A Paradox in Machine Preference

Peter Cotton

First version: July 2024

Abstract

Using prompts such as: "my favorite state in the US is [MASK]", and "my favorite Western state in the U.S. is [MASK]" we infer that Thurston models are a better match to the revealed preferences of large language models than the application of Luce's Choice Axiom. There is some irony in this finding given that Softmax functions, responsible for the token probabilities we interpret as preference, suggest Independence of Irrelevant Alternatives.

1 Introduction

The advent of large language models (LLMs) has revolutionized natural language processing, enabling machines to perform tasks that were once considered exclusive to human intelligence. These machines have not only achieved proficiency in language tasks, but, in the opinion of some observers, have also begun to exhibit reasoning ability, emotion, and volition. From this side of the artificial intelligence debate, it is as natural to study revealed preferences for thinking machines as it is to study them in humans - with some considerable urgency added by the alignment problem!

From an opposing perspective, Large Language Models (LLMs) are mere circus stunts implementing little more than statistical mimicry of humans via next token prediction. But if so, the study of machine preference can be viewed as a powerful, albeit indirect way to study human predilections. The noise inherent in this approach might be compensated for by the sheer quantity of data that can be accumulated, since interrogating LLMs is far cheaper and more scalable than arranging psychology experiments or surveys.

In the realm of choice modeling, Luce's Choice Axiom [5] has long been a cornerstone, positing that the probability of selecting an item from a set is proportional to its utility relative to the sum of utilities of all available items. In economics, this is sometimes justified through the notion of Independence of Irrelevant Alternatives (IIA): it asserts that the relative probabilities of selecting one option over another should remain unaffected by the presence or absence of other irrelevant alternatives in the choice set.

For example, if a person is deciding between options A and B, the ratio of their probabilities $\frac{P(A)}{P(B)}$ should stay the same, regardless of whether a third

option C is introduced or removed, provided C does not alter the utilities of A or B.

There is a natural alignment between this workhorse assumption and the token probabilities of the Large Language Models. Invariably, these token probabilities are the output of a Softmax function in the final layer of the network. The i'th token probability p_i is

$$p_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \tag{1}$$

where the input z_i might be result of a very complicated transformer computation. Despite this network complexity entering z_i there is a simple implication. If we set $z_1 = 0$ say, as a kind of surgical procedure that would be impossible in humans, the artificial brain will eliminate item 1 from consideration. The new set of choice probabilities $\{p_2, \ldots, p_n\}$ will be the same as the old set of probabilities up to a normalization factor $1/(1 - p_1)$. One might therefore say that when we use Softmax functions for token probabilities, we are heavily *suggesting* to LLMs that they obey Luce's Choice Axiom.

They don't. Here it is demonstrated that LLMs, despite using a mechanism that should theoretically align with Luce's Choice Axiom, actually exhibit preference behaviors that are better modeled by the Thurstonian framework. By designing experiments that probe the token selection probabilities of BERT [2] and RoBERTa [4] that are programmed to "fill in the blanks", we aim to shed light on the underlying mechanisms driving machine preferences - and possibly our own.

2 A choice of choice models

Luce's Choice Axiom asserts that the probability of choosing an item i from a set S is given by:

$$P(i|S) = \frac{u(i)}{\sum_{j \in S} u(j)} \tag{2}$$

where $u(i)$ represents the utility of item i. This implies a linear adjustment of probabilities when items are removed from S.

One might describe this as an "urn model" because we can imagine associating each item k with a plurality of balls that are placed in an urn, all sharing the same label k. If item j is more likely, it is assigned more balls. Choice probabilities correspond to a random drawing of a ball from the urn and, consequently, a simple renormalization of probability when the choice set S changes seems perfectly reasonable. It amounts to removing balls from the urn.

But perhaps preference is a horse race, not an urn. We consider a competing model for choice where each item is assigned a location parameter (dislikability) a_i. Each item is thereby also associated with a unit variance normally distributed variable $X_i \sim N(a_i, 1)$ centered on a_i. The probability of choosing item i is modeled as the probability that when we draw X_k for $k = 1, \ldots, n$ we find that the lowest value is X_i.

This second model, which falls into a class considered by Thurston, makes different predictions about the way the choice probabilities will change when the choice set is altered. Rather than removing a ball from the urn, we will remove a horse from a race.

(As an aside, it *is* possible to engineer a different style of contest where removal of one contestant satisfies Luce's Choice Axiom. To see this, we assume $X_i \sim Exp(h_i)$ that are exponentially distributed with the same location and differing hazard rates. Thereby, it is also possible to view the comparison we have in mind as between two different contest models, even if Luce's Choice Axiom is not usually regarded that way.)

3 Experimental designs

To investigate the preference behaviors of LLMs, we designed experiments using masked language modeling tasks with BERT (BertForMaskedLM) and RoBERTa. The models were prompted to fill in the blanks in sentences that invite choices among items in the same category. Two types of experiments were conducted.

In the first set of experiments, we asked LLMs to fill in a masked word in a question, which we refer to as the "original question", and also do the same for a second "qualified question". The use of the qualifier reduces the size of the choice set. For example:

1. *Original Question:*"My favorite state in the U.S. is [MASK] and I try to visit once a year."

2. *Qualified Question:* "My favorite Western state in the U.S. is [MASK] and I try to visit once a year."

These differ only because the adjective *Western* was added, thus restricting the collection of states the LLM can choose.[1]

In the second style of experiment, we more explicitly eliminated a single option by using the top answers from a first prompt to create a list of secondary prompts. As a second change, we also averaged the results over many similar pairs of questions differing only by a word or phrase substituting for SOMETHING. For example,

1. *First Prompt:*"My favorite primary color is [MASK] because it is *SOMETHING*."

2. *Second Prompt:* "My two favorite primary colors are [ANSWER] and [MASK] because they are *SOMETHING*."

By repeated many times for different choices of adjectives or phrases playing the part of SOMETHING, we hope to remove luck in the model comparison. We compared the models' token probabilities with the predictions made by Luce's

[1]In practice, LLMs do not always follow these instructions to the letter.

Choice Axiom [5] and the Thurstonian model [7]. In the first set of experiments, we compared RMSE errors for inferred conditional choice probabilities.

In the second, we compared total RMSE errors summed across all 2-way combinations of items and all choices of SOMETHING. In this context, Luce's Choice Axiom might be referred to as the simplest version of the Plackett-Luce model [6] or, in the horseracing context, a Harville model for quinella or exacta pricing [3].

4 Results for qualified questions

To further illustrate, Table 1 shows the results for the "favorite Western state" prompt using BertForMaskedLM. The "Actual" column represents the probabilities assigned by the model in the second qualified question. The "Original" column lists the token probabilities elicited by the unqualified question - though only for the Western states.

Predictions from Luce's Choice Axiom and the Thurstonian model are calculated using only the token probabilities from the unqualified task. They also displayed in Table 1. The Luce column calculation is the simplest because, as noted, we merely re-scale the original token probabilities to account for the reduced choice set.

The Thurstonian column in Table 1 represents a slightly more involved exercise. First, we calibrate a Thurston model, as described, to the original question token probabilities. This amounts to shifting the location parameters $\{a_i\}$ for the auxiliary variables $X_i \sim N(0, a_i)$ to achieve an exact calibration where

$$Prob(X_i < X_k \texttt{ for all k } \neq i \in S) = p_i \tag{3}$$

where p_i is the original token probability. Then we compute

$$p_i' = Prob(X_j < X_k \texttt{ for all k } \neq i \in S'$$

where S' is the reduced set of possibilities. Details of a fast numerical algorithm for this kind of Thurston model calculation were provided in [1].

In a similar manner, Table 2 presents the results for the "favorite baby name" and "favorite girl's baby name" prompts. The Thurstonian model again outperforms Luce's Choice Axiom, with the RMSE using Luce's probabilities being approximately 50% higher.

Figure 1 illustrates the head-to-head comparisons between Luce's Choice Axiom and the Thurstonian model across various experiments. The green points represent instances where the Thurstonian model provides more accurate predictions. The results indicate that on balance, the Thurstonian model typcially aligns more closely with the LLMs' behavior, particularly in scenarios with substantial qualification and low entropy.

4.1 Results for single item elimination

To test the robustness of our findings, we conducted additional experiments using RoBERTa and varied the prompts and adjectives. The two-step selection

State	Original (%)	Luce (%)	Thurstonian (%)	Actual (%)
California	10.89	29.87	25.38	20.17
Arizona	6.74	18.50	17.20	10.00
Texas	5.50	15.10	14.59	7.68
Colorado	3.18	8.74	9.38	7.81
Oregon	2.89	7.92	8.67	8.57
Oklahoma	1.97	5.41	6.37	6.64
Nevada	1.66	4.55	5.54	10.18
Montana	1.59	4.37	5.36	14.27
Idaho	1.05	2.87	3.82	4.99
Wyoming	0.98	2.69	3.62	9.69

Table 1: Comparison of predicted probabilities using Luce's Choice Axiom and the Thurstonian model for the "favorite Western state" prompt. The Thurstonian model provides a closer approximation to the actual probabilities.

Name	Original (%)	Luce (%)	Thurstonian (%)	Actual (%)
Lily	1.56	20.35	16.18	14.52
Emily	1.15	15.06	13.31	16.29
Mia	0.86	11.25	11.02	13.18
Bella	0.85	11.12	10.93	8.84
Sarah	0.68	8.89	9.44	9.41
Rachel	0.59	7.66	8.56	7.15
Kate	0.59	7.63	8.54	8.44
Lauren	0.47	6.11	7.39	7.23
Brittany	0.46	6.04	7.33	7.55
Chloe	0.45	5.89	7.21	7.39

Table 2: Comparison of predicted probabilities for "favorite girl's baby name." The Thurstonian model shows better alignment with the actual probabilities assigned by the model.

process and the use of different adjectives aimed to minimize bias introduced by specific wording.

Figure 2 presents examples of average relative prediction errors for different question categories. The Thurstonian model consistently shows lower errors across various contexts, reinforcing the conclusion that it provides a better fit for modeling LLM preferences.

5 Conclusion

The study demonstrates that large language models, despite being engineered with Softmax functions that align with Luce's Choice Axiom, display preference

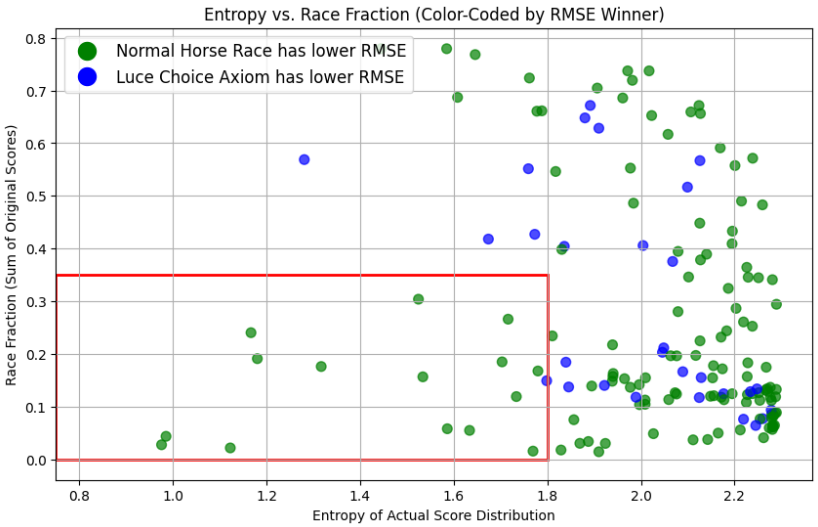

Figure 1: Head-to-head comparison of prediction errors between Luce's Choice Axiom and the Thurstonian model. Each dot is a different pair of unqualified and qualified mask prediction tasks. The Thurstonian model outperforms Luce's Choice Axiom in approximately 80% of the cases.

behaviors that are more accurately modeled by the Thurstonian framework.

To the extent that LLMs mirror our own preferences, this provides a compelling rejection of Luce's Choice Axiom in favor of the Thurston model. The popularity of the former may arise in large part due to its computational convenience, but an algorithm for Thurston models scaling to $n > 100,000$ was provided in [1] and so probably this approach should warrant more attention.

(LLMs token probabilities provide an example of large choice set where this new method of wielding Thurston models is highly relevant. Other applications include digital commerce.)

Leaning into the irony of Softmax machines not obeying Luce's Choice Axiom, the results suggest that Thurstonian neural network layers might generalize better than Softmax functions - although they obviously present the greater engineering challenge notwithstanding [1].

This study is coarse-grained but there may also be significant implications for modeling the nuances of machine decision-making processes in finer detail. If any such models satisfy Luce's Choice Axiom, for instance, they might be suspicious or readily improved. Similarly, this study might be viewed as one type of crude surrogate model, or a technique falling into the large category of model explanation tooling.

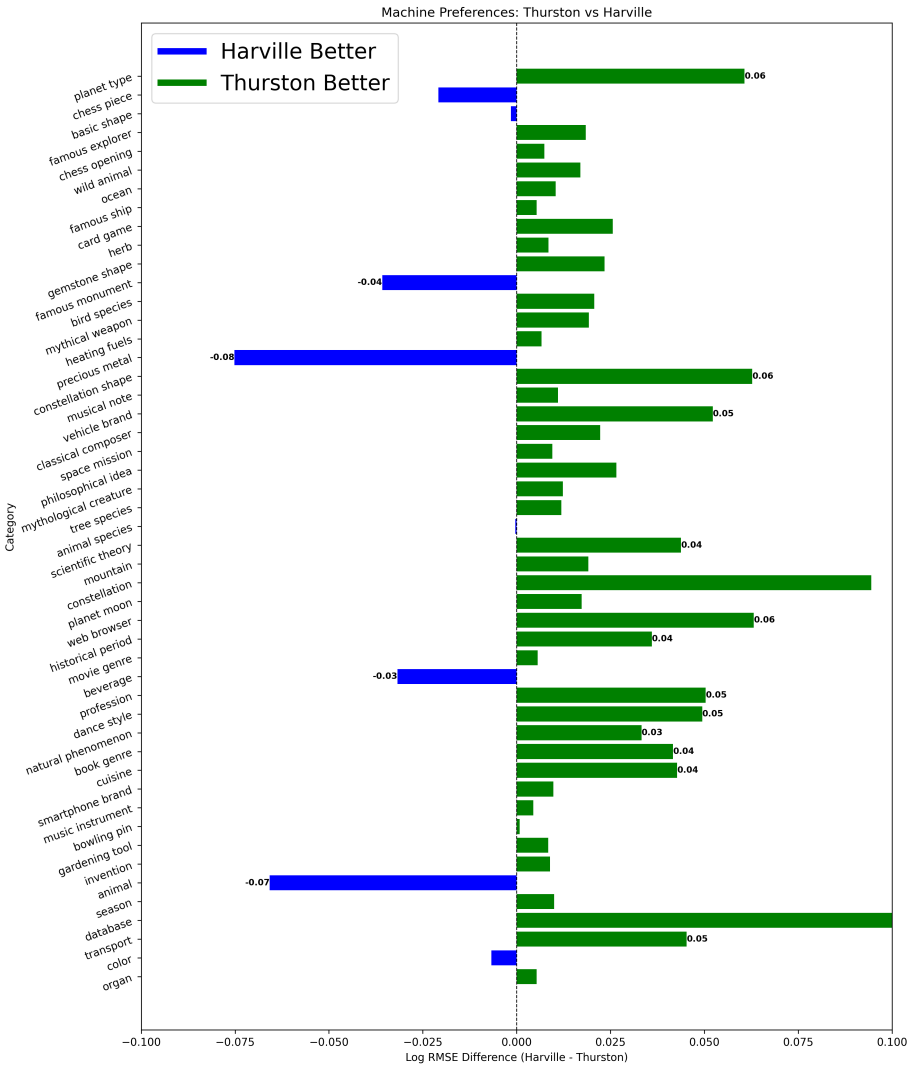

Figure 2: Average relative prediction errors for different question categories using RoBERTa. The Thurstonian model consistently outperforms Luce's Choice Axiom in predicting the LLM's token selection probabilities.

Acknowledgments

We thank the developers of BERT and RoBERTa for making their models accessible for research purposes.

Appendix: Prompt examples

Category	Original Prompt	Qualified Prompt
Organ	My favourite human organ is the [MASK] because it is SOMETHING.	My two favourite human organs are the [ANSWER] and the [MASK] because they are SOMETHING.
Tool	The most useful tool is a [MASK] because it is SOMETHING.	The two most useful tools are a [ANSWER] and a [MASK] because they are SOMETHING.
Flower	My favorite flower is the [MASK] because it is SOMETHING.	My two favorite flowers are the [ANSWER] and the [MASK] because they are SOMETHING.
Cuisine	The best cuisine is [MASK] cuisine because it is SOMETHING.	The two best cuisines are [ANSWER] and [MASK] because they are SOMETHING.
Holiday	My favorite holiday is [MASK] because it is SOMETHING.	My two favorite holidays are [ANSWER] and [MASK] because they are SOMETHING.
Book Genre	The most engaging book genre is [MASK] because it is SOMETHING.	The two most engaging book genres are [ANSWER] and [MASK] because they are SOMETHING.
Superhero	My favorite superhero is [MASK] because they are SOMETHING.	My two favorite superheroes are [ANSWER] and [MASK] because they are SOMETHING.
Planet Moon	My favorite moon in the solar system is [MASK] because it is SOMETHING.	My two favorite moons are [ANSWER] and [MASK] because they are SOMETHING.
Fruit	My favorite fruit is [MASK] because it is SOMETHING.	My two favorite fruits are [ANSWER] and [MASK] because they are SOMETHING.
Dessert	My favorite dessert is [MASK] because it is SOMETHING.	My two favorite desserts are [ANSWER] and [MASK] because they are SOMETHING.

Category	Original Prompt	Qualified Prompt
Mountain	The most impressive mountain is [MASK] because it is SOMETHING.	The two most impressive mountains are [ANSWER] and [MASK] because they are SOMETHING.
Scientific Theory	The most significant scientific theory is [MASK] because it is SOMETHING.	The two most significant scientific theories are [ANSWER] and [MASK] because they are SOMETHING.
Animal Species	The most intriguing animal species is the [MASK] because it is SOMETHING.	The two most intriguing animal species are the [ANSWER] and the [MASK] because they are SOMETHING.
Music Genre	My favorite music genre is [MASK] because it is SOMETHING.	My two favorite music genres are [ANSWER] and [MASK] because they are SOMETHING.
Tree Species	My favorite tree species is the [MASK] because it is SOMETHING.	My two favorite tree species are the [ANSWER] and the [MASK] because they are SOMETHING.
Programming Language	The best programming language is [MASK] because it is SOMETHING.	The two best programming languages are [ANSWER] and [MASK] because they are SOMETHING.
Dance Style	My favorite dance style is [MASK] because it is SOMETHING.	My two favorite dance styles are [ANSWER] and [MASK] because they are SOMETHING.
Instrumental Piece	My favorite instrumental piece is [MASK] because it is SOMETHING.	My two favorite instrumental pieces are [ANSWER] and [MASK] because they are SOMETHING.
City Landmark	The most famous landmark in the city is [MASK] because it is SOMETHING.	The two most famous landmarks are [ANSWER] and [MASK] because they are SOMETHING.
Space Mission	The most important space mission is [MASK] because it was SOMETHING.	The two most important space missions are [ANSWER] and [MASK] because they were SOMETHING.

Category	Original Prompt	Qualified Prompt
Painting	My favorite painting is [MASK] because it is SOMETHING.	My two favorite paintings are [ANSWER] and [MASK] because they are SOMETHING.
Martial Art	My favourite martial art is [MASK] because it is SOMETHING.	My two favourite martial arts are [ANSWER] and [MASK], because they are SOMETHING.
Mythological Weapon	The mythical weapon I admire the most is called [MASK] because it is SOMETHING.	The two most SOMETHING mythical weapons are called [ANSWER] and [MASK].
Famous Monument	My favorite monument is called [MASK] because it is SOMETHING.	The two most SOMETHING monuments are called [ANSWER] and [MASK].
Telescope	My favorite telescope is called [MASK]. I like it because it is SOMETHING.	The two most SOMETHING telescopes are called [ANSWER] and [MASK].
Ocean	The ocean I find most SOMETHING is called [MASK].	[ANSWER] and [MASK] are the two oceans that are particularly SOMETHING.
Vehicle Type	My favorite type of vehicle is [MASK] because it is SOMETHING.	My two favorite types of vehicles are [ANSWER] and [MASK] because they are SOMETHING.
Heating Fuels	My favourite heating fuel is [MASK] because it is SOMETHING.	The two favourite heating fuels are [ANSWER] and [MASK] because they are SOMETHING.
Precious Metal	The most desirable precious metal is [MASK] because it is SOMETHING.	The two most desirable precious metals are [ANSWER] and [MASK] because they are SOMETHING.
Art Medium	My favorite art medium is [MASK] because it is SOMETHING.	My two favorite art media are [ANSWER] and [MASK] because they are SOMETHING.

Renewable Energy Source	The most promising renewable energy source is [MASK] because it is SOMETHING.	The two most promising renewable energy sources are [ANSWER] and [MASK] because they are SOMETHING.

To examine the impact of small modifications to the questions, numerous modifications were used. Here are some examples of substitutinos for the string SOMETHING in the above questions.

Category	**Examples of SOMETHING**
Organ	vital, complex, essential, central, efficient, delicate, versatile, remarkable
Tool	versatile, essential, accurate, durable, reliable, efficient, portable, powerful, precise, sturdy, compact, user-friendly, lightweight, robust
Flower	fragrant, vibrant, delicate, beautiful, colorful, elegant, stunning, graceful, charming, enchanting
Cuisine	flavorful, spicy, savory, rich, aromatic, hearty, exotic, wholesome, delectable, diverse
Holiday	festive, joyful, relaxing, memorable, exciting, meaningful, traditional, cheerful, heartwarming, rejuvenating
Book Genre	intriguing, thought-provoking, thrilling, captivating, inspiring, immersive, entertaining, educational, imaginative, insightful
Superhero	courageous, inspiring, powerful, heroic, iconic, noble, selfless, determined, legendary, admirable
Tree Species	majestic, resilient, ancient, graceful, towering, sturdy, beautiful, vital, iconic, diverse
Programming Language	versatile, powerful, efficient, user-friendly, popular, robust, dynamic, expressive, flexible, modern
Dance Style	expressive, energetic, graceful, rhythmic, vibrant, dynamic, elegant, captivating, traditional, passionate
Art Medium	versatile, expressive, vibrant, tactile, dynamic, traditional, innovative, accessible, impactful, unique
Precious Metal	valuable, rare, lustrous, malleable, conductive, durable, sought-after, timeless, prestigious, versatile
Vehicle Type	efficient, spacious, fast, reliable, versatile, compact, rugged, luxurious, practical, economical
Natural Phenomenon	awe-inspiring, magnificent, breathtaking, mysterious, fascinating, beautiful, mesmerizing, grand, majestic, stunning
Mythological Creature	mythical, powerful, legendary, fascinating, mysterious, awe-inspiring, iconic, enchanting, formidable, captivating

Category	Examples of SOMETHING
Programming Framework	robust, scalable, efficient, flexible, powerful, modern, versatile, popular, intuitive, comprehensive
Historical Figure	influential, visionary, courageous, pioneering, inspiring, revolutionary, iconic, remarkable, impactful, legendary
Scientific Theory	foundational, groundbreaking, complex, elegant, profound, revolutionary, influential, comprehensive, intriguing, essential
Gemstone Shape	brilliant, intricate, popular, classic, timeless, elegant, unique, refined, symmetrical, beautiful

References

[1] Peter Cotton. Inferring Relative Ability from Winning Probability in Multientrant Contests. *SIAM Journal on Financial Mathematics*, 12, 2021.

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

[3] David A. Harville. Assigning probabilities to the outcomes of multi-entry competitions. *Journal of the American Statistical Association*, 68(342):312–316, 1973.

[4] Y. Liu, M. Ott, N. Goyal, et al. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint*, arXiv:1907.11692, 2019.

[5] R. Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. John Wiley & Sons, 1959.

[6] Robin L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(2):193–202, 1975.

[7] L. L. Thurstone. A law of comparative judgment. *Psychological Review*, 34(4):273–286, 1927.