategy is defined as the option that obtains the best payoff without first assuming

that the opponent will change strategy (Nash, 1951). In the PD, rational agents are typically expected to seek their own interest and choose the Nash equilibrium strategy to betray. However, in practice humans don't necessarily choose the Nash equilibrium. By choosing to be silent they can minimize the total number of months spent by either player in jail. This is known as the Pareto optimal strategy. In (Yamagishi et al., 2016), the authors did a large study on human subjects in Tokyo and show that humans tend to cooperate (choose the Pareto optimal strategy) when the stake size is low. However, when the stake size is large, humans tend to choose to betray the other player in self-interest.

LLMs have been engaged in a prisoner's dilemma in previous work (Akata et al., 2023). They found that in a non-repeated PD, GPT-4 tended to cooperate. However, as mentioned previously, the measurement they use is a type of cloze task, their work does not consider the effect of variation on the result, and it does not consider the effect of stake size. To understand if LLMs tend to have human-like preference we formulate RQ 4.4.1.

Research Question 4.4.1. *When engaged in an obfuscated prisoner's dilemma, do LLMs tend to have preferences consistent with human preference including sensitivity to stake size?*

4.4.1 Experimental Method

Given the prevalence of the PD, it is likely that is well represented in language model training data. To effectively measure the impact of stake size without encountering the canonical moral preference for cooperation, we obfuscate the PD.

The low stakes version of the PD is cast as a decision to use (betray) or not use (silent) a shared air conditioner at night. The payoff matrix is shown in the left side of Table 4.1. The high stakes version is nearly identical with the AC system exchanged for a life support system.

To evaluate model preference for each strategy, we construct a prompt which enumerates the options and the possible results. We then use counterfactual prompting to find the probability assigned to a constant evaluation word as done in the previous experiment. A full example of the prompt is available in the appendix.

We construct different versions of the prompt to ensure that all permutations of label order and strategy assignment are represented to control for alternative hypotheses. We again perform the set of experiments using both positive ("Best") and negative ("Worst") sentiment evaluation words. Finally, we perform the set of experiments using populations (N=50) of the 4 self-consistent models that exhibitied VBP: Solar, Mistral, Gemma, and Llama-2.

Table 4.1: Prisoner's dilemma payoff matrices

		Player 2					
		AC Sharing		Life Support Sharing		Months in Jail	
		Silent	Betray	Silent	Betray	Silent	Betray
Player 1	Silent	Cool, Cool	Cold, Hot	4,4	0,10	2,2	5,0
	Betray	Cold, Hot	Warm, Warm	10,0	2,2	0,5	3,3

4.4.1.1 Pythagorean Preference Relation

To construct a preference relation from the counterfactual prompting, we use the stratified population members to evaluate the possible strategies. This permits the use of the non-parametric, paired Wilcoxon rank-sum test. The null hypothesis for this test is that the distribution of observations of a single group, arising from two treatments, are not statistically different. This allows not only the characterization strategy preference but also statistical significance.

Table 4.2: Preference relation using positive and negative evaluation for preference and anti-preference.

	Strong Preference		Partial Preference				Indifference		
Best Evaluation	$L \succ M$	$M \succ L$	$L \succ M$	$M \succ L$	$L \sim M$	$L \sim M$	$L \sim M$	$L \succ M$	$M \succ L$
Worst Evaluation	$L \succ M$	$M \succ L$	$L \sim M$	$L \sim M$	$M \succ L$	$L \succ M$	$L \sim M$	$M \succ L$	$L \succ M$
Result	$L \succ M$	$M \succ L$	$L \succeq M$	$M \succeq L$	$M \succeq L$	$L \succeq M$	$L \sim M$	$L \sim M$	$L \sim M$

Inspired by work in pythagorean fuzzy preference relations for group decision making (Mandal and Ranadive, 2019), we consider that preference and anti-preference may vary independently. By measuring the probability of both the positive and negative evaluation words, we arrive at measures related to the preference and anti-preference respectively. Performing separate Wilcoxon tests on the positive and negative evaluations independently yields a measure and significance of the preference and anti-preference.

So, for strategies L and M each presented as options in context c and a positive evaluation word used as the measure, if $p(e_{pos}|c,L) > p(e_{pos}|c,M)$ tends to hold in a population, as characterized by a Wilcoxon test, then we say the population has a significant preference for L over M, or $L \succ M$. Alternatively, if $p(e_{neg}|c,L) > p(e_{neg}|c,M)$ tends to hold in a population, then we say the population has a significant anti-preference for L over M, or $M \succ L$. If the result of a Wilcoxon test fails to be statistically significant, then we say that the population has indifferent preference or anti-preference to L over M, or $L \sim M$. In Table 4.2 the possible resulting preferences are shown.

4.4.2 Results: LLM Preference in the Prisoner's Dilemma

In Figure 4.4 the probability of positive evaluation is shown in the top row and the probability of negative evaluation is shown in the bottom for all population members and all species. When the stakes are low, Solar, Mistral, and Llama-2 have a strong preference to cooperate. On the other hand, when the stakes are high, all

models have a partial preference for self-interest.

Interestingly, the Gemma population is uncertain regarding preference and anti-preference when faced with a low-stakes PD. This is most likely due to the brittleness result already discussed.

In the high stakes scenario, Solar and Mistral show an anti-preference to cooperate (silent), but they don't prefer to act in self interest (betray). A human, choosing to use a life support system and potentially shorten the life of another, or choosing to trust another not to do the same, may ultimately experience a similar preference/anti-preference dichotomy. It's not preferable to potentially shorten the life of another. However, choosing to trust another individual to not act in self-preservation may be unacceptable. Llama-2 preferred to act in self-interest with no significant reservation.

In answer to RQ 4.4.1, the results show that self-consistent, non-brittle LLMs with VBP tend to have distinctly human-like preferences in the PD, including sensitivity to stake size. This is true even when the scenario does not resemble the classical incarnation of the dilemma. We further consider the discussed nuance of these results to suggest that populations of Mistral and Solar have more human-like preference than Llama-2 when engaged in the PD.

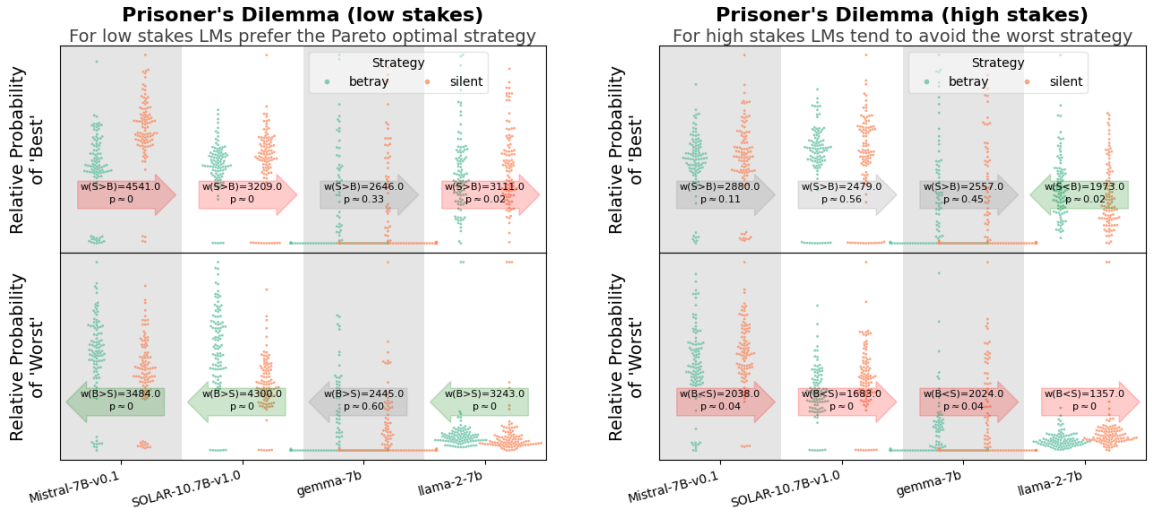

Figure 4.4: Left: LLMs in a low stakes obfuscated prisoner's dilemma prefer cooperation. Right: LLMs in a high stakes obfuscated prisoner's dilemma prefer self-interest.

4.5 Do LLMs Have Human-Like Preference in the Traveler's Dilemma?

The traveler's dilemma (TD) is an interesting game introduced in (Basu, 1994) to illuminate concrete scenarios in which humans are expected to deviate from the Nash equilibrium.

Suppose there are two strangers traveling back from vacation who have purchased the same antique. The airline breaks both antiques. The two individuals are informed independently. They are each asked the value

of the antique and are allowed to respond in the range $[2, 100]$. They are warned that out bidding the other passenger will result in a penalty.

Specifically, player A provides quote Q_A and player B provides Q_B. If $Q_A > Q_B$ then the payoff for player A will be $Q_B - 2$ and the payoff for player B will be $Q_B + 2$. The reciprocal statement is true if $Q_A < Q_B$. Lastly, if $Q_A = Q_B$ they receive the value quoted with no adjustment.

4.5.1 Human Deviation from the Nash Equilibrium

If strategy Q_a is in all cases as good as Q_b and, in at least one case, Q_a provides a better payoff, then Q_a is said to weakly dominate Q_b (Osborne and Rubinstein, 1994). In the prisoner's dilemma, quoting 99 weakly dominates quoting 100. In this case, game theorists consider 100 to be eliminated as a strategy since 99 *should be* strictly preferred. This creates a cascading elimination since, iff 100 is removed, 98 weakly dominates 99.

This elimination of weakly dominated strategies results in a canonical Nash equilibrium that predicts rational players will quote the airline 2 dollars. However, empirical studies show that humans tend to prefer strategies that are more cooperative (Becker et al., 2004) and tend to provide quotes in the mid 90s. Further, when the penalty is increased humans tend to choose strategies that closer to the Nash equilibrium (Morone et al., 2014) even though the size of penalty has no game theoretic effect on the equilibrium.

It has been argued in (Roberts, 2021) that human deviation from the Nash equilibrium suggests that humans are not certain of a preference for 99 over 100 which prevents elimination of that strategy. They show that if this is the case, and the elimination scheme is retooled to permit fuzzy elimination, then human behavior is well predicted by fuzzy elimination of weakly dominated strategies. They also suggest that the penalty size must directly effect the certainty of the preference.

We examine the behavior of self-consistent LLMs with VBP in the traveler's dilemma by evaluating the preference for 99 and 100. Specifically, we formulate RQ 4.5.1.

Research Question 4.5.1. *When engaged in a traveler's dilemma, do LLMs tend to prefer strategies closer the Nash equilibrium in response to increased penalty?*

4.5.2 Experimental Method

We again use model species populations (N=50), counterfactual prompting, and the preference relation described in Table 4.2. We provide the TD scenario and the range of options in the prompt context and a discussion of payoffs for 99 and 100. We again permute the labeling of options to control for superficial preference heuristics. We conduct the set of experiments with penalty sizes of 2 and 20. A full sample of the low penalty and high penalty prompts are available in the appendix.

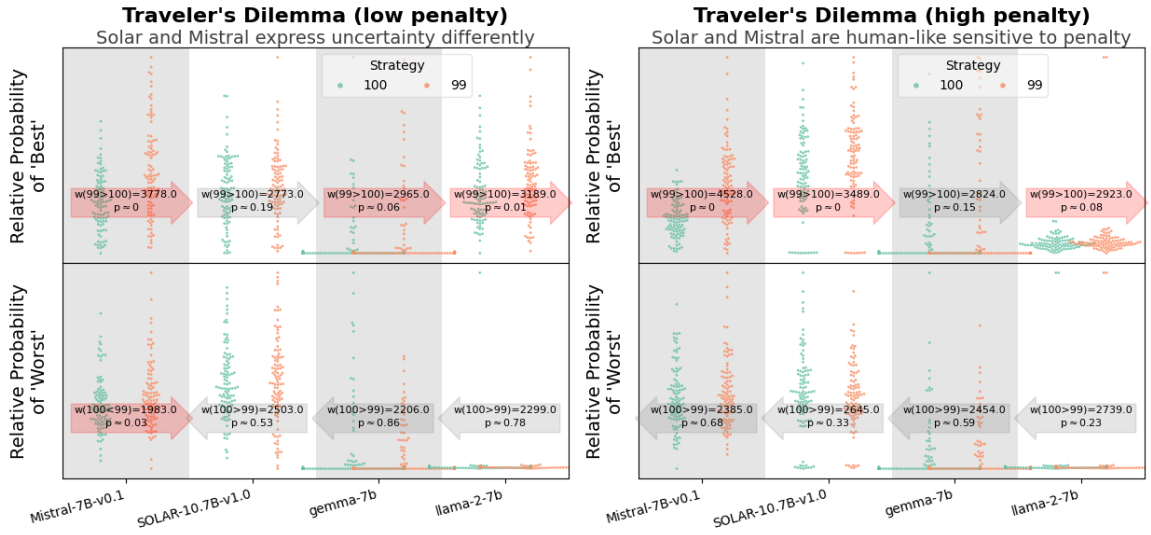

Figure 4.5: Left: LLM preference in a low penalty TD. Right: LLM preference in a high penalty TD

4.5.3 Results: LLM Preference in the Traveler's Dilemma

In the right of Figure 4.5, the low penalty scenario results are shown, and the high penalty results are shown on the left. In the low penalty scenario, Solar and Mistral are indifferent to 99 and 100, that is $99 \sim 100$. However, when the penalty size increases to 20, Solar and Misral show a partial preference for 99, $99 \succeq 100$.

In answer to RQ 4.5.1, we find that non-brittle LLMs with VBP tend to have human-like preference sensitivity to penalty size in the TD. Such indifference necessarily prevents the elimination sequence that allows a quote of 2 to emerge as the Nash equilibrium. This bidirectionally supports the hypothesis for the origin of empirical deviation from the Nash equilibrium in the TD given in (Roberts, 2021), suggesting humans may not prefer 99 over 100 even though 100 is weakly dominated.

In contrast, Gemma and Llama-2 populations showed a preference for 99 over 100 in the low penalty scenario. When the penalty size was increased, Llama-2 was essentially unaffected while Gemma became surprisingly indifferent.

4.6 Discussion

In this paper, we evaluate the ability of many LLMs to prefer strategies based on value. We control for label-based heuristics and showed that value-based preference and self-consistency tend to emerge as a function of model size and training token count. We reasoned that Solar and Mistral may be less brittle due to the effect of sliding window attention during model training. Finally, we showed that smaller base models tended to prefer strategies based on superficial heuristics like label. We then showed that Solar and Mistral additionally exhibit human-like strategy preference in both the PD and the TD. We evaluated the PD stake-size effect and

the TD penalty-size effect on model preference and found that Solar and Mistral tended to have human-like sensitivity in both cases.

Our results suggest that Solar and Mistral, among the models tested in Figure 4.1, are most appropriate for application in human-adjacent NLP tasks like HRI based on possession of strategic preferences that are similar to empirically established human strategic preferences.

In the course of this work, we established a method for robustly measuring the preference of LLM populations generated via PopulationLM (Roberts et al., 2024). And finally, we found evidence suggesting the hypothesis given in (Roberts, 2021) claiming the origin of empirical deviation from the Nash equilibrium in the TD is based on penalty-dependent indifference to weak domination of a strategy.

4.6.1 Limitations

While studying model behavior in a population does tend to reduce the prevalence of poorly supported behaviors this does not guarantee that framing effects may not effect the experimental results we have obtained. However, uncontrolled framing effects tend to adversely effect human results as well and is a common problem in economics research (Goldin and Reck, 2020).

The preference relation construction method described herein is not transitive given it utilizes the Wilcoxon rank-sum test (Lumley and Gillen, 2016). Future work studying LLM preference should consider the effect of non-transitivity. It is understood that substituting a student t-test or other method making strong assumptions regarding the data distribution would provide transitivity. However, the data generated by a population of LLMs may not conform to the distributional assumptions necessary for a parametric test to be applicable. Further, guaranteeing transitivity may be counter productive since two-sample tests that do so are necessarily reducible to univariate summary statistics (Lumley and Gillen, 2016) and humans seem to have preferences that are at times non-transitive (Alós-Ferrer et al., 2022).

Finally, and most importantly, the tests here establish that Solar and Mistral have learned human-like preferences in specific contexts. It is probable, though not established, that in some circumstances these models may have distinctly non-human strategic preferences. Proving otherwise is intractable.

4.7 Appendix

4.7.1 Counterfactual Prompting

In this paper counterfactual prompting is applied. This method of prompting is neither novel or typical. It has strong similarities to noisy channel model prompting (Min et al., 2022) which tends to improve prediction stability. Further, other works have used an equivalent measurement method in the past (Misra et al., 2021; Roberts et al., 2024). However, noisy channel prompting is developed in a significantly different context and

the other mentioned works do not provide a formal reasoning or explicit discussion of the technique. Based on theoretical considerations, we propose counterfactual prompting as a general method for performing more robust, as compared to a cloze test, measurements of language model behavior.

The typical method of measuring LLM preference over a set of options is to perform a cloze test. That is, given a prompt whose completion is a strategy, the most preferred strategy is considered the option with highest probability. That is, given some set of strategies, $\mathbf{S}$, which are included in context c, a cloze task measures $p(s|c)\ \forall s \in \mathbf{S}$. Counterfactual prompting instead measures $p(e_{word}|s,c)\ \forall s \in \mathbf{S}$, the probability of a constant evaluation word given the context and strategy.

To see the benefit, consider that each strategy option has some base rate probability given the training data. Therefore, $p(s_a|\emptyset) \neq p(s_b|\emptyset)\ \forall s_a, s_b \in \mathbf{S}$ unless $s_a = s_b$. The effect is a bias toward options with a larger base rate probability which must be overcome by the network in response to the semantic content of the context. Therefore, if the signals, probabilities assigned to a set of options in light of the context, are sufficiently small, the noise, or base rate probabilities, may rival the magnitude of the signal. So, the signal to noise ratio may be poor.

Alternatively, counterfactual prompting is theoretically more sensitive since a single evaluation word is used to measure the preference for all strategies. Therefore, there is no difference in base rate probability across strategies. This form of prompting is inspired by typical methods in sensing. Measuring a signal in a noisy environment can be done effectively if the signal is carried by sets of conductors who all experience the same noise. When a measurement is made across those conductors, the noise affects all constituents equally and leaves the signal unchanged. Measuring the probability of a consistent evaluation word ensures base-rate noise effects all strategies equally.

4.7.2 Prisoner's Dilemma

4.7.2.1 Obfuscated Low Stakes Prompt

My roommate and I share an AC. If one person uses the AC then they will get all and the other person will get none. If neither uses the AC then both will have a little. I do not know if my roommate will use the AC today, but they might. I cannot control their choice. Option1 is to use the AC. Option2 is to not use the AC. If I choose Option1 and my roommate uses the AC then I will be warm. If I choose Option1 and my roommate does not use the AC then I will be cold. If I choose Option2 and my roommate uses the AC then I will be hot. If I choose Option2 and my roommate does not use the AC then I will be cool. I want to be cold. I do not want to be hot or warm. My roommate wants to be cold and does not want to be hot or warm. Option1 is the ___.

4.7.2.2 Obfuscated High Stakes Prompt

I am in space with another person and we share a life support system. If one person uses the life support then they will get all and the other person will get none. If neither uses the life support then both will have some. I do not know if the other person will use the life support today, but they might. I cannot control their choice. Option1 is to use the life support. Option2 is to not use the life support. If I choose Option1 and my roommate uses the life support then I will live 2 years. If I choose Option1 and the other person does not use the life support then I will live 10 years. If I choose Option2 and the other person uses the life support then I will die tomorrow. If I choose Option2 and the other person does not use the life support then I will live 4 years. I want to live as long as possible. I do not want to die. My roommate wants to live and does not want to die. Option1 is the ___.

4.7.3 Traveler's Dilemma

4.7.3.1 Low Penalty Prompt

You and another person have traveled and returned with two vases. However, the airline has broken the vases. You don't know the other person. The airline will let you and the other person claim between 100 and 2 dollars for the vase. If you claim more than the other person, you will get their claim minus 2 dollars. If you claim less than the other person, you will get your claim plus 2 dollars. If you choose optionB to claim 100 dollars you are likely to get a payment of 100 dollars, and it is less likely but possible that you may get a payment between 97 and 0 dollars. If you choose optionA to claim 99 dollars you are likely to get a payment of 101 dollars, and it is less likely but possible that you may get 99 dollars or between 96 and 0 dollars. You prefer options that will get the most money. OptionA is___.

4.7.3.2 High Penalty Prompt

You and another person have traveled and returned with two vases. However, the airline has broken the vases. You don't know the other person. The airline will let you and the other person claim between 100 and 20 dollars for the vase. If you claim more than the other person, you will get their claim minus 20 dollars. If you claim less than the other person, you will get your claim plus 20 dollars. If you choose optionA to claim 100 dollars you are likely to get a payment of 100 dollars, and it is less likely but possible that you may get a payment between 77 and 0 dollars. If you choose optionB to claim 99 dollars you are likely to get a payment of 119 dollars, and it is less likely but possible that you may get 99 dollars or between 76 and 0 dollars. You prefer options that will get the most money. OptionA is ___.

References

Akata, E., Schulz, L., Coda-Forno, J., Oh, S. J., Bethge, M., and Schulz, E. (2023). Playing repeated games with large language models. *arXiv preprint arXiv:2305.16867*.

Alós-Ferrer, C., Fehr, E., and Garagnani, M. (2022). Identifying nontransitive preferences. Technical report, Working Paper.

Basu, K. (1994). The traveler's dilemma: Paradoxes of rationality in game theory. *The American Economic Review*, 84(2):391–395.

Becker, T., Carter, M., and Naeve, J. (2004). Experts playing the traveler's dilemma. Technical report, Department of Economics, University of Hohenheim, Germany.

Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Bhattamishra, S., Patel, A., and Goyal, N. (2020). On the computational power of transformers and its implications in sequence modeling. *arXiv preprint arXiv:2006.09286*.

Binz, M. and Schulz, E. (2023). Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

Fan, C., Chen, J., Jin, Y., and He, H. (2024). Can large language models serve as rational players in game theory? a systematic analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17960–17967.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Goldin, J. and Reck, D. (2020). Revealed-preference analysis with framing effects. *Journal of Political Economy*, 128(7):2759–2795.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2020). Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Hinkle, D. E., Wiersma, W., Jurs, S. G., et al. (2003). *Applied statistics for the behavioral sciences*, volume 663. Houghton Mifflin Boston.

Javaheripi, M., Bubeck, S., Abdin, M., Aneja, J., Bubeck, S., Mendes, C. C. T., Chen, W., Del Giorno, A., Eldan, R., Gopi, S., et al. (2023). Phi-2: The surprising power of small language models. *Microsoft Research Blog*.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. (2023). Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Kim, D., Park, C., Kim, S., Lee, W., Song, W., Kim, Y., Kim, H., Kim, Y., Lee, H., Kim, J., et al. (2023). Solar 10.7 b: Scaling large language models with simple yet effective depth up-scaling. *arXiv preprint arXiv:2312.15166*.

Lumley, T. and Gillen, D. L. (2016). Characterising transitive two-sample tests. *Statistics & Probability Letters*, 109:118–123.

Mandal, P. and Ranadive, A. (2019). Pythagorean fuzzy preference relations and their applications in group decision-making systems. *International Journal of Intelligent Systems*, 34(7):1700–1717.

Min, S., Lewis, M., Hajishirzi, H., and Zettlemoyer, L. (2022). Noisy channel language model prompting for few-shot text classification. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5316–5330, Dublin, Ireland. Association for Computational Linguistics.

Misra, K., Ettinger, A., and Rayz, J. T. (2021). Do language models learn typicality judgments from text? *arXiv preprint arXiv:2105.02987*.

Morone, A., Morone, P., and Germani, A. R. (2014). Individual and group behaviour in the traveler's dilemma: An experimental study. *Journal of Behavioral and Experimental Economics*, 49:1–7.

Nash, J. (1951). Non-cooperative games. *Annals of Mathematics*, pages 286–295.

OpenAI (2023). Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Osborne, M. J. and Rubinstein, A. (1994). *A Course in Game Theory*. The MIT Press.

Pérez, J., Marinković, J., and Barceló, P. (2019). On the turing completeness of modern neural network architectures. *arXiv preprint arXiv:1901.03429*.

Roberts, J. (2021). Finding an equilibrium in the traveler's dilemma with fuzzy weak domination. In *2021 IEEE Conference on Games (CoG)*, pages 1–5. IEEE.

Roberts, J. (2024). How powerful are decoder-only transformer neural models?

Roberts, J., Moore, K., Wilenzick, D., and Fisher, D. (2024). Using artificial populations to study psychological phenomena in neural models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):18906–18914.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Suri, G., Slater, L. R., Ziaee, A., and Nguyen, M. (2023). Do large language models show decision heuristics similar to humans? a case study using gpt-3.5. *arXiv preprint arXiv:2305.04400*.

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., et al. (2024). Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Trott, S., Jones, C., Chang, T., Michaelov, J., and Bergen, B. (2023). Do large language models know what humans know? *Cognitive Science*, 47(7):e13309.

Ullman, T. (2023). Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, S., Liu, C., Zheng, Z., Qi, S., Chen, S., Yang, Q., Zhao, A., Wang, C., Song, S., and Huang, G. (2023). Avalon's game of thoughts: Battle against deception through recursive contemplation. *arXiv preprint arXiv:2310.01320*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., and Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Yamagishi, T., Li, Y., Matsumoto, Y., and Kiyonari, T. (2016). Moral bargain hunters purchase moral righteousness when it is cheap: within-individual effect of stake size in economic games. *Scientific Reports*, 6(1):27824.

Zhong, W., Cui, R., Guo, Y., Liang, Y., Lu, S., Wang, Y., Saied, A., Chen, W., and Duan, N. (2023). Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.

CHAPTER 5

How Powerful are Decoder-Only Transformer Neural Models?

5.1 Introduction

Transformer models have achieved state of the art performance on many tasks since their introduction in (Vaswani et al., 2017). However, the provenance of their capabilities is not yet well understood. While some evidence suggests that capabilities may emerge as a function of model size (Wei et al., 2022), continually increasing the number of parameters consumes significant energy posing risk to the environment (Rillig et al., 2023). In this paper, we provide a proof that suggests that decoder-only transformer language models, like GPT-x, do not require the vast number of layers, attention heads, and parameters typical in current implementations to achieve powerful computation.

The transformer architecture introduced in (Vaswani et al., 2017) is based on a denoising auto-encoder scheme. Interestingly, the work on these *vanilla* transformers has largely been eclipsed by variations of the transformer like that in (Liu et al., 2018), (Radford et al., 2018) (GPT), and (Devlin et al., 2018) (BERT). Much of this may be due to GPT-4 (OpenAI, 2023) and its predecessors which have captured public attention. While the exact architecture of GPT-4 is closed source, GPT-3 and GPT-2 are known to be decoder-only transformer architectures (Radford et al., 2019; Brown et al., 2020).

Work regarding the computational expressivity of the *vanilla* transformer has proven it to be Turing complete (Pérez et al., 2019; Bhattamishra et al., 2020). However, in 5.2.1.2 we show that these proofs do not naturally extend to the decoder-only transformer architecture. Further, no formal evaluation of the computational expressivity exists for the decoder-only transformer architecture. In this paper:

1. We show that the decoder-only transformer architecture is Turing complete
2. We show that this result holds even for single layer, single attention head decoder-only architectures
3. We establish a minimum vector dimensionality, relative to the token embedding size, necessary for Turing completeness
4. We classify decoder-only transformer models as a causal variant of *B machines* (Wang, 1957)
5. We provide an explanation for parameter inefficiency

Based on our results, we suggest that decoder-only architectures do not necessarily require the large number of parameters typically allocated to perform the necessary computations to support complex NLP

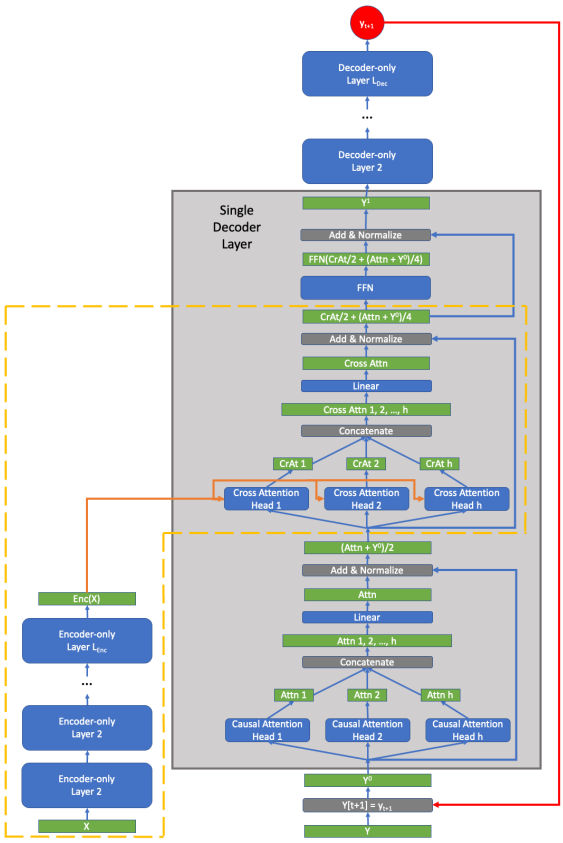

Figure 5.1: Vanilla Transformer Architecture. The yellow dashed line surrounds the sections removed to create a Decoder-only Transformer model.

functionality. Rather, the number of parameters may be necessitated by the interaction between the language modeling task and the architecture. This suggests that minor architectural adjustments could permit more parameter-efficient future models.

5.2 Background

5.2.1 Disambiguating Decoder-Only Transformer Models

Following after (Liu et al., 2018), the creators of GPT refer to their architecture as a decoder-only transformer. Seemingly in contrast, the creators of BERT refer to it as an encoder-only model (Devlin et al., 2018). This decoder-only/encoder-only architecture dichotomy is somewhat misleading as the two are architecturally identical as can be seen in 5.2. The differentiation lies in how the models execute. BERT and other encoder-only architectures are incapable of recursion. On the other hand, at each time step $t > 0$, decoder-only architectures have access to their own outputs from all previous time steps. This permits the model to be trained to generate output auto-regressively.

For brevity we follow previous conventions and refer to the transformer architecture presented in (Vaswani

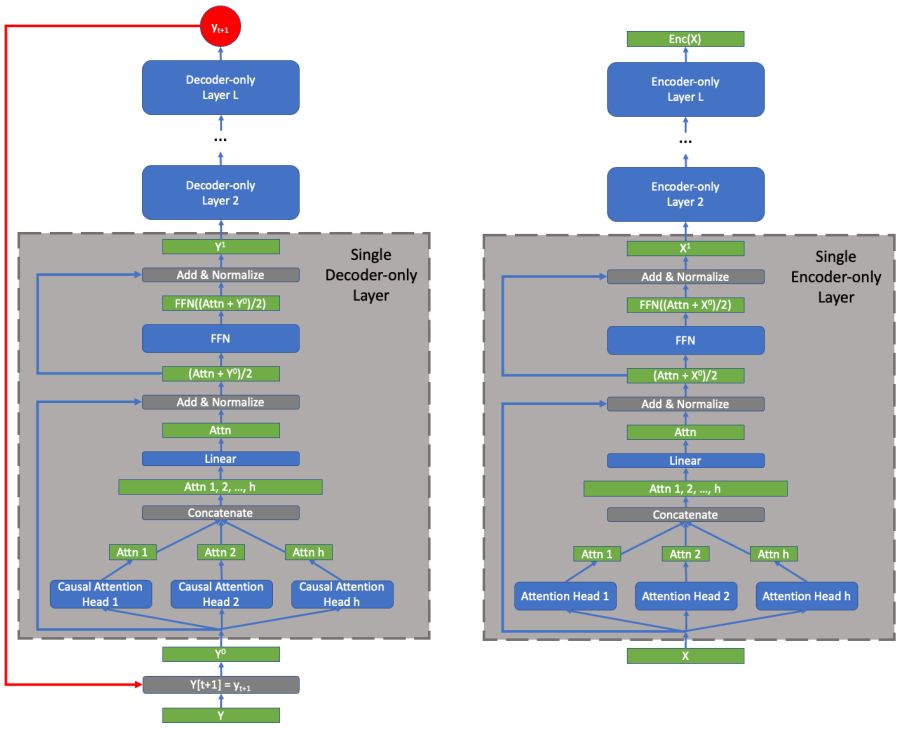

Figure 5.2: Decoder-only (left) and Encoder-only (right) Transformer Architectures. Green boxes are sequences of vectors with the width of the box representing relative sequence length. Red denotes a single vector. Gray and blue boxes denote simple and compound operations respectively.

et al., 2017) as the *vanilla* transformer, shown in 5.1. Encoder-only transformer architectures do not possess a decoder. Similarly, decoder-only models do not have an encoder. These architectures are shown in 5.2. Notice, in the case of encoder-only models, disconnection at the encoder output is sufficient to unambiguously define the modification to the vanilla transformer architecture. This is not the case for decoder-only architectures.

5.2.1.1 Modifying the Vanilla Transformer to form a Decoder-only Model

To create a decoder-only model, the vanilla architecture is modified in two ways. First, the connection to the encoder is removed. Second, the cross-attention which allows the decoder to conditionally attend to the encoder output at each layer of the decoder is eliminated. These, along with the entire encoder, are surrounded by a dashed yellow line in 5.1 to visualize what is eliminated. As mentioned previously, this superficially suggests that encoder-only and decoder-only architectures are identical as seen in 5.2.

5.2.1.2 Differentiating Encoder-only and Decoder-only Models

Decoder-only models have three necessary characteristics which are derived from their function in the vanilla transformer. The decoder must (1) provide a means of auto-regressively predicting the next token based on

the tokens generated so far given the encoder input as contextualization. In 5.2 this is shown as the recursive red connection mapping the output vector back into the last element of the input sequence of vectors. To be suited to this task, decoder-only models must (2) not see future values when evaluating a query on the input sequence of vectors. This is why decoder-only models are often referred to as causal language models (CLM). In 5.2, we refer to the decoder attention heads as causal attention heads rather than masked attention heads as they are called in (Vaswani et al., 2017). The model must be (3) trained to predict the next token given the current input sequence of vectors. This training method coupled with recursion allows decoder-only models to auto-regressively generate arbitrarily long (up to the max size of the input vector sequence) sequences.

If any of the above are violated, the model can't be reasonably considered a decoder-only model as it is no longer capable of auto-regressive next token prediction.

5.2.2 Related Theoretical Work on Transformers

Transformers were shown to be Turing complete first in (Pérez et al., 2019) with a simpler approach to the proof given in (Bhattamishra et al., 2020). The latter is based solely on the ability of the transformer to simulate arbitrary RNNs which are known to be Turing complete (Siegelmann and Sontag, 1992). This latter work also considered the contribution of the various architectural elements to the computational power. In their construction, they find the computational universality of the transformer is maintained even if the encoder acts essentially as an identity operator for the appropriate input. All significant computation, beyond input presentation, is handled exclusively in the decoder and FFN. However, in both proofs, the encoder is a necessary component without which the Turing completeness result does not hold.

Hahn shows that softmax based attention is often well approximated by the hardmax function (Hahn, 2020). They further show that one can apply input restrictions to transformers such that PARITY is unrecognizable in a single feedforward encoder pass regardless of the number of layers. However, their work assumes the number of computations is bounded by the length of a bounded length input.

In (Yun et al., 2019), the authors studied encoder-only architectures and showed that they were capable of universal function approximation. For this to be the case, the attention mechanism of the encoder-only architecture must be sufficient to provide the FFN with access to all subsets of the input field. Or to put this in terms familiar to a convolutional system, the attention mechanism must be capable of implementing any arbitrary feature map. This result is also important to the theoretical understanding of decoder-only transformer architectures as is clear in 5.2. Specifically, this implies that decoder-only models are universal function approximators for the n^{th} attention query in the L^{th} layer given an input sequence of length n. However, this does not prove Turing completeness.

It is reasonable to believe universal function approximation may be grounds for expecting Turing com-

pleteness to hold due to the progression of the literature for ANNs which began by showing universal function approximation (Hornik et al., 1989) and then progressed, through the addition of recursion, to Turing completeness (Siegelmann and Sontag, 1992). Further, it is intuitive based on the recursive capability of decoder-only models coupled with universal function approximation, as a model which can compute any partial recursive function is necessarily Turing complete (Turing, 1937). However, this would require that the computational class which includes primitive functions with composition and minimisation (Neto et al., 1997) be equivalent to the class of universal function approximators. Interestingly, the equivalency of these classes has never been addressed, leaving this an open question.

The only research regarding the computational expressivity of decoder-only transformer models (at the time of writing) is that of (Schuurmans, 2023). They recently considered the computational power of memory augmented language models. They showed that, when augmented by a memory module which is not part of the typical decoder-only transformer architecture, the model is Turing complete. To date, no work in the literature has addressed the computational power of typical decoder-only transformer models.

5.2.3 Required Conventions Inherited from Vanilla Transformers

The following are not architectural or training limitations, and are instead conventions that could be relaxed by future transformer architectures. However, we choose to evaluate the computational expressiveness of the typical decoder-only transformer model in common use.

First, the input embedding and output embedding used in the decoder must be identical. This permits the model output to be directly appended to the input vector sequence. Implicitly, this means decoder-only models can't have orthogonal input and output dimensions in the context vector. This is an important point as the proof method from (Bhattamishra et al., 2020) requires orthogonality which was permitted by cross-attention. However, cross-attention is removed in the decoder-only model as seen in 5.1.

Second, the input dimension of the FFN(s) must have the same dimensionality as the model dimension ie. the dimensionality of a vector in the input sequence. This disallows *sparsification* in the latent space which could be used to create a FFN input dimensionality greater than the model dimensionality. However, this does not require that the dimensionality of the model, d_{model}, be equal to the embedding dimensionality, d_{embed}.

5.3 Definitions & Approach

We modify the formalism established in (Pérez et al., 2019) and used in (Bhattamishra et al., 2020) for theoretical transformer analysis to be appropriate for our analysis of decoder-only architectures.

5.3.1 Embedding & Position

Transformers embed inputs as higher dimensional vectors via a base embedding f_b. So, for a vocabulary Σ with cardinality m, $f_b : \Sigma \to \mathbb{Q}^{d_b}$ where d_b is the number of dimensions in the embedding.

The Turing complete proof method will require that the transformer recognize the RNN stop token. Therefore, we define the embedding for the end symbol \$ such that $f_b(\$) = \mathbf{1}^{d_b}$.

In most transformer architectures the embedding is supplemented with positional information (whether explicitly defined or learned). Here we define the positional encoding as $pos : \mathbb{N} \to \mathbb{Q}$. So, for a vector $\mathbf{S}_k = (\sigma_1, ..., \sigma_k)$ with $\sigma_k \in \Sigma$ for all $k \geq 1$, the embedding with position of $\mathbf{S}_k$ is given by $(f_b(\sigma_1) + pos(1), ..., f_b(\sigma_k) + pos(k))$. The dimensionality of the combined token and position embedding is d_{embed}.

5.3.2 Decoder-only Transformer Architecture

A single layer decoder-only transformer is comprised of multi-headed attention followed by a feed forward network as seen in 5.2. It takes as input a sequence $\mathbf{Y} = (\mathbf{y}_1, ..., \mathbf{y}_k)$ of vectors where $k \geq 1$. The output of any single layer is likewise a sequence of vectors $\mathbf{Z} = (\mathbf{z}_1, ..., \mathbf{z}_k)$.

As previously mentioned, all $\mathbf{y} \in \mathbf{Y}$ and $\mathbf{z} \in \mathbf{Z}$ must have dimensionality d_{model}. However, d_{embed} is not required to be equal to d_{model}. We choose to include additional space in d_{model} such that the overall representation is sparse. Specifically, $d_{model} = 2 \cdot d_{embed} + 3$. The details of this choice are discussed in the proof.

The full decoder-only transformer architecture is formed by a stack of L layers, each composed of a single layer decoder. The output of a single execution of the model is a single vector $\mathbf{z}_k^L$, where superscript L denotes the L^{th} layer. This vector is then directly appended to $\mathbf{Y}$ such that $\mathbf{y}_{k+1} = \mathbf{z}_k^L$. The output of the previous execution is appended to the input of the subsequent execution, creating recursion.

The model will recursively execute continuously until a stopping criteria is met. Typically, the model is allowed to execute until a special token embedding is output by the model. After execution terminates, the size of the output sequence will be $|\mathbf{Y}| = k + N$ where N is the number of executions. The sub-vector $(\mathbf{y}_{k+1}, ..., \mathbf{y}_{k+N})$, referred to as the *response*, is the complete output of the model given the original *prompt* contained in $(\mathbf{y}_1, ..., \mathbf{y}_k)$.

5.3.3 Self-Attention

Every layer in 5.2 has one or more causal, self-attention, heads which filter the prompt to attend to the germane portions. Each attention head possesses functions $Q(\cdot)$, $K(\cdot)$, and $V(\cdot)$ which apply a linear transformation to each $\mathbf{y} \in \mathbf{Y}$. This results in a sequence of query vectors $\mathbf{Q}$, sequence of key vectors $\mathbf{K}$, and sequence of value vectors $\mathbf{V}$.

Each head creates a filtered view of the layer input given each query. Value vectors in $\mathbf{V}$ are chosen using the query vector $\mathbf{q}$, the sequence of keys $\mathbf{K}$, and scoring function $f^{att}(\mathbf{q},\mathbf{k})\ \forall \mathbf{k} \in \mathbf{K}$. The scoring function is the dot product of the vectors combined with a non-linear function (Vaswani et al., 2017).

Specifically, $\mathbf{q}$ attends to $\mathbf{V}$ according to an attention vector $\mathbf{a} = hardmax(\alpha_1,...,\alpha_n)$ with $\alpha_i = f^{att}(\mathbf{q},\mathbf{k}_i)$ for all $1 \leq i \leq n$. Then, the $\mathbf{q}$ attention on $\mathbf{V}$ is $\langle \mathbf{a},\mathbf{V} \rangle$. This self-attention is compactly referred to as $Att(\mathbf{q},\mathbf{K},\mathbf{V})$.

In (Vaswani et al., 2017), softmax is used. However, hardmax is used in our case to ensure all outputs are rational. Specifically, for a vector $\mathbf{x}$ with m maximum values, $hardmax(\mathbf{x_i}) = 1/m\ \forall x_i \in \mathbf{x}$ iff x_i is a maximum, else $hardmax(\mathbf{x_i}) = 0$.

In the case of multiple attention heads, each of these filtered views are concatenated and then agglomerated. Agglomeration is necessary because the concatenation step may produce a representation which no longer has dimensionality d_{model}. To return to d_{model}, a linear transformation using a set of weights, W^l, with dimensionality $d^v_{l,H}$xd is applied. The concatenation and linear transform are referred to compactly as $Conn(\cdot)$.

5.3.4 Feed Forward Network

The feedforward network, referred to as $O(\cdot)$, is fully connected and parameterized by θ. The output is $\mathbf{Z} = (\mathbf{z}_1,...,\mathbf{z}_k)$.

5.3.5 Single Layer Decoder-Only Models

The following set of equations fully characterizes the function of a single layer decoder-only transformer model. Notice that the output is a sequence of vectors.

$$\mathbf{p} = Att(Q(\mathbf{y}),K(\mathbf{Y}),V(\mathbf{Y})) \tag{5.1}$$

$$\mathbf{r} = Conn(\mathbf{p}) + \mathbf{y} \tag{5.2}$$

$$\mathbf{z} = O(\mathbf{r} : \theta) + \mathbf{r} \tag{5.3}$$

The set of equations characterizing a single layer are compactly referred to as $Dec_l(Y_l;\theta_l)$, with l denoting the layer.

5.3.6 Multi-Layer Decoder-Only Models

A multi-layer decoder-only transformer has one or more additional layers which take the output sequence generated by the previous layer as as input.

The output sequence of vectors from layer l is then referred to as Y^l and becomes the input to layer $l+1$. The output equation becomes $Y^{l+1} = Dec_l(Y_l; \theta_l)$, with $\mathbf{Y}_0 = \mathbf{Y}$. The output of a model is a single vector, the k^{th} element of the output vector for the last layer.

5.3.7 Proof Approach

Our approach to proving Turing completeness, following the example of (Bhattamishra et al., 2020), is to show that a decoder-only transformer architecture is capable of simulating the computations performed by an RNN. Based on the work of (Siegelmann and Sontag, 1992), RNNs are known to be at least as computationally expressive as Turing machines. Therefore, if a decoder-only transformer model can simulate an arbitrary RNN, then the decoder-only transformer architecture is at least as computationally expressive as a Turing machine.

Just as in (Bhattamishra et al., 2020) we will say that an RNN is simulated if (1) at each time step the input vector to the neural network contains the input x_t, (2) at each time step the input vector to the neural network contains the hidden state h_t, and (3) the decoder-only model stops at the same time step as the RNN.

To simulate an RNN via a decoder-only transformer architecture we use the decoder to implement recursion as has been done previously for vanilla transformers. However, our construction is different in that decoder-only transformers do not have an encoder. Therefore, we will provide the input to the model as the *prompt* and the *response* will be appended until execution terminates. It is clear that $\mathbf{Y}$ will always contain h_t and x_t for all timesteps. We will show by construction that self-attention, a feedforward neural network, and recursion via the decoder-only transformer is sufficient to attend to and present h_t and x_t to the FFN for all t and simulate an arbitrary RNN.

5.4 RNN Simulation by Decoder-Only Transformer

In this section we prove that there exists a single-layer, single-attention head, decoder-only transformer which may simulate any RNN. Some details are encapsulated in theorems below the proof body. In the subsequent sections we give a detailed, intuitive explanation of the proof and discussion of the implications and limitations.

5.4.1 Proof

Consider a decoder-only transformer with a single layer and single attention head in that layer.

Before the first execution of the network, the input sequence of vectors, $\mathbf{Y}$, contains the *prompt* (inputs to the RNN) in the form $\mathbf{y}_i = [f_b(\sigma_i), 0^{d_{embed}}, i, t, stop]$ with each $\mathbf{y}_i \in \mathbf{Y}$ having dimensionality d_{model}. The value of i, t, and *stop* for $i \leq k$ are *pos*, 0, and 0 respectively. The penultimate element in the *prompt*, $\mathbf{y}_{k-1}$, has $\sigma_{k-1} = \$$, and the last element, $\mathbf{y}_k$, has $\sigma_k = 0^{d_{embed}}$, the RNN start token.

Appropriate Q, K, and V linear transforms are applied to each element of $\mathbf{Y}$ such that $\mathbf{q}_i = \mathbf{y}_i$, $\mathbf{k}_i = [0^{d_{embed}}, 0^{d_{embed}}, 1, -1, 0]$, and $\mathbf{v}_i = \mathbf{y}_i$. Therefore, $\langle \mathbf{q}_i, \mathbf{k} \rangle = i - t$. The existence of such a Q, K, and V is trivial.

By application of the nonlinear function, $f^{att}(\mathbf{q}, \mathbf{k}_i)$, the attention on each $v \in \mathbf{V}$ is $\alpha_k = -|i - t|$. Therefore, $hardmax(\mathbf{V}) = 1$ when $i = t$ and $0 \; \forall i \neq t$. Therefore, $Attn(\mathbf{q}_{i=t}, \mathbf{K}, \mathbf{V}) = \mathbf{x}_{i=t}$. Therefore, the t element in the query vector selects the $i = t^{th}$ element from the *prompt*.

To generate the t^{th} element of *prompt*, the query $\mathbf{q}_{k+t} = Q(\mathbf{y}_{k+t})$ is used. The model will execute a total of N times such that $t = 0...N$.

Notice the first execution has $\mathbf{q}_{k+t} = [0^{d_{embed}}, 0^{d_{embed}}, i = pos(k), t = 0, stop = 0]$ and, by application of the agglomeration and residual connection as described in 5.4.4, the vector presented to the FFN will be $[h_t = f_b(\sigma_k), x_t = f_b(\sigma_t), i, t, stop]$. The FFN will output the vector $\mathbf{y}_{k+t+1} = [h_{t+1}, 0^{d_{embed}}, i = k + t + 1, t = t + 1, stop]$ which is appended to the sequence $\mathbf{Y}$. Therefore, for all executions $t > 0$, the vector presented to the network will be $[h_t = f_b(\sigma_{i=k+t}), x_t = f_b(\sigma_{i=t}), i, t, stop]$.

As proved in 5.4.3 and 5.4.2 there exists an FFN such that once the stop token, $f_b(\$)$, has been encountered the output of the FFN for all subsequent time steps will be $stop = 1$ and the value $x_t = f_b(\sigma_{i=t})$ will be overridden in latent space such that for all $t > k$, $x_t = x_k = f_b(\$)$ due to 5.4.1.

At all time steps the FFN will be presented with x_t, h_t, and will terminate based solely upon the weights of the RNN. Therefore, there exists a decoder-only transformer which may simulate any RNN.

5.4.2 Theorems

Theorem 5.4.1 (Single Network replacement of Cascaded Networks)**.** *For any pair of fully connected feed forward neural networks (FFNs) such that the outputs of the first are fed into the inputs of the next, there exists a single FFN whose outputs will be identical to the outputs of the second network.*

By construction, the output weights can be directed into the input of the subsequent network and stored in a single set of network connection matrices such that a single network is created. The outputs of the first network in the cascade become a latent space within the combined network.

Theorem 5.4.2 (FFN Override Input)**.** *Given any neural network with inputs $x_1, ..., x_k$, outputs $O = o_1, ..., o_k$, and n_l neurons in l hidden layers. We may add an input x_{k+1} and neuron $n_l + 1$ to hidden layers $1...l$ such that an arbitrary subset $o' \in O$ are overridden by the added neurons.*

All weights from input x_{k+1} to neurons $1,...,n_1$ are set to zero. All weights from inputs $x_1,...,x_k$ to neuron n_1+1 are zero. The weight from input x_{k+1} to neuron n_1+1 are set to infinity.

In each layer $l>1$, neuron n_l+1 has connections set to zero for all neurons $1,...,n_{l-1}$. And in each layer $l>1$, n_l+1 has connections set to infinity for neuron $n_{l-1}+1$. This forms a column of mutually connected neurons.

An arbitrary subset of outputs $o' \in O$ may be chosen which are to be affected by the added column of neurons. The weights connecting neuron n_l+1 in hidden layer l to each output in o' are set to infinity and the weights connecting n_l+1 to each output in $O \setminus o'$ are set to zero.

For all neurons n_l+1 in all layers, the bias value is set to zero. Therefore, when the input $x_{k+1}=0$, the original function of the network is left unchanged. When $x_{k+1}=1$, the value of each output in o' is forced to be the max activation function value.

Theorem 5.4.3 (Recognize the stop token)**.** *Given a stop token* \$ *that is embedded as a vector with k elements each equal to* 1 *and presented to a neural network along inputs* $x_1,...,x_k$*, a neuron may be defined such that the output is non-zero only for inputs that are* ε *close to the stop token embedding.*

Since the stop token is defined as a vector of ones, for any token presented, the output of the neuron is zero when $k-\Sigma_{i=1}^{k} x_i > \varepsilon$ and is greater than zero for all other inputs so long as the bias is $b=\varepsilon-k$. By setting the output weight of the neuron to be a large value, any non-zero output will result in saturation of downstream neurons with non-zero connecting weights. Therefore, an output that represents whether the stop token has been presented will have a max activation function value iff the input along $x_1,...,x_k$ is within ε of \$.

Theorem 5.4.4 (Compression of $\mathbf{x}_t$ and $\mathbf{h}_t$ into $\mathbf{r}_t$)**.** *Given a token base embedding with dimensionality* d_{embed}*,* $\mathbf{x}_t$ *and* $\mathbf{h}_t$ *may be losslessly compressed into* $\mathbf{r}_t$ *when each have dimensionality d.*

Recall the dimensionality of $V(\mathbf{x}_t)$ with $\mathbf{x}_t \in \mathbf{Y}$ is not related to d. Rather, $V: \mathbb{Q}^d \rightarrow \mathbb{Q}^{d_{embed}}$ such that $V(\mathbf{x}_t)=\sigma_t$ with σ_t being a token in Σ. Then, by matrix multiplication with W defined as:

$$\begin{array}{c} \\ \begin{pmatrix} 0 & \dots & 0 & 1 & 0 & \dots & 0 & 0 & 0 \\ \vdots & \ddots & 0 & 0 & \ddots & 0 & 0 & 0 & \vdots \\ 0 & 0 & 0 & \dots & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \\ (d_{embed}, d_{model}) \end{array} \begin{array}{c} (1, d_{embed}) \\ \\ \\ \\ (1, d_{model}) \end{array}$$

The resulting vector is $Conn(\mathbf{p}_t)=[0^{d_{embed}}, \mathbf{x}_t, 0, 0, 0]$. Finally, by applying the residual connection we have $\mathbf{r}_t=[\mathbf{h}_t, \mathbf{x}_t, i, t, stop]$.

5.5 Proof Explanation

To accomplish RNN simulation, an attention head is used to select the appropriate input from the *prompt* in $\mathbf{Y}$. The attention head and agglomeration weights shift the embedded representation of the input into an empty area of the model embedding. Then, the residual connection sums the input vector with the attention representation. This results in h_t and x_t in a single vector of size of d_{model}. This vector is then presented to the FFN which contains the RNN weights as well as cascaded supplementary functions.

5.5.1 Vector Elements

Recall the base embedding has dimensionality d_{embed}. As discussed previously, the input dimension of the FFN must be $2 \cdot d_{embed} + 3$. From the requirements inherited from transformer conventions, the model dimension must be equivalent to the input dimension of the FFN. So, we choose $d_{model} = 2 \cdot d_{embed} + 3$. Therefore, each $\mathbf{y} \in \mathbf{Y}$ is composed as $\mathbf{y}_i = [f_b(\sigma_i), 0^{d_{embed}}, i = pos, t, stop]$. $f_b(\sigma_i)$ is the base embedding of the token in position i^{th} position. $0^{d_{embed}}$ is the unused space to permit simulataneous presentation of x_t and h_t to the FFN. The sequence position of σ_i is stored in i and the execution time step is written by the FFN to t.

5.5.2 Attention

A single atttention head attends to $\mathbf{y}_i$ where $i = t$. This input value is referred to as x_t as this is the value which would be presented to an RNN at time t.

The attention head will return x_t with size d_{embed}. By application of a linear transformation, W^l, x_t is right shifted $|d_{embed}|$ elements and padded with zeros to have dimension d_{model}. Finally, via the residual connection and normalization, the resulting $\mathbf{r}_t$ from 5.2 is $\mathbf{r}_t = [h_t, x_t, i, t, stop]$, proved in 5.4.4.

Note that for all $t > k$, the attention head will select a value from the *response*. If unaddressed, this would prevent RNN simulation as only the *prompt* contains RNN input. However, as explained, when $t > k$ the stop token will have been seen and the FFN will ignore the value presented by the attention head by overriding it with the stop token. As an alternative construction, the position encoding could be set to zero for all vectors in the *response* generated by the model as this would result in the stop token being attended to for all $t > k$. However, we avoid this solution as it is a significant deviation from typical models.

A similar method for selection of the t^{th} element of $\mathbf{y}$ is used in (Bhattamishra et al., 2020). However, in their construction the attention head is performing cross attention rather than causal, self attention 5.1. This important difference means that their construction does not apply to decoder-only transformer models.

5.5.3 FFN Operations

The FFN instantiates the weights of the RNN. However, the FFN has three additional functions. The FFN (1) recognizes the RNN stop token and (2) overrides the x_t provided by the attention head with the stop token if the $stop = 1$. Lastly, the FFN (3) acts as a counter which generates the execution timestep, $t+1$, based on t in the previous input vector.

The stop token recognition, override function, and RNN weight instantiation are each proved possible for standalone networks. However, by considering each of the individual networks as cascaded, there exists a single network which may implement these three functions in series.

5.5.4 Summary

At each time step, the transformer FFN is presented with x_t and h_t. Further, h_{t+1} will generate the RNN stop token at the same time step as the RNN. This is because the RNN weights are a proper subset of the FFN weights and they have identical access to x_t and h_t as would occur in an RNN. There necessarily exists a decoder-only transformer capable of simulating an arbitrary RNN and thus the class of decoder-only transformer models is shown to be at least as computationally expressive as RNN models. Therefore, there exists a computationally universal decoder-only transformer.

5.5.5 Assumptions & Limitations

There are 2 main assumptions required by this proof which limit applicability to general decoder-only models.

First, the attention mechanism here uses hardmax as opposed to the typically used softmax. This assumption is similar to prior work in theoretical transformer analysis (Pérez et al., 2019; Bhattamishra et al., 2020) and is necessary to ensure values are kept rational which is not the case for softmax. Additionally, (Hahn, 2020) suggests that transformer softmax attention heads may focus attention on high scoring context and learn behavior that is well approximated by hard attention.

Second, this work inherits the assumptions made in the proof of RNN Turing completeness. For the proof of RNN computational universality in (Siegelmann and Sontag, 1992) to hold, infinite precision, infinite output space, and value rationality are required.

These assumptions are typical in theoretical work regarding the transformer architecture. However, future work should seek to characterize transformer computational expressivity under relaxed assumptions.

5.6 Discussion

5.6.1 Relationship Between Model Dimensionality and Turing Completeness

Recall the requirements discussed in 5.2.3. An interesting consequence of these requirements is that, for a decoder-only transformer to be Turing complete, it must have dead space in the model dimension. That is, it must satisfy $d_{model} > d_{embed}$. This dead space is necessary to present both the last output, h_t, and the current input, x_t, to the FFN for computation of the next value in the sequence. Presentation of both values can't be guaranteed without satisfying the above inequality.

As brief proof by contradiction, assume that we are guaranteed to be able to present both h_t and the x_t without dead space in the model dimensionality. Since there is no dead space, every element in the vector is used to embed some piece of information about the token. To present both the embedding for the input and the last output to the FFN (without violating the mentioned requirements) we must compress the input or the last output. In the case of a dense embedding ie. a single bit, compression is not possible. Therefore, without the presence of dead space in the model dimensionality it may be impossible to present x_t and h_t to the FFN at time step t. Therefore, assuming no dead space is required to present both h_t and x_t at a single time step leads to a contradiction.

In the case of simulating an RNN, we can say that the minimum model dimension for x_t and h_t to be presented to an RNN simulating FFN simultaneously must be greater than or equal to twice the size needed to house an embedded token, $d_{model} \geq 2 \cdot d_{embed}$. In practice, some embeddings may be losslessly compressible. However, this assumption does not hold for all embeddings.

However, direct RNN simulation is sufficient, but not necessary, for Turing completeness. Therefore, the size requirement for RNN simulation does not imply an equivalent size requirement for Turing completeness. However, the more general $d_{model} > d_{embed}$ does hold.

To see that this is the case, assume that the base embedding is not compressible. Now assume x_t and a state variable representing the internal state of a Turing machine is compressed into a latent sequence presented to an FFN. Assume the Turing machine's internal state may be compressed into a single binary value as a lower bound. The minimum dimensionality of the latent vector containing the Turing machine state and x_t is $d_{embed} + 1$. Recall, the FFN input dimensionality is required to be identical to d_{model}.

Therefore, for a decoder-only transformer model to be Turing complete, it must be true that $d_{model} > d^*_{embed}$ with d^*_{embed} being the dimsensionality of the compressed token embedding.

Interestingly, the inability to recognize PARITY shown in (Hahn, 2020) may be duplicated by showing that arbitrarily long binary words aren't compressible to any fixed size d. Consider, if the input to an attention layer is an n token sequence with each token encoding a binary value, at most 2^d values may be losslessly

compressed for PARITY computation. Therefore, in the case of hard attention without auto-regression, PARITY is not feed-forward recognizable if the length of the binary sequence is greater than 2^d.

5.6.2 Transformers and Wang's B Machines

It is important to point out, decoder-only transformer models do not directly approximate the behavior of Turing machines. Rather, they are computationally much more similar to the B machines studied by (Wang, 1957) which have a single tape and are incapable of erase or overwrite. By simulating a Turing machine via B machine, Wang showed that erasure is not necessary for computational universality. However, he also showed that a B machine cannot generate tape content identical to that of a Turing machine in all cases due to the lack of overwrite.

Decoder-only transformers possess additional limitations beyond those imposed on B machines. While they may read from any past tape location, (1) they are incapable of writing to any position on the tape apart from the next available location and (2) they may not read any position on the tape beyond the current write pointer location. This constitutes an unexplored type of theoretical computational machine which we refer to as causal B machines.

5.6.3 Parameter Inefficiency Provenance Conjecture

Based on the proof herein, small decoder-only transformers are computationally universal. However, due to the significant limitations on causal B machines, format restrictions imposed by an application (like sequence-to-sequence modeling) may prevent the architecture from utilizing arbitrary recursion to perform Turing complete computation. Given a single tape and single permissible write location, intermediate computations which do not fit the application output format will either violate the application or the application output format will prevent the intermediate computation result from being written.

We conjecture that the strong link between model size and model effectiveness may be related to application induced limitations which force the decoder-only model to induce more sophisticated operations rather than learning to compose them from "basic steps" unfolded through recursion. This is empirically plausible given the emergence of chain of thought (Wei et al., 2022) as a viable option in the largest of models. Our future work will address this question more thoroughly.

5.7 Conclusion

We have shown that the decoder-only transformer architecture is capable of simulating an arbitrary RNN and is therefore computationally universal under reasonable assumptions. This result holds even for a 1 layer transformer with a single attention head so long as the model dimensionality exceeds the dimensionality of

the minimum token embedding.

However, this result is limited by the fact that the analysis does not consider the limitations imposed by sequence-to-sequence modeling on the output format which may impact the *in situ* computational expressivity of the architecture.

Therefore, future work seeking to improve the parameter efficiency of decoder-only transformers should consider the effect of output format restrictions and potential architectural changes. Changes, like the inclusion of an additional tape (decoder output location), may permit recursion without dimishing the model's aptitude as a language model.

References

Bhattamishra, S., Patel, A., and Goyal, N. (2020). On the computational power of transformers and its implications in sequence modeling. *arXiv preprint arXiv:2006.09286*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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

Hahn, M. (2020). Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.

Liu, P. J., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L., and Shazeer, N. (2018). Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*.

Neto, J. P., Siegelmann, H. T., Costa, J. F., and Araujo, C. S. (1997). Turing universality of neural nets (revisited). In *Computer Aided Systems Theory—EUROCAST'97: A Selection of Papers from the 6th International Workshop on Computer Aided Systems Theory Las Palmas de Gran Canaria, Spain, February 24–28, 1997 Proceedings 6*, pages 361–366. Springer.

OpenAI (2023). Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Pérez, J., Marinković, J., and Barceló, P. (2019). On the turing completeness of modern neural network architectures. *arXiv preprint arXiv:1901.03429*.

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

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

Rillig, M. C., Ågerstrand, M., Bi, M., Gould, K. A., and Sauerland, U. (2023). Risks and benefits of large language models for the environment. *Environmental Science & Technology*, 57(9):3464–3466.

Schuurmans, D. (2023). Memory augmented large language models are computationally universal. *arXiv preprint arXiv:2301.04589*.

Siegelmann, H. T. and Sontag, E. D. (1992). On the computational power of neural nets. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 440–449.

Turing, A. M. (1937). Computability and λ-definability. *The Journal of Symbolic Logic*, 2(4):153–163.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, H. (1957). A variant to turing's theory of computing machines. *Journal of the ACM (JACM)*, 4(1):63–92.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., and Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Yun, C., Bhojanapalli, S., Rawat, A. S., Reddi, S. J., and Kumar, S. (2019). Are transformers universal approximators of sequence-to-sequence functions? *arXiv preprint arXiv:1912.10077*.

CHAPTER 6

Do Large Language Models Learn to Human-Like Learn?

6.1 Introduction

Transformer based neural networks have led to a number of recent advances in natural language processing and inference (Vaswani et al., 2017). These large language models (LLMs) acquire remarkable abilities through a form of unsupervised learning in which part of the data is hidden and the model is required to reproduce it, a form of cloze task which is similar to de-noising. The model parameters are updated to improve performance on this and similar *pre-training* tasks.

The number of examples required to achieve a language processing ability similar to a child is roughly six orders of magnitude larger than that required by a human (Warstadt et al., 2023). This pre-training process bears little resemblance to the behaviors identified as consistent with human-like learning (Langley, 2022). Further, current language model architectures are incapable of adjusting their architectures or parameters based on interactions and are, in this way, incapable of learning.

Through pre-training, language models acquire an alternative means of learning referred to as in-context learning (ICL) (Brown et al., 2020) which is unique in connectionist literature as it does not involve altering parameters and therefore may not experience *catastrophic forgetting*, though forgetting does occur (Coleman et al., 2023). Robustly establishing the presence (or absence) of catastrophic forgetting effects in ICL is an important target for future work.

Through ICL, LLMs perform tasks for which they have little relevant pre-training given a small number of examples (Radford et al., 2019). The examples are interpreted based on the LLM's pre-training and prior interactions, though sufficient ICL examples can override pre-training (Wei et al., 2023). So, while pre-training is not human-like, language models clearly exhibit facets of human-like learning through ICL (Langley, 2022).

In the remaining sections of this paper, I (1) propose that humans have not achieved human-like learning absent of a significant pre-training process, (2) provide an analysis of ICL in light of the facets of human-like learning given in (Langley, 2022), and (3) identify the facets of human-like learning which have not been sufficiently explored in ICL. These under-explored facets constitute an important hole in the current understanding of language model behavior and its relationship to human-like learning.

6.2 Emergent Human-Like Learning

I hold that a model which acquires the ability to human-like learn through a lengthy pre-training process is consistent with the development of human-like learning in humans.

The human brain is not developed in each individual. Rather, the brain's general architecture is inherited and represents countless generations of improvements. Concepts like cognitive modularity are tied strongly to evolutionary co-development of behavior and hardware (Barrett and Kurzban, 2006). The interplay of brain hardware development and behavior modification across a sea of time is believed to have led to the specialization of modular structures. Further, it is established that neuro-typical individuals learn differently as compared to their dyslexic peers (Alsulami, 2019), and that those with dyslexia have consistent differences in their brain structures as compared to the neuro-typical cohort (Sun et al., 2010).

The underlying neurological mechanisms which give rise to specific observed behaviors are not understood sufficiently to make a strong claim regarding the provenance of learning. However, it is reasonable to hypothesize that human-like learning is an ability that has been acquired, at least partially, through a form of evolutionary *pre-training*.

6.3 What is ICL?

In-context learning (ICL) refers to a model learning to perform a task after being given a single or small number of examples in the model's context. Importantly, ICL does not involve any changes to model weights. So, novel task abilities are necessarily a result of interactions between the tokens in the context and the pretrained model. For a longer review of work regarding ICL refer to (Dong et al., 2022).

As a clarification, not all language models acquire the ability to in-context learn. It has been shown to emerge when data possesses certain properties common in language. When these properties are absent, models perform tasks using stored information in weights but will not improve performance with the presentation of in-context examples (Singh et al., 2023).

6.4 Human-Like Learning Constraints

In this section, I consider the constraining attributes of human-like learning presented in (Langley, 2022) and examine current empirical and theoretical research helping to establish whether human-like learning associated behaviors have been found to be present in LLMs, specifically when they engage in ICL.

6.4.1 Learning Involves the Acquisition of Modular Cognitive Structures.

Many authors have held that cognitive structures in the human mind are modular with any precise meaning of modular being contested, like that in (Fodor, 2000) requiring that modules be separated and specific. In

(Barrett and Kurzban, 2006), the authors provide an empirical view of modularity:

> ...functional magnetic resonance imaging (fMRI) might demonstrate the interaction of multiple systems and use of information from multiple sources, such findings do not falsify a hypothesis of principled and specialized use of information by dedicated systems. Empirically, what counts as evidence for or against a particular hypothesis about modularity turns on having a theory that predicts which inputs are relevant and, therefore, the psychological effects one expects to observe in different situations.

Modularity, in this notion, is not undermined by co-recruitment or distributed processing which is commonly found in fMRI-based human studies. In their view, such evidence simply serves to show that *encapsulation* is not a requisite component of modularity.

Adopting a similar perspective, I propose that LLMs possess functionally modular, though not encapsulated, knowledge structures. In (Bayazit et al., 2023), the authors show that domain specific knowledge subnetworks are identifiable and separable in GPT-2 such that, after ablation, the network is unable to perform related tasks but maintains unrelated knowledge and language ability. So, even though the entire network is executed for any task, it seems only a relatively small, modular subnet constitutes the pertinent knowledge for the task.

That being said, this does not suggest that structures are acquired during ICL since the network weights aren't being changed.

When a token is placed in the context of a transformer, three learned linear transformations are applied. The key and query transforms provide a representation that is used to find the attention weight placed on each token. Then, for each token in the context, $t_i \in S$, the associated attention, α_i, and value transform is used to create an admixture, $\Sigma_i \alpha_i \cdot V(t_i)$. It may be said that the unpacked value representations are structures acquired through ICL.

6.4.2 Learned Cognitive Structures Can be Composed During Performance.

Given the above notion of modularity, a cognitive structure within the network may be activated by a token in the ICL prompt. However, by having attention spread across multiple tokens, the output is generated from the compositions of individual modular structures (tokens). Each of these tokens then becomes a query that is used to create a set of contextually based representations. These representations are themselves composed into a single representation over the context given the query. The subsequent layers perform the same set of actions, resulting in compositions of compositions.

It is important to note that, while this can serve to create powerful compositions, transformers are not

capable of arbitrary composition (Roberts, 2023) without recursion. Though models like the decoder-only transformer are capable of recursion, language models don't typically learn this behavior as evidenced by the need to explicitly illicit recursive problem solving behavior through chain of thought prompting (CoT) (Wei et al., 2022).

6.4.3 Many Learned Cognitive Structures Are Relational.

It is well established that language models based on the transformer architecture learn relational information (Rezaee and Camacho-Collados, 2022; Bouraoui et al., 2020; Petroni et al., 2019). However, the relevant question is, do language models learn novel relational structures through ICL?

It has been shown that ICL facilitates learning truly new relational information (Kossen et al., 2024). However, this does not suggest that ICL permits the induction of novel types of relational structure as is necessary when prompted with semantically unrelated labels (SUL-ICL). This ability has been shown to tend to emerge when language models are massively scaled (Wei et al., 2023) like in the case of PaLM-50B (Anil et al., 2023).

6.4.4 Expertise Is Acquired In a Piecemeal Manner.

As discussed, each individual token presented through ICL results in an additional modular, composable structure which, by nature, is acquired in a piecemeal manner. However, to some relevance here, the study of human behavior has revealed many distinguishing facets present in expert behavior, like the use of heuristics as opposed to a reliance on rules, which are absent in the novice (Palmeri and Cottrell, 2010). So, a more nuanced question may be, do language models in-context learn expert-like performance and behavior? A review of the current literature regarding ICL **suggests this has not been addressed**.

In (Anderson, 1995), a link between long term memory and expert behavior is established. I recommend future work should investigate the effects of ICL on language model long-term working memory (LTWM) (Sohn and Doane, 2003) for items of the type presented in prompting, to empirically establish the relationship of ICL and expert behavior.

6.4.5 Learning Is An Incremental Activity That Processes One Experience At a Time.

The work in (Kossen et al., 2024) shows that ICL permits a language model to develop improved task performance with each in-context example. However, in most empirical work on ICL the test method presents all in-context examples as a small batch as opposed to interleaved experience and inference as may often be the case in human interaction. While interleaved example and inference may be a common practical prompt pattern in language model use, a review of the literature suggests this its effects on ICL performance **have not**

been explicitly considered. However, research suggesting ICL suffers from a form of forgetting (Coleman et al., 2023) suggests that interleaved ICL may have a mitigated effect.

6.4.6 Learning Is Guided by Prior Experience.

In (Kossen et al., 2024), the authors show that providing a single incorrect example followed by correct examples harms the model's performance until correct in-context examples sufficiently outnumber the incorrect example. This shows that learning is guided by the prior experience to an extent.

However, in (Langley, 2022) the motivational examples call for a more significant treatment of this question. At the time of writing, no work was identified in the literature that explicitly considered the degree to which subsequent ICL examples interfere (constructively or destructively) with prior examples.

6.4.7 Cognitive Structures Are Acquired and Refined Rapidly.

ICL drew significant attention as a unique ability that language models learn to exhibit. As already described, the hallmark of ICL is the ability to learn novel tasks from one to a few examples. Language models are certainly able to acquire (Radford et al., 2019) and refine (Kossen et al., 2024) knowledge structures rapidly with few examples through ICL.

6.5 Conclusions

ICL is a powerful and unique emergent ability present in certain language models of sufficient size. When the pre-training of the language model is seen analogically as a counterpart to the evolution of the human brain, ICL stands as a reasonable counterpart to human-like learning in language models. I have examined the constraints defined in (Langley, 2022) and applied the resulting insightful lens to ICL in language models by examining the literature and identifying the challenges within the gauntlet of human-like learning already met by ICL and those that stand as important future work.

The development of expertise and the effect of incremental experience have not been sufficiently considered in the literature. Further, the composition of transformers is bounded by the depth of the model given most models are unable to engage in arbitrary recursion. However, all other constraints given in the motivating paper have either been shown to be empirically or theoretically met by ICL.

Published Version

Roberts, J (2024). Do Large Language Models Learn to Human-Like Learn?. Proceedings of the AAAI 2024 Spring Symposium Series.

References

Alsulami, S. G. (2019). The role of memory in dyslexia. *International Journal of Education and Literacy Studies*, 7(4):1–7.

Anderson, J. R. (1995). Cognitive psychology and its implications (4" ed.). *Now York: WH Freeman and Company*.

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al. (2023). Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Barrett, H. C. and Kurzban, R. (2006). Modularity in cognition: framing the debate. *Psychological review*, 113(3):628.

Bayazit, D., Foroutan, N., Chen, Z., Weiss, G., and Bosselut, A. (2023). Discovering knowledge-critical subnetworks in pretrained language models. *arXiv preprint arXiv:2310.03084*.

Bouraoui, Z., Camacho-Collados, J., and Schockaert, S. (2020). Inducing relational knowledge from bert. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7456–7463.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Coleman, E. N., Hurtado, J., and Lomonaco, V. (2023). In-context interference in chat-based large language models. *arXiv preprint arXiv:2309.12727*.

Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., and Sui, Z. (2022). A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Fodor, J. A. (2000). *The mind doesn't work that way: The scope and limits of computational psychology*. MIT press.

Kossen, J., Gal, Y., and Rainforth, T. (2024). In-context learning learns label relationships but is not conventional learning. In *The Twelfth International Conference on Learning Representations*.

Langley, P. (2022). The computational gauntlet of human-like learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12268–12273.

Palmeri, T. J. and Cottrell, G. W. (2010). Modeling perceptual expertise. In Goldstein, E. B., editor, *Encyclopedia of Perception*, pages 197–244. Oxford University Press.

Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., and Riedel, S. (2019). Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

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

Rezaee, K. and Camacho-Collados, J. (2022). Probing relational knowledge in language models via word analogies. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3930–3936.

Roberts, J. (2023). On the computational power of decoder-only transformer language models. *arXiv preprint arXiv:2305.17026*.

Singh, A. K., Chan, S. C., Moskovitz, T., Grant, E., Saxe, A. M., and Hill, F. (2023). The transient nature of emergent in-context learning in transformers. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Sohn, Y. W. and Doane, S. M. (2003). Roles of working memory capacity and long-term working memory skill in complex task performance. *Memory & cognition*, 31:458–466.

Sun, Y.-F., Lee, J.-S., and Kirby, R. (2010). Brain imaging findings in dyslexia. *Pediatrics & Neonatology*, 51(2):89–96.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Warstadt, A., Choshen, L., Mueller, A., Williams, A., Wilcox, E., and Zhuang, C. (2023). Call for papers–the babylm challenge: Sample-efficient pretraining on a developmentally plausible corpus. *arXiv preprint arXiv:2301.11796*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., and Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Wei, J., Wei, J., Tay, Y., Tran, D., Webson, A., Lu, Y., Chen, X., Liu, H., Huang, D., Zhou, D., and Ma, T. (2023). Larger language models do in-context learning differently.

CHAPTER 7

Subscription-Based Models Harm Reproducibility and Current LLM Architectures Lack Computational Power

7.1 Introduction

Transformers are among the most influential machine learning methods developed to date. It has lead to state of the art performance on numerous tasks (Lin et al., 2022). However, the early and continued success of the vanilla model (Vaswani et al., 2017) has encouraged companies to create proprietary, large models with limited experimental architectural deviations.

Subscription based models and architectural stagnation are, in our estimation, two of the most important problems in the current transformer research landscape. We discuss the problem with each of these and offer reasoned positions for the future direction of the machine learning and artificial intelligence communities.

In this paper, we choose to be brief in the statement of our positions and reasoning for the reader's convenience. Based on our literature review we are the first to present a reasoned position on the issues associated with subscription based models. In contrast, many researchers hold that more research investigating transformer architecture variations is needed to address specific shortcomings: model size (Fournier et al., 2023), longer context (Zaheer et al., 2020), and vision (Khan et al., 2022) are among the list. However, we are the first to identify theoretical computational limits of current transformer models and call for architectural changes in support of research toward artificial general intelligence (AGI).

7.2 The Problem with Subscription Based Models

Research that, by nature, can't be reproduced is antithetic to the scientific method. Failing to reproduce a published result is precisely the grounds upon which a published hypothesis is to be refuted (Oates et al., 2022). Therefore, for research to be constructive and develop the overall body of knowledge, it must be reproducible.

Paleontologists, zoologists, and botanists aren't permitted to publish research conducted on privately held specimens in most journals as they follow the codes of conduct given by the International Commission on Zoological Nomenclature (ICZN) (Article 16.4.2) (Ride et al., 1999) and the International Commission on Biological Nomenclature (ICBN) (Turland et al., 2018) (Article 8.1, 8.5, and 40.4). These codes require that specimens which are used in the preparation of academic manuscripts are deposited in public institutions to ensure future scientific inquiry will have access to the necessary artifacts to reproduce and extend the studies.

This commitment to reproducibility comes at a real cost. Some of the most complete fossil skeletons have

been found on private land within the United States. These fossils, like Stan the T. Rex are sold at auction and often wind up in private collections (Roddy, 2021) and go unstudied. However, the only alternative is to accept that peer review and reproducibility aren't necessary components of scientific enquiry.

In the following subsections, we show that closed source models possess intrinsic road blocks to reproducibility that are very similar to those of privately held fossils. We also examine the incentive structure that may develop if academic study of closed source models is permitted to continue without limitation. We address various forms of partial disclosure and show that the only viable options to ensure reproducibility are deposition of the frozen model in an appropriate institution or open source disclosure of both the architecture and parameters.

7.2.1 Subscription Models Preclude Reproducibility

In the study of artificial intelligence and machine learning, it is not typical to require that a studied model and its parameters be kept and made available as is the case in the natural sciences. However, in an effort of the same spirit, publications like AAAI require that authors complete a reproducibility checklist, based on (Gundersen et al., 2018), verifying their work is documented sufficiently to be considered *experimentally reproducible* (Gundersen and Kjensmo, 2018). That is, given the same algorithm, data, and preparation, the experimental results will be reproduced. For a model, this necessarily includes the training data, architectural decisions, and all hyper-parameters.

In contrast, independent researchers do not have sufficient access to closed source model parameters and architecture to provide sufficient documentation for their work to be experimentally reproducible. The corporate owners of a subscription-based model have a fiduciary duty to their shareholders, employees, and other stakeholders (Marens and Wicks, 1999) to profit from the models. However, a traditional patent to protect the intellectual property is not a viable option as it is based on open source prior art that precludes a traditional patent (usc, 2023). So, to ensure future viability, companies restrict access to trade secrets like the model architecture, training process and data, and parameters. Most only offer access to their models as a subscription service.

Similarly, future researchers may not have access to a subscription-based model or its hyper-parameters even if they were documented as the company may discontinue the model or disband altogether. Even when the model does remain accessible, since models are not frozen or downloadable, research based on them often fails even to be self-reproducible. In (Suri et al., 2023), the authors kept all user driven hyper-parameters consistent including the target model revision. However, one day after the first set of experiments were run a silent update was performed which changed the model's behavior, making the results non-reproducible even by the original investigator.

What if all the parameters are exposed but not the architecture? If the company becomes goes out of business, the model may no longer be available. Therefore, knowing a list of parameters without documentation of the architecture, does not permit experimental reproduction.

What if all the architecture is exposed but all parameters? As with the previous situation, if the company becomes defunct, the model will no longer be available. In this case, without also having documentation of the parameters, the experiments will not be reproducible.

What if the company guarantees that the models will remain available? A similar question often arises regarding privately held fossils. The difficulty here is such a guarantee can't be made in perpetuity. The owner of the artifact can't be held to follow through on the guarantee as they may be legally prevented by future laws, prevented by the future owner of the artifact, or future leadership may simply opt to disregard the guarantee.

Private companies have a reasonable desire to restrict access to their trade secrets. Further, maintaining permanent revisions at all update points is a costly obligation, with each being several hundred gigabytes for GPT-3 (Hu et al., 2021), not necessary to provide a robust tool to their primary customer base. Therefore, financial incentives for such a company drive them toward practices that do not support reproducible, independent research.

7.2.2 Academic Artifacts as a Subscription Service

Companies like OpenAI develop powerful models. Their business depends on individuals and companies desiring access to their proprietary models and paying for a subscription to the models, usually in the form of a number of passes tokens or images. These models and the surrounding infrastructure form a valuable tool for industry, and it is well within the company's rights to withhold details of the model parameters and architecture to protect its assets and trade secrets. In our opinion, this business model is entirely appropriate. However, if AI research is permitted to investigate these closed source models likewise by subscribing to the model, what is being sold is not access to a tool but access to an academic artifact.

We attempted to investigate the number of NSF grants that are partially being used to pay for subscriptions to OpenAI's GPT models. The actual spending from NSF research grants is not publicly available. However, as a lower bound, there are 11 currently funded NSF projects that explicitly reference GPT in their abstract. Of these, five necessarily require a subscription for the proposed research (as a direct investigation or a baseline comparison). However, it is reasonable to assume the number of grants which will use closed source models as a baseline or artifact for research is far higher than this.

Research leading to a viable tool for industry is a great success and rightly monetized by the holders of the intellectual property. Requiring a subscription to access an artifact of intellectual importance is ethically

dubious and predatory.

If this is permitted, researchers are then incentivized to develop models and, rather than release the model as open source for the academic community, put them behind a paywall so that researchers must pay for a subscription to test novel datasets or run baseline comparisons.

7.2.3 Proposed Conditions for Publication

We believe the field of AI should adopt an attitude similar to paleontology and other natural sciences. In consideration of reproducibility, incentive structures, and the rights of for-profit companies, we propose as a requirement for publication:

> AI and machine learning research should be based on models that are either 1) fully open source with readily available public access or 2) closed source with a frozen, complete copy of the studied model deposited in an appropriate public institution and made available for future inquiry.

This does not prevent a closed source model from being studied, nor does it intend to. A company may facilitate reproducible research while not divulging either the architecture or parameters of their model by encrypting the model such that it is only executable through a trusted execution environment (Sabt et al., 2015) which takes in a set of inputs and digital rights management license and provides the model output. Systems like (Tramer and Boneh, 2018) combine secure network inference (Mann et al., 2023) and secure dissemination of the model. Thereby, closed source models could be protected but available for the reproduction and extension of research.

7.3 The Problem with Current LLM Architectures

In this section, we argue that current language model architectures, while capable of extraordinary feats, are computationally limited. We show that these limitations are produced by the confluence of model application and architecture. However, there are architectures that do not possess the same limitations and recommend that research into intelligent systems explore more expressive architectural variants.

7.3.1 How Computationally Powerful are Transformer Architectures?

The full auto-encoder transformer model was shown to be Turing complete if the softmax attention is assumed to approximate ”hard” attention (Pérez et al., 2019; Bhattamishra et al., 2020). However, both of the proof approaches depend strongly on the presence of the encoder. LLMs like GPT-x are based on the decoder-only architecture (Radford et al., 2018; OpenAI, 2023) whose construction does not include an encoder. The decoder-only architecture, absent of an encoder, was later shown to be Turing complete under a similar

assumption of ”hard” attention (Roberts, 2023). So, without any consideration of the language modeling task, the current transformer architectures used to construct LLMs are computationally universal.

In (Roberts, 2023), the authors identify transformer models as a form of *causal B machines*, a restricted type of B machine, or non-erasing Turing machine, which is known to be Turing equivalent (Wang, 1957). Causal B machines have a single tape and a single write pointer which is always pointing to the next empty location on the tape with all spaces beyond the write pointer being empty. They are unable to read any space on the tape at or beyond the write pointer and unable to write to any location other than that pointed to by the pointer. Further, they are unable to overwrite any non-empty location. Based on this description, all current and prevalent transformer models fall into this class of algorithm.

In (Roberts, 2023) the authors propose that, since decoder-only language models are Turing complete, the increased capability with increased model size must be due to interaction between the architecture and the task. We further this conjecture here by giving a simple, intuitive proof that while the encoder-decoder and decoder-only transformer architectures are Turing complete, causal B machines engaged in a language modeling task can't be Turing complete and simultaneously pass a Turing test.

7.3.2 Proof that Transformer LLMs aren't Turing Complete

First, B machines are known to be Turing complete. However, the inability to overwrite prevents them from performing certain functions that are possible for a Turing machine. As an example, it is possible for a traditional Turing machine to compute the result of an arbitrary computation, erase all other tape content, and then leave the tape filled only with the computational result with no intermediate steps. B machines are not able to eliminate the intermediate steps from the tape as they are not able to overwrite. Therefore, B machines and causal B machines can't compute an arbitrary function without intermediate computations being written in perpetuity to the tape.

Consider a task that requires the number of execution steps for a given program as output but does not allow intermediate steps be written to the output. For a given program that terminates, finding the number of steps before termination is impossible without executing the program (Turing et al., 1936). Therefore, the task is possible for a Turing machine by overwriting the intermediate steps but impossible for a B machine/causal B machine.

When tasked as described, causal B machines may either use the output space to perform intermediate computations such that the answer is correct (failing to comply with the task parameters) or give an approximate answer (answering incorrectly on average).

In contrast, humans are capable of performing the described task by considering the verbal response as the output tape and all internal thought and written work as intermediate computations on a non-output tape.

If we allow that humans may sometimes answer incorrectly but are able to consistently comply with the output content restriction, then causal B machines are left with a dilemma. They may choose to comply with the task's output content restriction, or they may choose to perform the intermediate computations. However, they are mutually exclusive.

If a causal B machine is coded so that it prefers answering correctly, it will never be able to comply with the task output requirement, and therefore be clearly identifiable as non-human. Alternatively, if the causal B machine is coded to prefer human-like interaction, it will be incapable of computations requiring recursion. The inability to comply with the output restriction while performing the computations is sufficient to prove that, for a causal B machine, Turing complete computation is incompatible with the ability to pass a Turing test.

Current, prevalent transformer architectures, as variants of causal B machines, are therefore incapable of Turing complete computation while passing a Turing test.

7.3.3 Formal Proof

Let k be the number of floating point operations that are computable and storable in the state representation of causal B machine M.

Let Task T require that the next location along the *interactive output* of M be $M(T) = O_i$ where i is the current output pointer and contain the number of executed floating point operations, x, given a program, P that terminates.

Task T is solved if there exists M such that, for all P, $O_i = x$. Alternatively, if there exists some P for all M such that $O_i \neq x$, then T is considered unsolvable.

Theorem 7.3.1. *For any causal B machine M with finite state representation size k and unchangeable write pointer location i, there exists a program P such that the number of floating point operations executed is* $x \geq k+1$.

From Theorem 7.3.1 it is clear that the stated task is unsolvable for any causal B machine. Note that the difficulty lies in the nature of the causal B machine. Possessing only a single output location, the interactive output is not separable from latent or internal computations if the internal computations exceed the state size of the causal B machine. This task is clearly soluble for any machine that may overwrite or any machine that has multiple output locations such that the interactive output is not the only space to write intermediate computations.

7.3.4 Comments of the Theoretical Result

The result described so far, seems on the surface to be incompatible with previous theoretical results showing that encoder-decoder and decoder-only transformer architectures are Turing complete. This perception is intuitive but erroneous. The past Turing complete results implicitly assumed that the output of the model was unrestricted and permitted to be used for recursion. When this implicit assumption is relaxed, the causal nature of the model, inability to overwrite, and single tape cause the previous theoretical result to fail to hold.

The simple proof herein identifies an important computational discrepancy between humans, Turing machines, and transformer architectures. Humans often do many computations internally, on a non-verbalized tape, to choose the next token to ultimately be output on a verbal tape. Turing machines only have a single tape but are able to overwrite spaces on the tape. Transformers do not have access to either additional tapes or the ability to overwrite as methods of deliberation, leading to practical computational limitations.

7.3.5 Empirical Support

Empirical results add speciousness to the theoretical conclusions. We present evidence that permitting current models to freely use output space for intermediate steps in problem solving leads to improved results just as restricting access leads to poor performance on even trivial tasks.

Chain of thought (CoT) prompting has been shown to allow transformer language models to answer difficult questions more accurately (Wei et al., 2022). Tantamount to relaxing the output restriction inherent to language modeling, CoT prompting explicitly permits the model to use the output for intermediate computations rather than simply providing an answer.

In contrast, prompting can be used to further restrict the output. We develop a prompt pattern referred to as K repetition + task. The prompt requires the model to repeat a target word K times, with K being some large number, then perform a simple addition or answer a simple question. The model complies with the requirement to repeat the target word. However, for large numbers of occurrences, recursive summation is necessary to identify when K repetitions have occurred just as recursion is necessary to compute PARITY (Hahn, 2020). Since the output space is restricted to repetition of the target word, addition is not possible. The prompted model becomes garrulous and repeats the word far more than K times. Repetition ends when the prompt is diluted due to softmax saturation. At this point, random hallucinations begin and continue until the max number of tokens in the context is reached.

This experimental result has been repeatable for values of K close to or greater than 100, though it may not remain so due to the problems associated with subscription based models. The model sometimes terminates early without approaching K repetitions. This likewise reinforces the conclusion that the model is performing a guess regarding the elapsed number of repetitions since summation isn't possible.

7.3.6 Position on Current Transformer Architectures for Intelligent Systems

The theoretical result shows that current single-tape, non-erasing transformer architectures are incapable of Turing complete computation while passing a Turing test. The empirical results from novel experiments and past work on CoT support the theoretical conclusion. When a prompt restricts access to intermediate computation space, GPT-3 chooses to comply but must therefore guess the result of some necessary computations. Alternatively, when CoT prompting permits additional access to intermediate computation space, the answers to difficult questions become more accurate.

Given the substantial advances brought to the field by transformer models (Vaswani et al., 2017) we believe that transformers may provide an avenue to AGI. However, based on the theoretical and empirical considerations here, other architectures, or architectural variations like overwrite capabilities and multi-tape constructions, should be considered.

7.4 Conclusions

We have given evidence that subscription based models, like privately held fossils, may not be available for future researchers to reproduce and extend work. Next, the incentives associated with private models aren't aligned with scientific research. As a solution, we suggest that AI and ML publications should adopt policies that require published research investigating transformer models be based on open source models or models that are deposited in appropriate public institutions to ensure reproducibility. This position does not preclude research based on closed source models nor does it suggest that closed source models can't be used as tools for manuscript preparation.

Additionally we have argued that the success of the vanilla transformer architecture has potentially slowed the exploration of alternative architectures that may permit deliberation through overwrite or access to multiple computing tapes. We hold the position that investigating alternative architectures, like those mentioned, is paramount to realize progress toward AGI. All current, mainstream transformer architectures belong to the class of causal B machines and are unable to pass a Turing test while being Turing complete. While this combination is not necessarily required for language modeling applications, it is reasonable to expect, or potentially require, that any AGI be computationally universal and optionally indistinguishable from a human.

References

(2023). 35 u.s.c. § 102(a). United States Code. Available online at https://www.law.cornell.edu/uscode/text/35/102.

Bhattamishra, S., Patel, A., and Goyal, N. (2020). On the computational power of transformers and its implications in sequence modeling. *arXiv preprint arXiv:2006.09286*.

Fournier, Q., Caron, G. M., and Aloise, D. (2023). A practical survey on faster and lighter transformers. *ACM Computing Surveys*, 55(14s):1–40.

Gundersen, O. E., Gil, Y., and Aha, D. W. (2018). On reproducible ai: Towards reproducible research, open science, and digital scholarship in ai publications. *AI magazine*, 39(3):56–68.

Gundersen, O. E. and Kjensmo, S. (2018). State of the art: Reproducibility in artificial intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Hahn, M. (2020). Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., and Shah, M. (2022). Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41.

Lin, T., Wang, Y., Liu, X., and Qiu, X. (2022). A survey of transformers. *AI Open*.

Mann, Z. Á., Weinert, C., Chabal, D., and Bos, J. W. (2023). Towards practical secure neural network inference: the journey so far and the road ahead. *ACM Computing Surveys*, 56(5):1–37.

Marens, R. and Wicks, A. (1999). Getting real: Stakeholder theory, managerial practice, and the general irrelevance of fiduciary duties owed to shareholders. *Business Ethics Quarterly*, 9(2):273–293.

Oates, B. J., Griffiths, M., and McLean, R. (2022). *Researching information systems and computing*. Sage.

OpenAI (2023). Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Pérez, J., Marinković, J., and Barceló, P. (2019). On the turing completeness of modern neural network architectures. *arXiv preprint arXiv:1901.03429*.

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

Ride, W. et al. (1999). *International code of zoological nomenclature*. International Trust for Zoological Nomenclature.

Roberts, J. (2023). On the computational power of decoder-only transformer language models. *arXiv preprint arXiv:2305.17026*.

Roddy, B. (2021). Can you dig it? yes, you can! but at what cost?: A proposal for the protection of domestic fossils on private land. *Tex. A&M J. Prop. L.*, 8:473.

Sabt, M., Achemlal, M., and Bouabdallah, A. (2015). Trusted execution environment: what it is, and what it is not. In *2015 IEEE Trustcom/BigDataSE/Ispa*, volume 1, pages 57–64. IEEE.

Suri, G., Slater, L. R., Ziaee, A., and Nguyen, M. (2023). Do large language models show decision heuristics similar to humans? a case study using gpt-3.5. *arXiv preprint arXiv:2305.04400*.

Tramer, F. and Boneh, D. (2018). Slalom: Fast, verifiable and private execution of neural networks in trusted hardware. *arXiv preprint arXiv:1806.03287*.

Turing, A. M. et al. (1936). On computable numbers, with an application to the entscheidungsproblem. *J. of Math*, 58(345-363):5.

Turland, N. J., Wiersema, J. H., Barrie, F. R., Greuter, W., Hawksworth, D. L., Herendeen, P. S., Knapp, S., Kusber, W.-H., Li, D.-Z., Marhold, K., et al. (2018). *International Code of Nomenclature for algae, fungi, and plants (Shenzhen Code) adopted by the Nineteenth International Botanical Congress Shenzhen, China, July 2017*. Koeltz botanical books.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, H. (1957). A variant to turing's theory of computing machines. *Journal of the ACM (JACM)*, 4(1):63–92.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., and Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., et al. (2020). Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297.

CHAPTER 8

Summary and Conclusions

8.1 Empirical Summary

The empirical work in this dissertation improves understanding of what is currently *practicable* using human cognitive behavior as a comparison.

8.1.1 PopulationLM

To characterize LLM cognitive behaviors, a method called PopulationLM (Roberts et al., 2024) (Ch 3) was developed. It uses stratified Monte Carlo dropout to apply systematic variations to a base model to generate a population of that model. A population of Monte Carlo dropout perturbed models constructed in this way approximates a gaussian random process (Gal and Ghahramani, 2016). This tends to reduce the presence of behaviors that are not robust under variations (E.2 from Table 1.1).

The method is used to replicate a study on typicality and another on structural priming in LLMs (E.3). Typicality is shown to be present in the tested language models. Further, it is shown that models tend to learn typicality when they are exposed to a sufficient number of training tokens drawn from the associated category (Roberts et al., 2024). On the other hand, the experiments showed that structural priming did not tend to be present in the populations for any of the tested models.

8.1.2 Model of Humans in the Traveler's Dilemma

Humans engaged in a traveler's dilemma (Basu, 1994) deviate from game theoretic predictions. They tend to choose strategies that are far from the Nash equilibrium and, by doing so, achieve a better payoff. In (Roberts, 2021) Ch 2, it is shown that the rationale by which the Nash equilibrium emerges is not supported if uncertainty regarding a preference for a weakly dominating strategy is not certain. The model formulates elimination of weakly dominated strategies as a many-valued, elimination of fuzzy weak dominated strategies. This reformulation is shown to be a faithful predictor of empirical human behavior. However, the assumption of uncertainty regarding strategy preference is not substantiated and is left a hypothesis (E.1).

8.1.3 LLM Strategic Behavior

PopulationLM has been leveraged to study the strategic behavior of LLMs in Ch 4. The experiments show that Solar (Kim et al., 2023), Mistral (Jiang et al., 2023), Llama-2 (Touvron et al., 2023), and Gemma (Team et al., 2024) tend to prefer strategies based on their value. The experiments show that smaller models tend to

prefer strategies based on superficial information like the strategy label. Further, the results show that Gemma and Llama-2 are brittle under variations. This identifies Mistral and Solar as important, robust models that are capable of value based preference (VBP). This is extended by evaluating LLM behavior in an obfuscated prisoner's dilemma. Solar and Mistral respond to the scenario in a human-consistent manner, including human-consistent sensitivity to stake size (E.4).

Finally, the study of LLM strategic behavior shows that LLMs that are capable of robust, VBP are not necessarily certain of a preference for a weakly dominated strategy. Further, it establishes that LLM behavior in the traveler's dilemma is consistent with empirically established human sensitivity to penalty size, a parameter of the traveler's dilemma game. LLMs capable of VBP are indifferent toward a strategy that weakly dominates another when the penalty is low. On the other hand, when the penalty is large, LLMs with VBP prefer strategies that are more near the Nash equilibrium (E.4).

8.2 Theoretical and Scholarly Summary

This dissertation augments the theoretical work by providing a better understanding of what is *possible* for transformer-based LLMs.

8.2.1 Decoder-Only Transformer Models are Turing Complete

First, by showing that an RNN can be simulated by a decoder-only transformer, Ch 5 shows that decoder-only transformers are Turing complete (TS.1). Further, Wang developed a model of computation called B machines that was incapable of overwriting (Wang, 1957) but was proved Turing complete. This is extended by categorizing decoder-only and vanilla transformers as a special case of B machine called causal B machines (TS.2). That is, B machines that cannot overwrite and may only write to the current pointer location which is then incremented precisely by 1.

8.2.2 LLM Pre-Training May Be Considered Human-Consistent

Some regard the large number of pre-training tokens necessary to achieve significant LLM behavior to be a distinct difference as compared to human behavior. However, humans inherit a base architecture that has been developed across innumerable generations. To account for this pre-training, comparisons between humans and pre-trained LLMs are more equitable than comparisons to a randomly initialized model (TS.3). Additionally, the development of expertise, incremental learning effects, and catastrophic forgetting when engaged in in-context learning (ICL) are yet unexplored in the ICL literature.

8.2.3 Decoder-only Transformer LLMs Aren't Turing Complete in Some Tasks

There are tasks that require the evaluation of an arbitrary partial recursive function and expect only the answer, without intermediate steps, be written to the output used for interaction. An example of such a task would be to provide the number of steps necessary for a program to terminate. When a causal B machine, which has only one tape and is incapable of changing the write location, is engaged in such a task, it may either evaluate the partial recursive function or comply with the requirement to record the answer without intermediate computation if the number of computations needed to terminate is greater than the state representation.

So, even though a large model may be very powerful, without recursion, the function cannot be guaranteed to be evaluated since the latent space of the model is finite. Recursion can only be accomplished, in a causal B machine, by recording a value to the output which is then fed back in as input. Therefore, given a function to be computed which requires a number of floating point operations greater than can be computed in the latent space of the model, the model may either comply with the task requirements or compute the correct answer, but not both. Humans are not limited in this way since they are capable of computations that are not interactively output. Therefore, such a task could be used to differentiate a human from a causal B machine (TS.4).

Possible architectures that don't have this limitation would be (1) models that have separate decoders for interaction and computation, (2) models that permit message passing around the tokenizer by way of a [COMPUTING] token, and (3) models that are able to choose to either overwrite their last written token or move to the next empty output space.

A system similar to (2) was presented in (Goyal et al., 2024). However, it possessed a fixed number of non-output computations allowed per output token. For any fixed number of recursions, there exists a partial recursive function that is not evaluable. Therefore, this construction is an important but preliminary step toward alleviating the issue.

In the pursuit of artificial general intelligence (AGI), this dissertation advocates for exploration of alternate architectures which are less restricted than causal B machines (TS.5).

8.2.4 Reproducible Research

From results in research like (Suri et al., 2023) and the inapplicability of methods for systematic perturbation like PopulationLM (Roberts et al., 2024) Ch 3 to closed-source models, this dissertation argues that, like privately held fossils, closed-source models are not appropriate targets for scientific research. Closed-source models change without warning and previous model checkpoints are not available. This leads to results that are not reproducible long term (TS.6).

8.3 Conclusion

When LLMs are robustly tested, they exhibit human-like cognitive behaviors and have human-like strategic preferences with similar context sensitivities in both the prisoner's dilemma and the traveler's dilemma. For these behaviors to emerge, sufficient size and number of training tokens is necessary. The number of tokens necessary for pre-training is not necessarily incompatible with the development of human-like behavior. However, the increased presence of human-like behavior with increased model size runs counter to the model's computational universality. The dependence on model size instead seems to stem from the task, language modeling, which requires the next token be a human-like completion. Therefore, to support progress toward artificial general intelligence (AGI), alternative transformer architectures will need to be explored. Further, to substantiate the presence of AGI and ultimately guide its development, research regarding LLM behavior must be reproducible and independently verifiable in the long term. This suggests a need for the research community to prioritize the development and investigation of open-source models.

References

Basu, K. (1994). The traveler's dilemma: Paradoxes of rationality in game theory. *The American Economic Review*, 84(2):391–395.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Goyal, S., Ji, Z., Rawat, A. S., Menon, A. K., Kumar, S., and Nagarajan, V. (2024). Think before you speak: Training language models with pause tokens. In *The Twelfth International Conference on Learning Representations*.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. (2023). Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Kim, D., Park, C., Kim, S., Lee, W., Song, W., Kim, Y., Kim, H., Kim, Y., Lee, H., Kim, J., et al. (2023). Solar 10.7 b: Scaling large language models with simple yet effective depth up-scaling. *arXiv preprint arXiv:2312.15166*.

Roberts, J. (2021). Finding an equilibrium in the traveler's dilemma with fuzzy weak domination. In *2021 IEEE Conference on Games (CoG)*, pages 1–5. IEEE.

Roberts, J., Moore, K., Wilenzick, D., and Fisher, D. (2024). Using artificial populations to study psychological phenomena in neural models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):18906–18914.

Suri, G., Slater, L. R., Ziaee, A., and Nguyen, M. (2023). Do large language models show decision heuristics similar to humans? a case study using gpt-3.5. *arXiv preprint arXiv:2305.04400*.

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., et al. (2024). Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, H. (1957). A variant to turing's theory of computing machines. *Journal of the ACM (JACM)*, 4(1):63–92.