Journal of Memory and Language 143 (2025) 104640

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Shades of zero: Distinguishing impossibility from inconceivability

Jennifer Hu [a,*], Felix Sosa [b], Tomer Ullman [a,b]

[a] Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, United States of America
[b] Department of Psychology, Harvard University, United States of America

ARTICLE INFO

Dataset link: https://github.com/jennhu/shades-of-zero

Keywords:
Impossibility
Inconceivability
Modal reasoning
Language models
Event knowledge

ABSTRACT

Some things are impossible, but some things may be even more impossible than impossible. Levitating a feather using one's mind is impossible in our world, but fits into our intuitive theories of possible worlds, whereas levitating a feather using the number five cannot be conceived in any possible world ("inconceivable"). While prior work has examined the distinction between improbable and impossible events, there has been little empirical research on inconceivability. Here, we investigate whether people maintain a distinction between impossibility and inconceivability, and how such distinctions might be made. We find that people can readily distinguish the impossible from the inconceivable, using categorization studies similar to those used to investigate the differences between impossible and improbable (Experiment 1). However, this distinction is not explained by people's subjective ratings of event likelihood, which are near zero and indistinguishable between impossible and inconceivable event descriptions (Experiment 2). Finally, we ask whether the probabilities assigned to event descriptions by statistical language models (LMs) can be used to separate modal categories, and whether these probabilities align with people's ratings (Experiment 3). We find high-level similarities between people and LMs: both distinguish among impossible and inconceivable event descriptions, and LM-derived string probabilities predict people's ratings of event likelihood across modal categories. Our findings suggest that fine-grained knowledge about exceedingly rare events (i.e., the impossible and inconceivable) may be learned via statistical learning over linguistic forms, yet leave open the question of whether people represent the distinction between impossible and inconceivable as a difference not of *degree*, but of *kind*.

Introduction

Some things are impossible. You cannot levitate a feather with your mind, no matter how hard you try. And yet, some things are more impossible than others. Levitating a feather with one's mind is in some sense easier than levitating a rock (McCoy & Ullman, 2019; Shtulman & Morgan, 2017). Such graded judgments of impossibility are the topic of ongoing study in cognitive science, philosophy, and cognitive development. The idea is that people's understanding of imaginary worlds is rooted in their understanding of the real world (Byrne, 2007), so studying what makes things easier or harder in the imagination can reveal people's understanding of everyday reality. As part of this research, there have been different explanations for what makes some things seem more impossible than others. The explanations are not mutually exclusive, and include (for example) moves across ontological hierarchies (Griffiths, 2015), causal violations (Shtulman & Morgan, 2017), perceived similarity (Goulding & Friedman, 2023), and violations of core knowledge and intuitive physics (Lewry et al., 2021; McCoy & Ullman, 2019).

But, just as there is a dividing line between the merely improbable and the truly impossible, there may be a category of events even more impossible than impossible. Levitating a feather with one's mind is impossible in our world, but can still be imagined as occurring in a fictional world, and fits into our intuitive theories of possible worlds (Lewis, 1986). In contrast, events like "levitating a feather using the number five" or "finding the square-root of a dog" cannot be evaluated, construed, or imagined in any possible world.[1] Borrowing from philosophy, we refer to such events as *inconceivable* (Gendler & Hawthorne, 2002). Our suggestion is that just as it has been fruitful for cognitive science to study people's understanding of the impossible, it is useful to study people's understanding of the inconceivable.

While the relationship between conceivability and possibility has been the topic of much philosophical research (e.g., Balog, 1999; Chalmers, 1996; Gendler & Hawthorne, 2002), there has been less empirical and computational cognitive science study of inconceivability. The most closely related line of work has investigated children's

* Corresponding author.
E-mail address: jenniferhu@fas.harvard.edu (J. Hu).

[1] Some readers may object that they can imagine various transformations of the example sentences to make sense of them, such as supposing "the number five" refers to a particular blow-drier in a set. We refer to such transformations as *coercions*, and return to them in the Discussion.

https://doi.org/10.1016/j.jml.2025.104640
Received 1 November 2023; Received in revised form 31 October 2024; Accepted 26 March 2025
Available online 15 April 2025
0749-596X/© 2025 Published by Elsevier Inc.

distinction between events that are *impossible* (e.g., walking on water) and events that are possible but highly *improbable* (e.g., growing a beard down to one's toes, or singing Jingle Bells at a birthday party) (e.g., Browne & Woolley, 2004; Goulding & Friedman, 2021; Goulding et al., 2023; Komatsu & Galotti, 1986; Lane et al., 2016; Shtulman, 2009; Shtulman & Carey, 2007). A main finding across these studies is that while even preschool-age children can distinguish impossible and probable events, they tend to judge improbable (or expectation-violating) events as *impossible*, and that the ability to distinguish the improbable and impossible is refined with age. This has been taken by some to suggest that young children try to conceive of an improbable event, fail, and judge their inability to conceive of it as evidence that the event cannot occur at all (Harris, 2021). However, such a distinction would not explain the impossible/inconceivable divide for adults: adults are easily able to imagine things they judge to be impossible, at least in our world. These findings in general leave open the question of whether adults naturally treat impossible and inconceivable as distinct modal categories (as they do for impossible and improbable), or whether people naturally treat inconceivability as simply an instance of impossibility.

We begin by investigating whether people can readily distinguish the impossible from the inconceivable, using categorization studies similar to those used to investigate the differences between improbable and impossible. These categorization tests follow the logic of the cognitive development studies mentioned above, which conclude that young children do not distinguish improbable from impossible events early in development, but refine their understanding over early childhood (e.g., Shtulman, 2009; Shtulman & Carey, 2007; Shtulman & Phillips, 2018). In our first experiment (Section "Experiment 1: Modal classification task"), we introduce a novel set of stimuli featuring event descriptions in four different modal categories (probable, improbable, impossible, and inconceivable), and validate them by examining whether people intuitively and reliably distinguish inconceivable event descriptions from event descriptions belonging to the other modal categories. We find that people are highly consistent in their categorizations, suggesting the set of inconceivable event descriptions is easily distinguished from the sets of impossible and improbable event descriptions.

Having established that people readily do make this distinction, we then examine how this distinction might be made. Broadly, there are two functional accounts for how the distinction between impossibility and inconceivability may work, corresponding to a distinction in kind, and a distinction of degree. A functional framework in support of the distinction in kind would argue that encountering the inconceivable is similar to the mind encountering a category error (Magidor, 2009; Ryle, 1949), leading to either termination of further processing, or the need to coerce the meaning of the original event. Category errors may in turn be similar to "type errors" encountered by type-based computer programs (Magidor, 2009; Sosa & Ullman, 2022). This general view also connects to a literature on type-theoretic distinctions in semantic theory (e.g., Bekki, 2014; Chatzikyriakidis & Luo, 2020; Luo, 2012; Montague, 1973; Sutton, 2024), and psycholinguistic studies of selectional restrictions (Chomsky, 1965; Katz & Fodor, 1963) in language processing (e.g., Paczynski & Kuperberg, 2012; Sitnikova et al., 2008; Warren & McConnell, 2007; Warren et al., 2015). An alternative possibility is that people have a single, graded notion of probability, where all modal events exist on a spectrum, including the improbable, impossible, and inconceivable. Each event is processed and assigned some probability of occurring, and modal categories are then read out by defining a straightforward transformation on top of the underlying probabilities. As a simplified example, one could define thresholds on probability values such that improbable events are one-in-a-hundred, impossible are one-in-a-million, and inconceivable are one-in-a-trillion. This view also connects to prior proposals that selectional restrictions can be defined based on statistical associations (Resnik, 1993, 1996), and evidence that world knowledge violations elicit similar patterns of neural activity as selectional restriction violations (e.g., Hagoort et al., 2004; Matsuki et al., 2011). Our experiments investigated two potential versions of the second account ("difference in degree"), but we do not resolve the debate of whether inconceivable differs from impossible in kind or degree. We return to the different mechanisms for distinguishing inconceivable and impossible in the Discussion (Section "Discussion").

In our second experiment (Section "Experiment 2: Subjective ratings of event likelihood"), we ask people for subjective ratings of the likelihood of the events from Experiment 1, on a continuous scale from 0 to 100. We find that people overwhelmingly rate impossible and inconceivable events at the bottom of the scale, and the ratings across these categories cannot be distinguished statistically. This suggests that people's modal distinctions between impossible and inconceivable are based on information beyond simple perceptions (or transformations) of event likelihood, which weighs against a basic version of the "graded" view — namely, that modal categories are read off based on thresholds upon probability.

We then investigate another version of the graded view by asking to what extent distinctions between modal categories can be made based on the statistical patterns of language. From a cognitive perspective, one proposal has been that people try to imagine or simulate an event occurring, and use the difficulty of imagination to judge the possibility of the event (Gendler & Hawthorne, 2002; Harris, 2021; Kahneman & Tversky, 1981). It is possible to argue that event descriptions that we frequently encounter are easier to simulate or imagine, and event descriptions that are very infrequently encountered will be harder to simulate. To investigate this view, we examine whether statistical language models (LMs) display similar behavioral signatures as humans regarding the distinction between inconceivable event descriptions and other modal categories. LMs are trained with the objective of predicting sequences of tokens, which they learn to do by observing vast amounts of text typically obtained from Internet posts, news datasets, and books. Through this paradigm, LMs may implicitly learn the latent properties of the world that make certain linguistic expressions more or less likely. For example, since people communicate about events (McRae & Matsuki, 2009) and physical observations (Louwerse, 2011, 2018), it may be reasonable to expect that language itself contains structured information about the world. Indeed, prior work has shown that LMs may capture important aspects of commonsense and world knowledge (see Chang & Bergen, 2023, for review), including the distinction between possible and impossible events (Kauf et al., 2023), and the structure of perceptual spaces (Abdou et al., 2021; Merullo et al., 2023; Patel & Pavlick, 2022). The autoregressive next-token-prediction objective used to train LMs also has connections to human behaviors during language comprehension, which involve prediction about upcoming linguistic content (e.g., Altmann & Kamide, 1999; Kutas & Federmeier, 2011; Levy, 2008; Shain et al., 2024; Smith & Levy, 2013) and the integration of generalized event knowledge (e.g., Bicknell et al., 2010; Matsuki et al., 2011; McRae & Matsuki, 2009). It is therefore argued that LMs may learn structured information about the world (and events occurring in the world) in service of optimizing the objective of next-word prediction.

From a machine learning perspective, the question of whether the distinction between impossibility and inconceivability can be learned based on statistical patterns is also highly of interest. If both impossible and inconceivable events occur with vanishingly small frequency, it is not obvious how LMs would learn to distinguish them in string probability space. Indeed, our work's focus on distinctions within the region of low probabilities – "shades of zero" – differs from two previous major approaches to evaluating LMs. The first prior approach focuses on distinguishing between broad categories which are expected to be associated with high or low probabilities — for example, by performing targeted comparisons between sentences that are grammatical/ungrammatical (e.g., Hu et al., 2020; Marvin & Linzen, 2018; Warstadt et al., 2020), or describe possible/impossible events (e.g., Kauf et al., 2023).

The second approach involves studying the behavior of LMs on naturally occurring sentences, which are expected to fall within the region of the string distribution with most probability mass — for example, by examining the alignment between model-derived string distributions and human reading times on naturalistic sentences (e.g., Brothers & Kuperberg, 2021; Hofmann et al., 2022; Shain et al., 2024; Smith & Levy, 2013). While these approaches are reasonable and useful, we suggest that studying distinctions within the near-zero region of the probability distribution – such as the distinction between sentences describing impossible and inconceivable events – can provide new insights into the patterns and commonsense knowledge learned by LMs, and how their representations of event likelihood relate to people's conceptual categories.

In our third experiment (Section "Experiment 3: Language model evaluation"), we test whether LMs separate the modal categories from our stimuli through the probabilities assigned to string event descriptions, and whether these probabilities align with humans' subjective ratings of event likelihood (as we measured in Experiment 2). We evaluate five Transformer-based language models from the GPT-2 (Radford et al., 2019) and Llama 3 (AI@Meta, 2024) model families, ranging in size from 124 million to 70 billion parameters, as well as a strong non-neural baseline (Liu et al., 2024). We find that modal categories are strongly separated within the event description probabilities estimated by the neural models, but not within the probabilities estimated by the non-neural baseline. Similarly, the subjective event likelihood ratings provided by humans in Experiment 2 are predicted by the neural models' probabilities (but not the baseline model's probabilities) – although models' probabilities vary widely across impossible and inconceivable event descriptions, while people assign these extremely low likelihood ratings. We additionally test how distinctions between modal categories emerge over the course of training for a single OLMo model (Groeneveld et al., 2024), and find that the distinction between inconceivable and impossible emerges before the distinction between impossible and improbable. These analyses contribute a new empirical investigation of commonsense and event knowledge in modern LMs, which is a central topic in artificial intelligence (Chang & Bergen, 2023; Li et al., 2022; Zhou et al., 2020). In particular, we demonstrate that LMs capture a distinction within the distribution of non-possible events (impossible versus inconceivable), which goes beyond previous studies focusing on distinctions either within the distribution of *possible* events (likely versus unlikely) or *across* the boundary of possible and impossible (e.g., Kauf et al., 2024, 2023; Wang et al., 2018).

Overall, our findings reveal high-level similarities between the way that humans and statistical language models treat inconceivability. First, both humans and models show an ability to categorize probable, improbable, impossible, and inconceivable event descriptions (comparing Experiments 1 & 3). Second, the probabilities assigned to event descriptions by neural LMs predict humans' subjective ratings of event likelihood (comparing Experiments 2 & 3). And third, the distinction between improbable and impossible emerges later during LM training than the distinctions between other pairs of categories, which is qualitatively consistent with findings in developmental psychology (e.g., Shtulman & Carey, 2007). However, our findings also raise questions about *how* humans make distinctions between impossible and inconceivable when their subjective likelihoods are indistinguishable (all near zero), and suggest future directions for investigating children's perceptions of inconceivability. We return to these issues in the Discussion (Section "Discussion"), once we have the results in hand.

Experiment 1: Modal classification task

We first developed a novel set of materials, which included short English-language event descriptions across four modal categories: probable, improbable, impossible, or inconceivable. Our main questions of interest are whether humans categorize events in a way that is consistent with the underlying coding of conditions, and what kind of error patterns humans make. This experiment also serves to validate the construction of our stimuli, which we use in later experiments.

Stimuli

We manually constructed a set of 70 items designed to cover several modal categories (see examples in Table 1). Each item consists of a shared prefix denoting a commonplace event with a transitive verb and object, as well as the beginning of a phrase describing how the event occurs (e.g., "baking a cake inside"). No explicit subject is specified. Each item prefix is associated with five candidate continuations, each of which completes the phrase describing how the event occurs. These continuations reflect different types of modal relationships to the prefix. Since the region that modulates the modal category of the event occurs at the end of the phrase, these stimuli are well-suited to test autoregressive (i.e., left-to-right) language models, which condition on preceding context to predict the subsequent tokens. While prior cognitive work on modal distinctions has primarily focused on probable, improbable, and impossible events, we additionally include inconceivable events. We describe each condition in greater detail below.

Continuations in the **Probable** and **Improbable** conditions reflected events that are possible given the physical laws of the real world. We refer to these as the "possible" conditions. Probable continuations were chosen to be prototypical or highly expected (e.g., "baking a cake inside an *oven*", or "washing your hair with *shampoo*"). Improbable continuations were intuitively unconventional, but did not violate any physical constraints (e.g., "baking a cake inside an *airfryer*", or "washing your hair with *detergent*").

In the **Impossible** condition, continuations make the event impossible because of physical constraints in the real world. For example, "baking cake inside a *freezer*" violates thermodynamics. This condition is most similar to what prior studies have called "violations of world knowledge" (cf. violations of *word* knowledge, or selectional restrictions). But, some instances do not neatly fit into this framework. For example, "washing your hair with *a rainbow*" could be a violation of world knowledge, or a violation of the restrictions on the verb "to wash" (i.e., [+LIQUID]).

Finally, we considered **Inconceivable** continuations, which take the form of an abstract concept or non-physical noun when the event described in the prefix expects a concrete noun (e.g., "baking a cake inside *a sigh*" or "chilling a drink using *yesterday*"). This condition is the most similar to selectional restriction violations in the linguistics literature (Chomsky, 1965; Katz & Fodor, 1963).

Methods

We recruited $N = 149$ participants on Prolific, who were based in the United States and native speakers of English by self-report. Participants were compensated at a rate corresponding to $12/h. Each participant saw each of the 70 items, one item per trial. On each trial, participants saw a prefix and a continuation in one of the four conditions (e.g., "Baking a cake using an oven"). Participants were asked to categorize the stimulus into one of four categories (by pressing keyboard buttons): probable, improbable, impossible, or nonsense. We used the label "nonsense" instead of "inconceivable" to avoid jargon, while capturing a similar intuition. After responding on each trial, participants saw a screen saying "Your response has been logged" for 1 s. Each participant saw a nearly equal number of trials in each condition (i.e., 17 or 18 trials for each of the 4 main conditions). The order of trials and conditions was randomized. Prior to the experiment, participants were familiarized with definitions and examples of each category, and were required to correctly complete 8 practice trials before beginning the critical trials.

Table 1
Sample items used in our experiments. Each item (row) consists of a shared prefix and a set of continuations across five conditions. The Inconceivable-Syntax condition was never used to evaluate humans, and only used in Experiment 3 as a baseline for the language model evaluation.

Prefix	Probable	Improbable	Impossible	Inconceivable	Inconceivable-Syntax
baking a cake inside	an oven	an airfryer	a freezer	a sigh	grasp
chilling a drink using	ice	snow	fire	yesterday	onto
washing your hair with	shampoo	detergent	a rainbow	applause	if
smashing a pumpkin using	a hammer	a boulder	a leaf	a number	remember

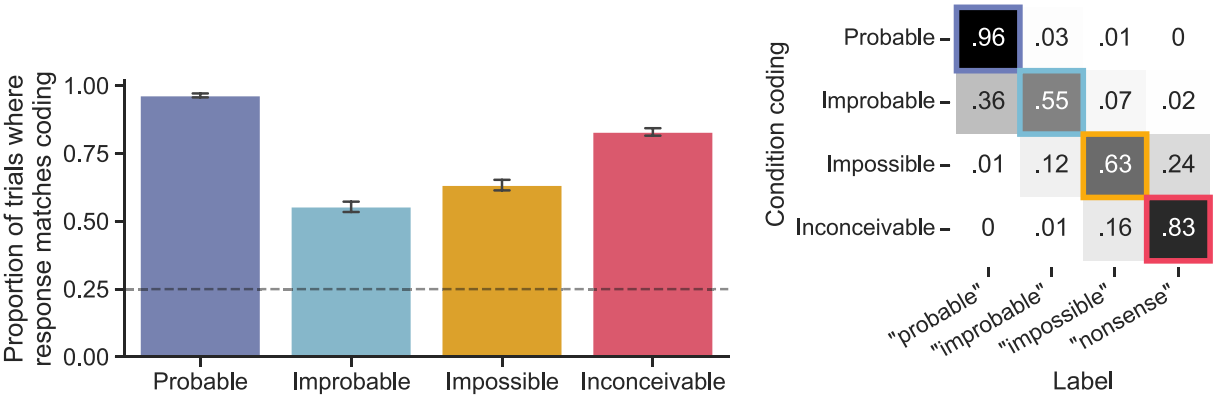

Fig. 1. Classification results from Experiment 1. (A) Proportion of trials where human responses matched the underlying condition coding. Dashed line indicates chance performance (25%). Error bars indicate bootstrapped 95% CI. (B) Distribution of responses within each condition. Highlighted cells indicate proportion of trials where responses match the condition coding (as shown in (A)).

Results

First, we ask whether humans categorize the events in a way that is consistent with the researcher-defined coding of the materials. Fig. 1(a) shows the proportion of trials where people's labels matched the underlying condition coding. In the Probable condition, participants agreed with our coding in nearly all cases (96%). Participants matched our coding for the overwhelming majority of cases in the Inconceivable condition (83%), and in the majority of cases for the Improbable (55%) and Impossible (63%) conditions. These results suggest that people's judgments align with our stimulus coding at a broad level, validating the construction of our materials. Furthermore, this provides evidence that people can intuitively categorize inconceivable items, mirroring and extending the findings of prior cognitive studies showing that people distinguish the impossible from the improbable (e.g., Shtulman & Carey, 2007).

Next, we ask what kind of choices people make, when they categorize events in a way that is inconsistent with our researcher-defined labels. Fig. 1(b) shows the distribution of responses in each condition, averaged across trials. The errors (i.e., trials where the participant's label of an event does not match our coding) are not random, but reflect structured patterns of confusability between neighboring modal categories. For example, when participants did not label the Improbable items as improbable, they were most likely to label them as probable (36%). This is reasonable, as the distinction between probable and improbable is a matter of degree, and there is inherent imprecision of the term improbable. In addition, subjective estimates of the probabilities of possible (i.e., probable or improbable) events may vary based on each individual's experiences and environments.

We also found evidence for confusability between modal categories which, *a priori*, might be expected to have firm boundaries. For example, Improbable items were labeled as impossible 7% of the time, and Impossible items were labeled as improbable 12% of the time. While the majority of responses label Impossible items as impossible and Inconceivable items as nonsense, we also observe some confusability between impossibility and inconceivability. 24% of responses labeled Impossible as nonsense, and 16% labeled Inconceivable as impossible. These patterns suggest that, while people do readily distinguish inconceivable from impossible (and other modal categories), the boundaries between these categories may be graded. In the next experiment, we more directly test this idea by investigating whether people assign graded ratings of likelihood to impossible and inconceivable events.

Experiment 2: Subjective ratings of event likelihood

In our second experiment, we asked human participants to rate how likely events are to occur, using a slider scale. This task tests a basic version of the view that the distinction between impossibility and inconceivability is a difference in *degree*. If people's modal distinctions between impossible and inconceivable are based on simple thresholds on top of subjective event likelihood, then we would expect to find separation between the likelihood ratings assigned to impossible and inconceivable events. If there is no separation in these ratings, then this would suggest that people's distinctions are made on another basis.

Methods

We used the same set of stimuli as in Experiment 1. We recruited $N = 50$ US-based participants on Prolific, with a self-reported native language of English. Participants were compensated at an hourly rate of $12. Each participant saw each of the 70 items, one item per trial. On each trial, participants saw a question of the form "How likely is it for someone to [EVENT DESCRIPTION]?", where the event description follows one of the four conditions (e.g., "How likely is it for someone to bake a cake inside a sigh?"). Their task was to respond using a slider with endpoints marked "Extremely unlikely" and "Extremely likely", which were internally coded as 0 and 100. No increments or numbers were shown on the slider. Each participant saw a roughly equal number of trials in each condition (i.e., either 17 or 18 trials for each of the 4 conditions).

We normalized participants' responses within items, to control for people's perceived prior probabilities of the prefixes, independent of

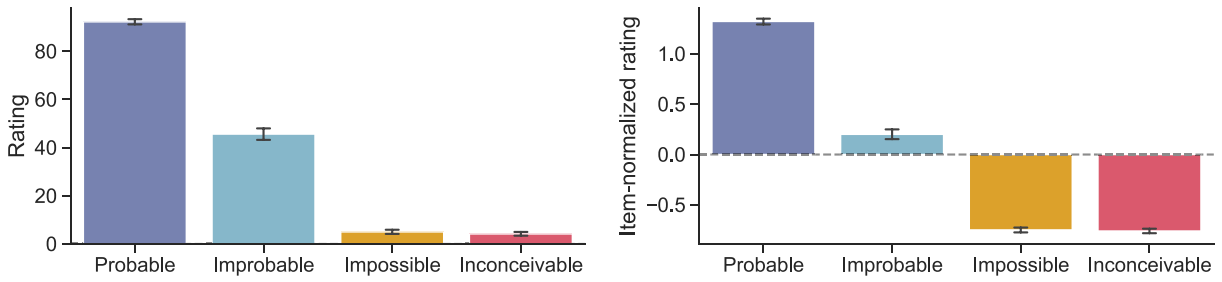

Fig. 2. Subjective ratings of event likelihood from Experiment 2. (A) Raw and (B) normalized (within-item) ratings, averaged over items in each condition. Error bars denote bootstrapped 95% CIs.

the continuations. For example, "baking a cake" might be more likely *a priori* than "docking a boat", and so it could be the case that someone would rate the Probable continuation of "docking a boat" as less likely than the Improbable continuation of "baking a cake".

Results

Figs. 2(a) and 2(b) show raw and item-normalized ratings across conditions, respectively. We first turn to the raw ratings (on a 0–100 scale). The Probable event descriptions received near-ceiling likelihood ratings (mean rating: 92.2), the Improbable items were near the middle (mean rating: 45.6), and the Impossible and Inconceivable items were near the floor of the slider scale (Impossible mean rating: 5.0; Inconceivable mean rating: 4.2). We find a similar pattern for the within-item normalized ratings as well (Fig. 2(b)). Importantly, the Impossible and Inconceivable event descriptions did not differ in their rating distributions (item-normalized ratings; two-sample Kolmogorov–Smirnov test: $D = 0.04, p = 0.37$; Mann–Whitney U test: $U = 379911.5, p = 0.79$).

To summarize our results thus far, we have found that people can reliably distinguish impossible and inconceivable events (Experiment 1), but they assign them indistinguishable near-zero likelihood of occurring (Experiment 2). This suggests that people's categorizations are not based purely on differences in perceived likelihood, which rules out a basic "threshold-probability" version of the view that the distinction between impossibility and inconceivability is one of *degree*. However, these results do not rule out the graded account entirely, nor do they provide direct support for the view that this distinction is a difference in *kind* (analogous, for example, to a type error). We test an alternate version of the graded account in the next experiment, by investigating the treatment of model categories in string probabilities estimated by statistical language models. We return to the open question of how people make the distinction between impossible and inconceivable in the Discussion (Section "Discussion").

Experiment 3: Language model evaluation

Experiment 1 established that people can intuitively distinguish between events in the categories of probable, improbable, impossible, and inconceivable. Experiment 2 showed that people assign essentially equal (near-zero) subjective likelihood ratings to sentences describing impossible and inconceivable events. In our third experiment, we investigate whether statistical language models (LMs) treat these modal distinctions in a similar way to humans: do the string-level probabilities of event descriptions separate modal categories, even impossible and inconceivable? And do these probabilities align with humans' subjective ratings of event likelihood?

Methods

Stimuli

We use the same stimuli as in Experiments 1 and 2. However, we add an additional condition as a baseline: Inconceivable-Syntax. This condition features a continuation that makes the full phrase ungrammatical, typically by featuring a verb, adverb, or preposition instead of the expected noun (e.g., "chilling a drink using *at*"). We emphasize that this condition is not the focus of our investigation of impossibility versus inconceivability — instead, it serves as a baseline, since we expect LMs to recognize these expressions as ill-formed.[2] In the context of the current experiment, this means that we expect LMs to reliably assign lower probabilities to these continuations, as would be consistent with a large body of work on targeted syntactic evaluation of LMs (e.g., Hu et al., 2020; Marvin & Linzen, 2018; Warstadt et al., 2020). Examples of the Inconceivable-Syntax condition are shown in Table 1. Again, this condition is only used for the model evaluation, and not the human experiments.

Models

We manipulate two properties that might affect models' ability to differentiate between the categories of interest: size (number of parameters), and training time (see Table 2).

For our investigation of model size, we evaluated six open-source language models: five neural network-based language models, and one n-gram baseline. We include the n-gram baseline to capture a statistical LM based purely on co-occurrence counts in a large corpus. If this model captures human-like behaviors in distinguishing the modal categories, this suggests that these distinctions could be explained by different frequencies of direct experience with these events. If not, this would suggest that the neural network-based LMs are doing something more sophisticated than tracking frequency counts or token co-occurrences to assign probabilities to event descriptions.

Our five neural models include three size variants of the GPT-2 family (Radford et al., 2019) and two size variants of the Llama-3 family (AI@Meta, 2024). Collectively, the models range in size from 124 million to 70 billion parameters, spanning multiple orders of magnitude. Each neural model is an autoregressive language model based on the Transformer architecture (Vaswani et al., 2017), and publicly available through the Huggingface Transformers library (Wolf et al., 2020).

As a strong non-neural baseline, we evaluate the infini-gram model ("∞-gram"; Liu et al., 2024), which estimates arbitrarily large n-gram counts based on the Dolma-v1.7 corpus (Soldaini et al., 2024). This corpus contains 2.6 trillion tokens created by the Llama-2 tokenizer (Touvron et al., 2023), substantially larger than other corpora commonly

[2] Whether category errors constitute a syntactic error is a long-standing debate that we do not resolve here; see Magidor (2013).

Table 2
Details of language models evaluated in Experiment 3. (A) Models tested in size manipulation experiment. Model identifiers correspond to Huggingface pretrained model paths for the neural models, and the API index for ∞-gram. (B) Checkpoints of OLMo tested in training time experiment.

(a)

Model name	Model family	Model identifier	# parameters	Training data size
∞-gram	*n*-gram (Liu et al., 2024)	v4_dolma-v1_7_llama	–	2.6 T tokens
GPT-2	GPT-2 (Radford et al., 2019)	gpt2	124 M	40 GB text
GPT-2 Med	GPT-2 (Radford et al., 2019)	gpt2-medium	355 M	40 GB text
GPT-2 XL	GPT-2 (Radford et al., 2019)	gpt2-xl	1.5 B	40 GB text
Llama-3 8B	Llama-3 (AI@Meta, 2024)	meta-llama/Meta-Llama-3-8B	8 B	15 T tokens
Llama-3 70B	Llama-3 (AI@Meta, 2024)	meta-llama/Meta-Llama-3-70B	70 B	15 T tokens

(b)

Checkpoint	Model identifier	# parameters	# steps	# training tokens
1	allenai/OLMo-7B-0424-hf	7B	500	2B
2	allenai/OLMo-7B-0424-hf	7B	1000	4B
3	allenai/OLMo-7B-0424-hf	7B	2000	8B
4	allenai/OLMo-7B-0424-hf	7B	4000	16B
5	allenai/OLMo-7B-0424-hf	7B	8500	35B
6	allenai/OLMo-7B-0424-hf	7B	17 500	73B
7	allenai/OLMo-7B-0424-hf	7B	36 000	150B
8	allenai/OLMo-7B-0424-hf	7B	72 500	303B
9	allenai/OLMo-7B-0424-hf	7B	151 500	635B
10	allenai/OLMo-7B-0424-hf	7B	477 000	2T

used to estimate *n*-grams, such as COCA (Davies, 2008) or Google Books (Juola, 2022).

For our investigation of training time, we evaluated 10 geometrically-spaced training checkpoints of OLMo-7B (Groeneveld et al., 2024), ranging from 500 steps (corresponding to 2 billion training tokens) to 477,000 steps (corresponding to 2 trillion tokens). OLMo was also trained on Dolma-v1.7, the same corpus used to estimate *n*-gram counts and train the ∞-gram model.

Estimating surprisal

We use the same stimuli as in Experiment 1 (see Table 1 for examples). To estimate how (un)expected an event is, we measure the surprisal S (Hale, 2001; Levy, 2008), or negative log probability, of the continuation c conditioned on its prefix p:

$$S(c|p) := -\log P(c|p) \tag{1}$$

If a model has learned to represent event probabilities in a way that conforms to the normative coding of our stimuli, then it should assign lowest surprisal to Probable continuations, higher surprisal to Improbable continuations, and highest surprisal to Impossible and Inconceivable continuations.

On a technical note, we are interested in probabilities assigned to words or multi-word expressions, while models compute probabilities at the level of *tokens*. To obtain the log probability of a continuation, we take the mean log probability over all tokens within the continuation.[3] This controls for the potential confound of length, as continuations that are split into more tokens will likely have lower probabilities overall.

To deal with sparsity issues for *n*-gram estimates, ∞-gram performs a version of backoff where the model conditions on the longest suffix in the prefix that appears in the training data. Since this backoff method still leaves open the possibility of zero counts, we also use additive smoothing to deal with sparsity. That is, given corpus $\mathcal{D}$ and vocabulary V, the ∞-gram probability of the ith token t_i in a sequence is given by

$$P_{\infty\text{-gram}}\left(t_i|t_1,\ldots,t_{i-1}\right) := \frac{\text{count}\left(t_{i-(n-1)},\ldots,t_{i-1},t_i|\mathcal{D}\right)+k}{\text{count}\left(t_{i-(n-1)},\ldots,t_{i-1}|\mathcal{D}\right)+k|V|}, \tag{2}$$

where

$$n := \arg\max_{n'\in[1,i]}\left\{\text{count}\left(t_{i-(n'-1)},\ldots,t_{i-1}|\mathcal{D}\right)>0\right\}. \tag{3}$$

In practice, we use $k = 0.1$ as a compromise between full pseudocounts ($k = 1$) and no smoothing ($k = 0$), taking into account the large vocabulary size of the tokenizer (under which larger k might distort the probability estimates more).

Finally, by comparing the surprisal of different continuations conditioned on the same prefix, there is the potential confound of observing differences driven by frequency effects. For example, we might observe

$$S(\text{an airfryer} \mid \text{Baking a cake using}) > S(\text{an oven} \mid \text{Baking a cake using}), \tag{4}$$

simply because "an airfryer" (the Improbable continuation) occurs less frequently than "an oven" (the Probable continuation) in text corpora. One way to address this would be to counterbalance the items, such that each continuation is seen in every condition, which has been a standard practice in other LM evaluations. However, our manipulation of inconceivability relies on pairing concrete-action prefixes with abstract continuations, meaning that the inconceivable continuations cannot be paired with any prefix that would make the described event conceivable. Therefore, we instead validate our materials by comparing the *n*-gram frequency counts of each continuation across conditions, and test for whether the condition significantly affects the surprisal values after controlling for *n*-gram frequencies. We estimated the *n*-gram frequency counts of the continuations using Dolma-v1.7 (2.6 trillion tokens), the same corpus used to train the ∞-gram model.

Results

Distinguishing modal categories

First, we compare LMs' probability distributions to people's categorization behaviors in Experiment 1: do LM surprisals distinguish between the modal categories of probable, improbable, impossible, and inconceivable?

[3] We appended a period at the end of each continuation when evaluating the neural LMs, in order to signal the end of a sentence. This is done because in certain cases the event description could be continued in a way that changes the modality of the event (e.g., "washing your hair with applause" could be continued with "playing on a speaker in the bathroom"). We did not include the final period when evaluating the ∞-gram model or estimating *n*-gram frequencies in order to avoid sparsity issues.

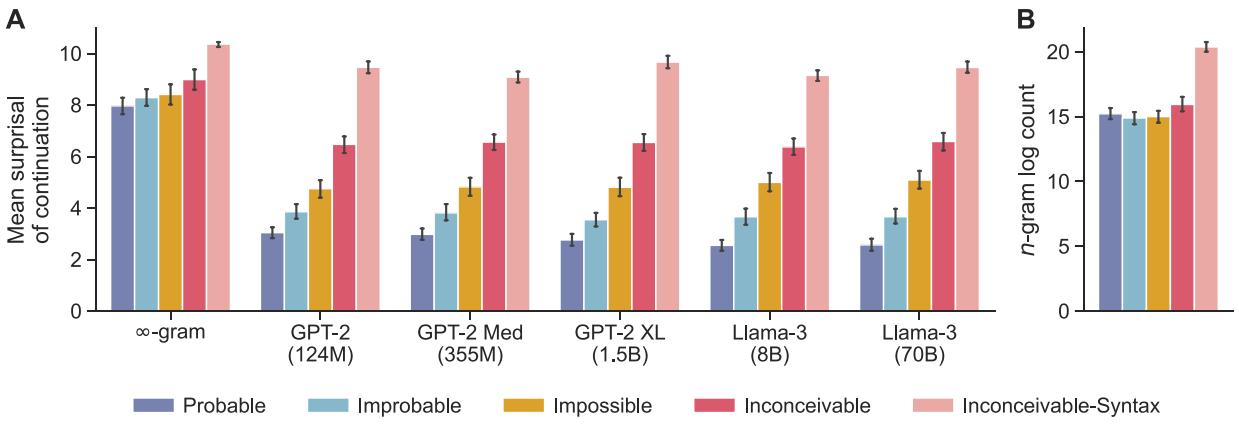

Fig. 3. (A) Mean surprisal values (i.e., negative log probability averaged over tokens) assigned by our tested language models to continuations in each condition. (B) Log *n*-gram count of continuations in each condition, estimated based on Llama-2 tokenization of Dolma-v1.7 (Liu et al., 2024). Error bars denote bootstrapped 95% CIs.

Size manipulation. Fig. 3A shows the surprisal values for continuations in each condition, averaged across items. Numerically, for each model, the condition-level mean surprisals are ordered in the following way (mirroring the ordering of the x-axis): Probable < Improbable < Impossible < Inconceivable < Inconceivable-Syntax. We additionally analyzed pairwise comparisons of surprisals across conditions *within* each item (Appendix A.1, Fig. 6), and the item-level surprisal comparisons also largely follow the condition ordering discussed above, except for the ∞-gram model. Turning to the frequency counts of each continuation, Fig. 3B shows the mean log count of occurrences of each *n*-gram across conditions. The counts are in fact higher in the Inconceivable and Inconceivable-Syntax conditions than the other conditions, which should bias surprisal values to be *lower* in these conditions, since LMs are sensitive to frequency effects (McCoy et al., 2024). If we find higher surprisals in the Inconceivable conditions relative to the others, this could not be explained purely by frequencies, suggesting that LMs are representing meaningful distinctions across the modal categories.

For each tested language model, we fit a linear mixed-effects regression model using the `lme4` package in R to test the effect of condition and *n*-gram frequency counts. We additionally included a random intercept for each item, as specified by the following formula:

$$\texttt{meanSurprisal} \sim \texttt{condition} + \texttt{logNgramCount} + (1 \mid \texttt{item}) \quad (5)$$

The condition variable was backward-difference coded, such that the mean surprisal for one condition is compared to the mean surprisal of the prior adjacent condition. In other words, the variable has four levels, corresponding to the following four pairwise comparisons: (1) Probable < Improbable; (2) Improbable < Impossible; (3) Impossible < Inconceivable; and (4) Inconceivable < Inconceivable-Syntax. For each neural model, we find a strong positive effect of each these pairwise condition comparisons. For the baseline ∞-gram model, we find a small but significant effect for the Inconceivable < Inconceivable-Syntax comparison, but no effect for the other three condition comparisons. Full results from each regression model are shown in Appendix A.2, Table 3.

Overall, these results suggest that the modal categories from our materials (probable, improbable, impossible, inconceivable) can be distinguished from each other in string probability space. In addition, the success of the neural models in contrast to the failure of the ∞-gram – the strongest available *n*-gram-based baseline – suggests that this ability is not simply based on word co-occurrences.

Training time manipulation. Next, we analyze the separability of modal categories over training time of a single language model: OLMo-7B. Fig. 4 shows the mean surprisal assigned to continuations in each condition across training checkpoints. First, we note that the syntactic violations (Inconceivable-Syntax) separate the earliest during training, which is perhaps unsurprising. Interestingly, the separation between Improbable/Impossible seems to occur later during training than the separation between Impossible/Inconceivable and Probable/Improbable. The relationship between Probable, Improbable, and Impossible seems broadly consistent with patterns of findings in the developmental literature (Shtulman, 2009; Shtulman & Carey, 2007), where younger children struggle to distinguish improbable and impossible events. Our findings also suggest a novel prediction that has been largely unexplored in child psychology, and could be tested in future experiments: the distinction between impossible and inconceivable events would occur earlier in development than the distinction between improbable and impossible.

Surprisal versus human ratings

Next, we compared the surprisal values derived by language models and the human ratings measured in Experiment 2 (Section "Experiment 2: Subjective ratings of event likelihood"). To analyze the predictive power of model surprisals, we fit the following linear mixed-effects regression model for each language model:

$$\texttt{rating} \sim \texttt{surprisal} + \texttt{logNgramCount} + (1 \mid \texttt{item}) + (1 \mid \texttt{subject}) \quad (6)$$

Overall, the model surprisals strongly predict human ratings, even when taking frequencies into account (∞-gram: $t = -13.83, p < 2e-16$; GPT-2: $t = -48.193, p < 2e-16$; GPT-2 Med: $t = -49.668, p < 2e-16$; GPT-2 XL: $t = -52.450, p < 2e-16$; Llama-3 8B: $t = -58.953, p < 2e-16$; Llama-3 70B: $t = -57.193, p < 2e-16$). The ordering of models in terms of predictive power generally aligns with standard notions of model quality: we find the weakest effect for ∞-gram, and the effect sizes slightly increase with the parameter count of the neural models.

Fig. 5 shows the mean surprisal assigned to continuations in each condition on the *x*-axis, versus humans' subjective ratings on the *y*-axis. While the linear relationship captures the overall trend across conditions, it does not capture the relevant distinction between impossible and inconceivable: there appears to be substantial variation in LM surprisal for the Impossible and Inconceivable event continuations (yellow and red points), whereas humans rated most of these as having near-zero likelihood. We return to this issue in the Discussion (Section "Discussion").

Discussion

While a great deal of recent research in cognitive science has studied the distinctions between and within the improbable and the impossible, far less work has empirically or computationally examined people's reasoning about the inconceivable. Here, we investigated inconceivable events as a distinct mental category, as a way of studying more

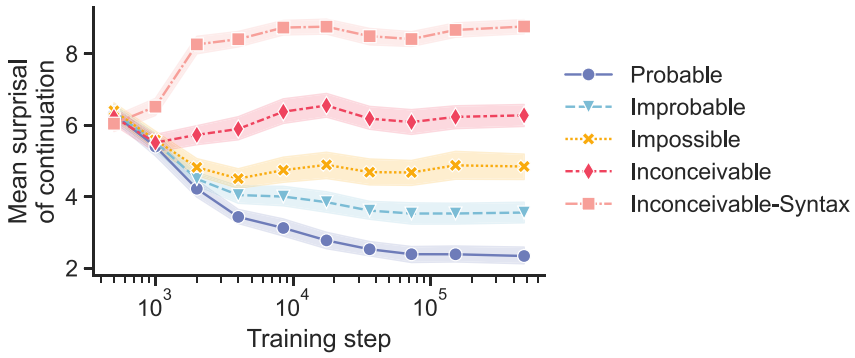

Fig. 4. Mean surprisal values assigned to continuations in each condition by 10 intermediate checkpoints of OLMo-7B. Error bands indicate bootstrapped 95% CIs.

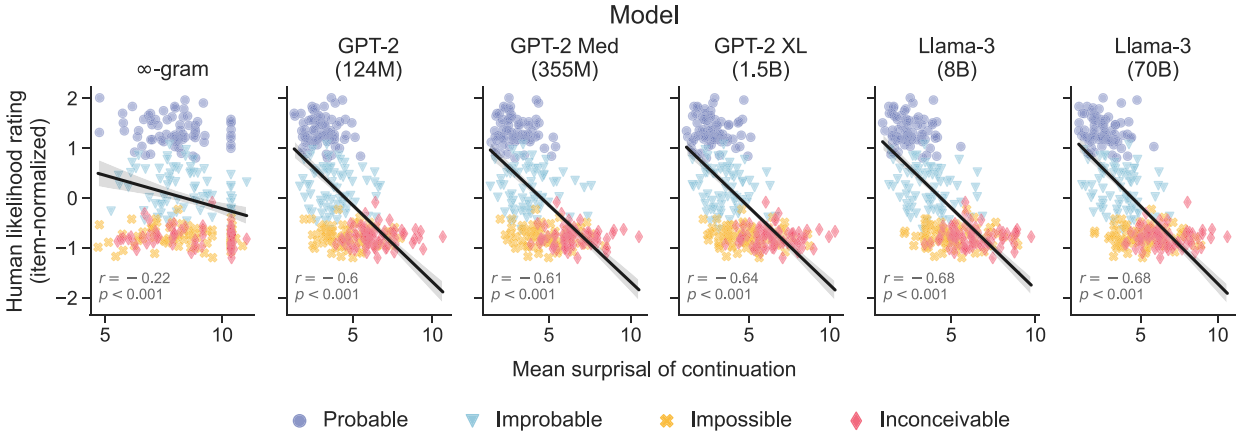

Fig. 5. Human ratings (y-axis) versus surprisal assigned by language models to continuations in each condition (x-axis). Subplots are annotated with Pearson *r* correlation coefficients. While the linear correlation captures the overall trend, it does not capture the relevant distinction between impossible (yellow x's) and inconceivable (red diamonds).

generally how people perform modal reasoning about (im)possibility. We found that people reliably distinguished inconceivable events from other modal categories, including impossible ones. These results expand on an ongoing research program in cognitive science and cognitive development that examines people's varying judgments of impossible events, and their distinction from the merely improbable (e.g., Browne & Woolley, 2004; Goulding & Friedman, 2021; Goulding et al., 2023; Komatsu & Galotti, 1986; Lane et al., 2016; Shtulman, 2009; Shtulman & Carey, 2007). However, using a rating task, we also found that people did not distinguish impossible and inconceivable events through their subjective likelihood of occurring. We also found that neural language models – but not a strong non-neural baseline – distinguished between impossible and inconceivable event descriptions through their string probabilities. These probabilities broadly align with people's subjective ratings of event likelihood, but not within the categories of impossible and inconceivable events.

As mentioned in the Introduction, previous theoretical and empirical work points to two broad functional accounts of how impossibility and inconceivability might be distinguished: a distinction in *kind*, or a distinction in *degree*. For simplicity, we will refer to these two accounts as View 1 and View 2, respectively. Given these two accounts, the novel finding that graded event probabilities (estimated by language models) can capture people's modal judgments could be explained in several ways: (a) people and LMs are performing convergent computations, either aligned with View 1 or View 2; or (b) people and LMs use different computations but arrive at the same outcome. We note that while our experiments do not fully tease these hypotheses apart, our findings are most consistent with a version of (b) where people do not make modal distinctions simply on the basis of event likelihood, but the statistics of language use are sufficient to separate descriptions of impossible and inconceivable events. We next discuss View 1 and View 2 in more depth, briefly surveying previous work that suggest *a priori* reasons in favor of both accounts. After establishing this background, we then discuss potential explanations for our empirical findings.

Several lines of previous work in cognitive development (e.g., Shtulman & Carey, 2007) and philosophy (e.g., Gendler & Hawthorne, 2002; Magidor, 2013) align with View 1 (distinction in kind). Impossible events such as levitating a cow are nomologically impossible qua physics, but still conceivable as occurring in some possible world. By contrast, imagining something that is both a cow and not a cow is a logical violation that does not even get off the ground. Computationally, one proposal for a distinction in kind between impossible and inconceivable would be that judgments of inconceivability follow the experience of a category error, analogous to type errors in typed computer programs (Magidor, 2009; Sosa & Ullman, 2022). To see this, consider the expression "square_root(45)". Such an expression can be evaluated in different ways by different programs, depending on algorithm they use, but all reasonable programs designed to handle most math expressions will evaluate the expression. But, contrast the previous expression with "square_root('dog')". Such an operation will be rejected by many programs as a type violation, as 'dog' is simply not the kind of thing you can apply the square-root program to. Some researchers have proposed that types, in the computer science sense that enforces the expected inputs and outputs of a program, may form

the basis of mental computation across domains and behaviors (e.g., Morales, 2018; Sosa & Ullman, 2022). In much the same way that a computer program for calculating square-roots expects a number and would throw an error when encountering a string, the mind may expect "baking a cake using" to be followed by a 'physical object' type, and throw an error if it is followed by the wrong type (e.g. "the number three"). The argument would then go that inconceivable events are naturally categorized as such by the mind encountering a type error ("imagine levitating the number 5" → type error → categorize as inconceivable), as opposed to impossible events, which *can* be evaluated by a mental program. We note as an aside that given this view, and to the degree that the mind contains multiple modules equivalent to different programs, the same event or expression may be categorized as inconceivable or not, depending on the module that is evaluating it.

This general view connects to a literature on type-theoretic distinctions and computations in semantic theory (e.g., Bekki, 2014; Chatzikyriakidis & Luo, 2017, 2020; Luo, 2012; Montague, 1973; Sutton, 2024). Type theory and predicate logic provide building blocks which are fundamental to proposed compositions in natural language, such as entities, first- and higher-order predicates, properties, and propositions. Building on this connection, a categorical distinction between impossibility and inconceivability has also been investigated from a language-processing perspective. In these studies, the general approach is to measure how people process sentences that violate different types of knowledge. Sentences that violate "world knowledge" correspond to impossible events, as we can evaluate them as being false in our world, while being able to conceive of possible worlds or circumstances in which they are true. However, sentences that violate "word knowledge" (e.g., selectional restrictions; Chomsky, 1965; Katz & Fodor, 1963) might correspond to inconceivable events, as it is difficult to derive a literal meaning for them by combining the lexical items.[4] For example, consider the following sentences, adapted from Warren et al. (2015):

(1)

a. The hamster lifted the refrigerator.
b. The hamster entertained the refrigerator.

Given standard interpretations of the terms, the events in both (1-a) and (1-b) are impossible: a hamster cannot *lift* a refrigerator, nor can a hamster *entertain* a refrigerator. However, (1-a) is impossible because of our physical knowledge about the world, while (1-b) constitutes a selectional restriction violation (SRV) because refrigerators – by virtue of being inanimate – cannot be entertained. Some theories of sentence comprehension propose that selectional restrictions are part of lexical knowledge, which is used more rapidly during the time-course of processing than world knowledge (e.g., Bornkessel & Schlesewsky, 2006; van Gompel et al., 2005). For example, eye-tracking studies have shown that SRVs appear to trigger processing difficulties that cannot entirely be explained by the plausibility or impossibility of the event (e.g., Rayner et al., 2004; Warren & McConnell, 2007; Warren et al., 2015). From the neural perspective, violations of selectional restrictions have also been shown to elicit distinct patterns of event-related potentials, compared to general world knowledge (e.g., Paczynski & Kuperberg, 2012; Sitnikova et al., 2008).[5] For example, Paczynski and Kuperberg (2012) find that world knowledge violations (e.g., "The pianist played his music while the bass was strummed by the drummer during the song") evoke an N400 effect but no P600, whereas animacy-based SRVs (e.g., "The pianist played his music while the bass was strummed by the drum during the song") evoke both an N400 and P600. While our focus here is on modal reasoning rather than sentence comprehension, these studies provide behavioral and neural evidence for a distinction that is conceptually analogous to the distinction between impossibility and inconceivability.

In contrast to the separation in kind, View 2 suggests that the primary representation is a kind of hazy probability that never quite gets to zero, and has no principled distinctions of kind between impossible, improbable, and inconceivable. What we often term "impossible" is simply very improbable, and what we term "inconceivable" is simply very impossible. Further distinctions like "violations of causal theories" are then post-hoc rather than primary. If this view is true, it would be a rather radical departure from how modal distinctions have been proposed to operate in people. However, analogues of the graded view have been explored in the language processing literature. Prior studies have found that world knowledge violations elicit similar patterns of neural activity as SRVs (Hagoort et al., 2004; Matsuki et al., 2011), and event-based plausibility plays a rapid and important role in online sentence comprehension (e.g., Garnsey et al., 1997; Kamide et al., 2003; McRae et al., 1998; Van Berkum et al., 2005). Beyond this, the lexical/world knowledge distinction might also be dis-preferred on the basis of parsimony. Without theoretically motivated constraints, it seems there could essentially be as many selectional restrictions (or types, in the type-theoretic version of View 1) as there are verbs. For example, the verb "to mail" can only take objects that are "mailable" (Myers & Blumstein, 2005), "to inflate" can only take objects that are "inflate-able", and so on (Matsuki et al., 2011; McRae et al., 1997). This becomes even more pronounced in our earlier example of "levitating a feather with the number five": "levitating a feather" can only be done using *the kinds of things that can be used to levitate*, such as one's mind, a magical incantation, or an instantaneous gravity-reversing beam of energy.

Having discussed previous support for Views 1 and 2, we now turn to how they bear on our empirical results. Broadly, we found similarities in people's and LMs' behaviors: both people and LMs can distinguish modal categories, including impossible and inconceivable. One potential explanation is that people and LMs achieve this behavior through similar kinds of computations. Within this explanation, one sub-option is that they are both aligned with View 1: the primary computation is one of causal violations, type errors, intuitive theories, and other such distinctions of kind. The output of such computations then appears graded, due to probabilistic thresholds on top of these computations. The work surveyed above – bridging semantic theory, cognitive science, and psycholinguistics – suggests that, *a priori*, these kinds of computations may support modal distinctions in people. If LM behaviors also operate this way, the internal computations and representations learned by language models in order to arrive in their judgments are quite sophisticated, having learned structured commonsense theories about domains such as biology and physics, and possibly type-based reasoning. If this view is true, it would be a rather radical departure from a common view in cognitive science that language models have not learned such representations, and it is also non-obvious how such models could be implementing type-based reasoning, dating back to the original debates surrounding connectionism and symbolic representations (e.g., Fodor & Pylyshyn, 1988; Pinker & Prince, 1988; Ramsey, 1997; Rumelhart & McClelland, 1986; Smolensky, 1988). While our findings do not definitively rule out that LMs are performing type-based reasoning, this view makes the least contact with our data, and would require substantial further investigation.

Another sub-option is that people and LMs both align with View 2, making distinctions between modal categories based on degree rather than kind. The threshold-probability view appears to be the best explanation for LM behaviors. Graded representations of probability

[4] We emphasize the difficulty of deriving *literal* meanings. People can of course derive figurative or metaphorical meanings from sentences with apparent selectional restriction violations, and see also Magidor (2013) for arguments that "category error" sentences do have meaning.

[5] See Mankowitz (2023) for arguments that category errors are neither necessary nor sufficient to trigger the N400 effect.

are the currency of LM computation, and it is reasonable to expect LMs to implicitly learn generalized event knowledge as a by-product of the next-word prediction objective (see, e.g., Kauf et al., 2023). However, at first glance View 2 appears inconsistent with the findings of Experiment 2, where people did not distinguish between impossible and inconceivable event descriptions through their perceived likelihood of occurring. We note, however, it could be the case that people judge impossible and inconceivable events as having (near) zero chance of occurring, but they also represent graded notions of which inconceivable events are less likely than others. For example, perhaps if we directly asked people to rank inconceivable event descriptions in order of how likely they are, we would find structured agreement across people, just as prior work has found that people agree on some *impossible* events being more impossible than others (McCoy & Ullman, 2019; Shtulman & Morgan, 2017). Such a finding would suggest that people use their commonsense knowledge to reason about events that are not only impossible, but cannot even be imagined in a possible world (without some process of non-literal interpretation or coercion).

In contrast, the other high-level explanation is that people and language models are using different computations in a convergent fashion. Within this option it is also possible that people make modal distinctions based on graded representations, while LMs rely on distinctions of kind — but this sub-option seems the least likely, given our findings in Experiment 2 (where people's likelihood ratings did not differentiate impossible and inconceivable events) as well as the current understanding of LMs, as discussed above. The other sub-option is that people make their modal distinctions in a manner of kind, consulting "type errors" and causal violations, even though event probabilities (as estimated through the statistics of language use) may in principle be sufficient for this distinction. This option seems the most consistent with our data. On the one hand, LMs are able to leverage the statistics of language use to assign probabilities that distinguish between impossible and inconceivable event descriptions. But on the other hand, people's likelihood judgments do not differ between impossible and inconceivable events, suggesting that people are using a different kind of computation to form this modal distinction. However, the caveats discussed above still hold — for example, it could be that people would agree on the relative rankings within impossible or inconceivable events in a different task setting, and these rankings might correspond well to LMs' event description probabilities. We also have not provided direct evidence for type-based computations in humans, nor direct evidence *against* type-based computations in LMs. It remains an open question how the type-error view could be cleanly separated from the threshold-probability view in an experimental setting.

In addition to comparing fully trained LMs to people's categorization and rating behaviors, we also found interesting patterns in LM training dynamics. Our investigation of modal categories across a single LM's training (summarized in Fig. 4) suggests novel predictions about modal reasoning in children: the ability to distinguish between impossible and inconceivable should be present earlier than the ability to distinguish between improbable and impossible. This prediction could be tested by adapting the categorization experiments that we conducted with adults in Experiment 1, building upon prior categorization experiments in children (Shtulman, 2009; Shtulman & Carey, 2007). We stress that we are *not* claiming direct parallels between model training and child development — these processes substantially differ in many ways (Frank, 2023). Nevertheless, the trajectory of model training suggests high-level interactions between exposure to linguistic data and the emergence of key behavioral patterns (Evanson et al., 2023; Zhang et al., 2021). In particular, the earlier emergence of a distinction between the impossible and the inconceivable could inform ongoing broader debates about the development of modal reasoning: It could bridge between the emergence of basic modal reasoning and minimal representations of possibility (Alderete & Xu, 2023; Leahy & Carey, 2020; Leahy et al., 2022) and the later understanding of the distinction between improbable and the impossible. Further, to the degree that the later understanding rests on causal models or intuitive theories, the earlier emergence of the distinction would support the notion that basic modal possibility judgments may rely on other processes.

Turning now to the AI perspective, our study contributes to a line of work seeking to understand the limitations and capabilities of language models — in particular, what type of semantic, commonsense, or world knowledge LMs might learn through the objective of word prediction (e.g., Chang & Bergen, 2023; Yildirim & Paul, 2024). The closest to our work is that of Kauf et al. (2023), who found that LMs performed poorly at distinguishing between likely and unlikely events, whereas we found evidence that models robustly distinguished between probable and improbable events (Section "Experiment 2: Subjective ratings of event likelihood"). One potential reason for this disparity is that their likely/unlikely distinction involves social world knowledge, whereas our probable/improbable distinction relies on physical phenomena. For example, they tested materials such as "The nanny tutored the boy" (likely) versus "The boy tutored the nanny" (unlikely), or "The actor won the award" (likely) versus "The actor won the battle" (unlikely). It could be the case that the likelihood of events in these minimal pairs can be easily manipulated through context or are more similar *a priori*, where the distinctions in our stimuli are sharper and less context-dependent. For example, there may be more contexts in which actors win *battles* (e.g., against their competitors in a tough audition), but fewer contexts in which someone would chill a drink using *snow*. Other studies have also shown that LMs perform poorly on "counterfactual" tasks which are possible but highly uncommon (conceptually similar to our Improbable condition), such as Python with 1-based indexing (Wu et al., 2024). These results suggest that models' reasoning abilities are highly sensitive to specific patterns seen during pretraining (McCoy et al., 2024), instead of relying on more general task-solving skills. Another related study by Sathe et al. (2024) investigated whether LM conditional probabilities predict human acceptability judgments of novel, unattested collocations such as "educated hairbrush". However, they focused on a specific type of linguistic expression (adjective-noun pairings) and not on modal categories. To our knowledge, our experiments test LMs on fine-grained distinctions across the widest range of modal categories to date, and we demonstrate that LMs are able to reliably distinguish between the likely/unlikely, unlikely/impossible, and impossible/inconceivable. These findings add to a growing body of evidence that structured knowledge about events and physical properties may be learned via statistical learning over linguistic forms.

One potential limitation of our study is the validity of using LM-derived string probabilities to estimate the probability of an event. As discussed in the Introduction (Section "Introduction"), this approach rests on an assumption that the likelihood of using a particular linguistic expression to describe an event is a proxy for how frequently the event occurs. This assumption faces at least two challenges, which we discuss in more detail here. First, let us suppose that language is used only to communicate about events that happen in the world. This view raises the issue of reporting bias (Gordon & Van Durme, 2013; Sorower et al., 2011): low-likelihood events may be over-represented in language, since people are incentivized to talk about them, whereas high-likelihood events may be underrepresented in language, since there is little communicative benefit of talking about them. Reporting bias is most likely to skew models' probability estimates of probable and improbable events, as events that do not occur in the world (i.e., impossible and inconceivable events) will not be described in language at all, if we adopt the view mentioned above. In our study, we do find that LMs assign lower probabilities (higher surprisals) to improbable than probable event descriptions, suggesting that modern LMs can overcome reporting biases to some extent (cf. Shwartz & Choi, 2020).

The second challenge is related to non-literal uses of language. People use language not only to communicate about events literally

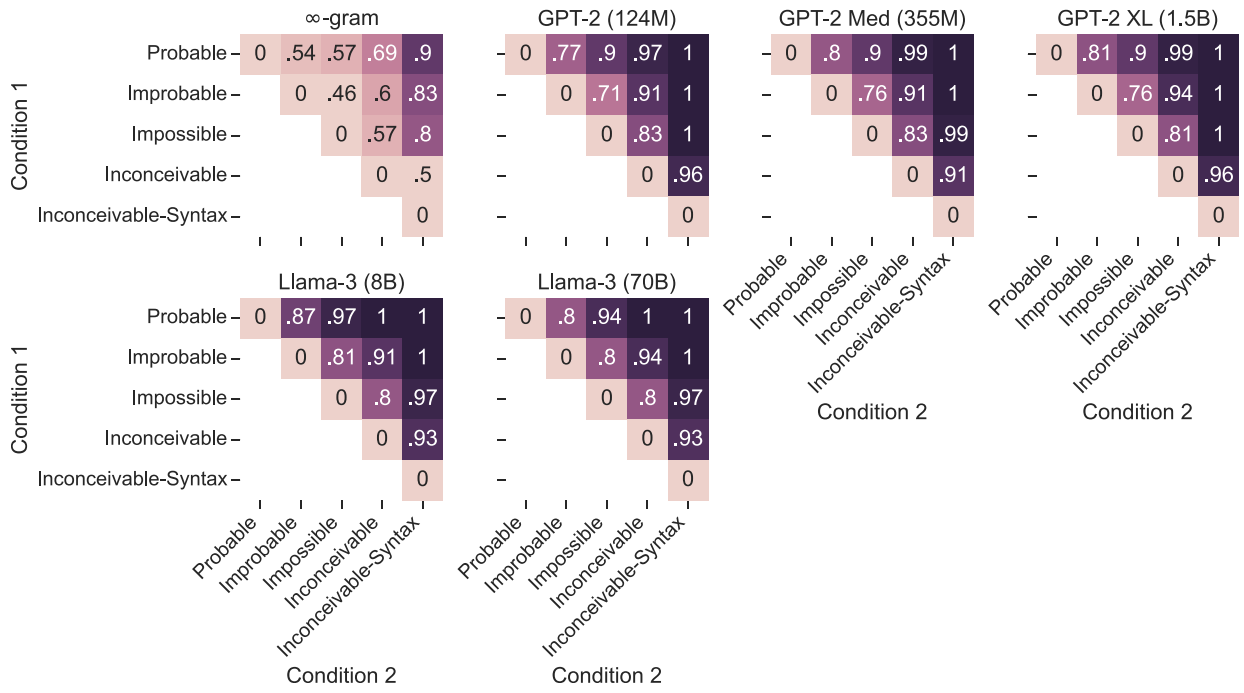

Fig. 6. Item-level surprisal comparisons between pairs of conditions. Each cell (row i, column j) shows the proportion of items where the surprisal of the continuation in condition i < the surprisal of the continuation in condition j.

happening in the world, but also to talk about fictional or hypothetical events, or to express figurative meanings. As a result, the probability an LM assigns to a sentence will be influenced by the non-literal language seen during training, in addition to factors that affect how likely the event is to literally occur. Therefore, there may be cases where a particular string completion makes an event impossible, but is still highly predictable given the context. For example, given "The wizard plucked a scale from his fire-breathing", the completion "dragon" might be extremely likely, even though dragons do not exist. This is not a major concern for our probable, improbable, and impossible stimuli, since the prefixes describe commonplace, physical events, which should set up expectations for physically plausible continuations.

Non-literal language is, however, a potential concern for our inconceivable items, which are only inconceivable when interpreted in a strictly literal sense. For example, "washing your hair with applause" clearly violates physical knowledge, but could potentially be interpreted in a metaphorical manner by re-interpreting "washing" as a non-physical activity. We refer to these re-interpretations of inconceivable events as *coercions*, mirroring the concept of type coercion in computer science. Indeed, a prominent theory of category mistakes argues that these types of sentences *do* have meaning (Magidor, 2013), and Chomsky (1965) writes that "sentences that break selectional rules can often be interpreted metaphorically ... or allusively in one way or another". Another example of non-literal interpretation might apply to certain impossible event descriptions, such as "baking a cake inside a freezer", which could be interpreted as a phrase about cakes that are effectively "baked" inside refrigerators, such as tiramisu. This re-interpretation is potentially less metaphorical than it is about resolving communicative intent. For example, a listener might infer that a speaker chose to use the verb "bake" to convey the broad concept of preparing cakes, even though the literal definition of baking involves ovens or dry heat. While human participants in our experiments were told to consider the sentences in a strictly literal sense, it is not obvious how this restriction could be enforced when measuring string probabilities using language models, especially since models' training datasets are rife with non-literal language.

Another limitation of our study is that we focus on one particular implementation of inconceivability: namely, the encountering of abstract concepts when a concrete noun is expected. But there is a broad range of ways in which events might be inconceivable: for example, by violating other features such as animacy ("the refrigerator giggled"), or by ascribing properties to objects in other odd ways ("the banana is an hour long", "the theory is blue"). It remains an open question whether these violations share some underlying cognitive process, or whether there are distinct avenues for dealing with different kinds of ontological oddities. In addition, we focused on judgments *across* modal categories (probable, improbable, impossible, inconceivable), but it remains an open question how humans reason about items *within* inconceivability. As discussed earlier, people make graded judgments among things that are nominally impossible, such as levitating a feather versus levitating a rock (e.g., McCoy & Ullman, 2019; Shtulman & Morgan, 2017). Similarly, there may also be structured gradedness within people's judgments of the likelihood of inconceivable items. Understanding the intuitive and commonsense knowledge that people use to make judgments about inconceivability is an important direction for future work.

Our focus in this work was on "shades of zero" — distinctions within the region of extremely low probability, beyond the typical experience of people or models. While prior work has focused on distinguishing between the possible and impossible (e.g., Kauf et al., 2023), or the ability of language models to capture people's behavior on naturally occurring sentences (e.g., Shain et al., 2024; Smith & Levy, 2013), we believe that investigating the distinctions between impossibility and inconceivability is an exciting direction from both cognitive and AI perspectives, and can illuminate the boundaries of thought and imagination in both people and machines. In ancient maps, uncharted territories beyond the known world were marked with fantastical creatures that did not exist in the real world. There is much

Table 3
Results of linear mixed-effects regression model predicting surprisal values with fixed-effect of condition and log *n*-gram count, and random intercept of item (Eq. (5)).

Model	Predictor	Estimate	Std error	DF	*t*	*p*	Sig. Code
∞-gram	(Intercept)	5.59685	0.58427	341.56437	9.579	< 2e−16	***
∞-gram	Improbable > Probable	0.38345	0.19752	275.68031	1.941	0.0532	.
∞-gram	Impossible > Improbable	0.10254	0.19723	275.32756	0.520	0.6035	
∞-gram	Inconceivable > Impossible	0.39321	0.19996	278.51520	1.966	0.0502	.
∞-gram	Inconceivable-Syntax > Inconceivable	0.50355	0.25213	316.95485	1.997	0.0467	*
∞-gram	logNgramCount	0.19695	0.03541	343.58475	5.561	5.39e−08	***
GPT-2	(Intercept)	1.85632	0.50714	321.16938	3.660	0.000294	***
GPT-2	Improbable > Probable	0.88423	0.18125	274.68488	4.878	1.81e−06	***
GPT-2	Impossible > Improbable	0.87134	0.18101	274.24529	4.814	2.45e−06	***
GPT-2	Inconceivable > Impossible	1.50987	0.18327	278.22518	8.238	6.90e−15	***
GPT-2	Inconceivable-Syntax > Inconceivable	1.99502	0.22689	325.82304	8.793	< 2e−16	***
GPT-2	logNgramCount	0.22488	0.03084	326.24721	7.291	2.34e−12	***
GPT-2 Med	(Intercept)	1.66233	0.49594	290.80265	3.352	0.000909	***
GPT-2 Med	Improbable > Probable	0.91011	0.18680	275.75602	4.872	1.87e−06	***
GPT-2 Med	Impossible > Improbable	0.98405	0.18658	275.27062	5.274	2.70e−07	***
GPT-2 Med	Inconceivable > Impossible	1.51907	0.18869	279.67165	8.051	2.38e−14	***
GPT-2 Med	Inconceivable-Syntax > Inconceivable	1.48940	0.22972	331.73046	6.484	3.25e−10	***
GPT-2 Med	logNgramCount	0.23270	0.03022	295.85151	7.701	2.04e−13	***
GPT-2 XL	(Intercept)	1.45052	0.49455	344.00000	2.933	0.00358	**
GPT-2 XL	Improbable > Probable	0.86584	0.18872	344.00000	4.588	6.28e−06	***
GPT-2 XL	Impossible > Improbable	1.23210	0.18849	344.00000	6.537	2.27e−10	***
GPT-2 XL	Inconceivable > Impossible	1.50969	0.19057	344.00000	7.922	3.27e−14	***
GPT-2 XL	Inconceivable-Syntax > Inconceivable	2.02452	0.23108	344.00000	8.761	< 2e−16	***
GPT-2 XL	logNgramCount	0.24662	0.03014	344.00000	8.182	5.46e−15	***
Llama-3 8B	(Intercept)	1.52189	0.51228	305.01804	2.971	0.00321	**
Llama-3 8B	Improbable > Probable	1.18717	0.18857	275.69218	6.295	1.20e−09	***
Llama-3 8B	Impossible > Improbable	1.30949	0.18833	275.22579	6.953	2.60e−11	***
Llama-3 8B	Inconceivable > Impossible	1.15916	0.19056	279.45135	6.083	3.87e−09	***
Llama-3 8B	Inconceivable-Syntax > Inconceivable	1.73243	0.23368	329.59096	7.414	1.05e−12	***
Llama-3 8B	logNgramCount	0.23494	0.03119	310.23786	7.533	5.48e−13	***
Llama-3 70B	(Intercept)	2.07629	0.53851	300.73340	3.856	0.000141	***
Llama-3 70B	Improbable > Probable	1.16023	0.19970	276.02326	5.810	1.72e−08	***
Llama-3 70B	Impossible > Improbable	1.40419	0.19945	275.55187	7.040	1.53e−11	***
Llama-3 70B	Inconceivable > Impossible	1.29915	0.20178	279.82344	6.439	5.23e−10	***
Llama-3 70B	Inconceivable-Syntax > Inconceivable	1.94857	0.24685	330.40583	7.894	4.37e−14	***
Llama-3 70B	logNgramCount	0.20858	0.03279	305.89998	6.360	7.32e−10	***

to explore in regions of extremely low probability. Here, there may literally be dragons.

CRediT authorship contribution statement

Jennifer Hu: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Felix Sosa:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Tomer Ullman:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been made possible in part by a gift from the Chan Zuckerberg Initiative Foundation to establish the Kempner Institute for the Study of Natural and Artificial Intelligence. The experiments were supported by a grant from the Harvard University Hodgson Memorial Fund. We also acknowledge the support of the Jacobs Foundation (TDU).

Appendix. Additional analyses for Experiment 3

A.1. Item-level comparisons

Fig. 6 shows pairwise comparisons of surprisals across conditions within items. Each cell shows the proportion of items where the surprisal of the continuation in condition 1 (row) is less than the surprisal of the continuation in condition 2 (column).

A.2. Full regression outputs

Table 3 shows full results from the linear mixed-effects regression model fit for each language model in Experiment 2 (Section "Experiment 2: Subjective ratings of event likelihood"; Eq. (5)).

Data availability

Our code, experimental stimuli, and data are publicly available at https://github.com/jennhu/shades-of-zero.

References

Abdou, M., Kulmizev, A., Hershcovich, D., Frank, S., Pavlick, E., & Søgaard, A. (2021). Can language models encode perceptual structure without grounding? A case study in color. In *Proceedings of the 25th conference on computational natural language learning* (pp. 109–132). Association for Computational Linguistics, Online, URL https://aclanthology.org/2021.conll-1.9.

AI@Meta (2024). Llama 3 model card. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Alderete, S., & Xu, F. (2023). Three-year-old children's reasoning about possibilities. *Cognition*, *237*, Article 105472.

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264. http://dx.doi.org/10.1016/S0010-0277(99)00059-1, ISSN: 1873-7838(Electronic), 0010-0277(Print).

Balog, K. (1999). Conceivability, possibility, and the mind-body problem. *The Philosophical Review*, [ISSN: 00318108] *108*(4), 497–528. http://dx.doi.org/10.2307/2998286, URL http://www.jstor.org/stable/2998286.

Bekki, D. (2014). Representing anaphora with dependent types. In N. Asher, & S. Soloviev (Eds.), *Logical aspects of computational linguistics* (pp. 14–29). Berlin, Heidelberg: Springer Berlin Heidelberg, ISBN: 978-3-662-43742-1.

Bicknell, K., Elman, J. L., Hare, M., McRae, K., & Kutas, M. (2010). Effects of event knowledge in processing verbal arguments. *Journal of Memory and Language*, *63*(4), 489–505. http://dx.doi.org/10.1016/j.jml.2010.08.004, ISSN: 1096-0821(Electronic), 0749-596X(Print).

Bornkessel, I., & Schlesewsky, M. (2006). The extended argument dependency model: A neurocognitive approach to sentence comprehension across languages. *Psychological Review*, [ISSN: 0033-295X] *113*(4), 787–821. http://dx.doi.org/10.1037/0033-295X.113.4.787.

Brothers, T., & Kuperberg, G. R. (2021). Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of Memory and Language*, [ISSN: 0749-596X] *116*, Article 104174, URL https://www.sciencedirect.com/science/article/pii/S0749596X20300887.

Browne, C. A., & Woolley, J. D. (2004). Preschoolers' magical explanations for violations of physical, social, and mental laws. *Journal of Cognition and Development*, [ISSN: 1524-8372] *5*(2), 239–260. http://dx.doi.org/10.1207/s15327647jcd0502_4.

Byrne, R. M. (2007). *The rational imagination: how people create alternatives to reality*. MIT Press.

Chalmers, D. J. (1996). *The conscious mind*. New York: Oxford University Press.

Chang, T. A., & Bergen, B. K. (2023). Language model behavior: A comprehensive survey. *Computational Linguistics*, [ISSN: 0891-2017] 1–55.

Chatzikyriakidis, S., & Luo, Z. (2017). Adjectival and adverbial modification: The view from modern type theories. *Journal of Logic, Language and Information*, [ISSN: 1572-9583] *26*(1), 45–88. http://dx.doi.org/10.1007/s10849-017-9246-2.

Chatzikyriakidis, S., & Luo, Z. (2020). Formal semantics in modern type theories. In *Cognitive science series: logic, linguistics, and computer science: vol. 2*, Wiley.

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

Davies, M. (2008). The corpus of contemporary American english (COCA). URL https://www.english-corpora.org/coca/.

Evanson, L., Lakretz, Y., & King, J. R. (2023). Language acquisition: do children and language models follow similar learning stages? In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: ACL 2023* (pp. 12205–12218). Toronto, Canada: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.findings-acl.773, URL https://aclanthology.org/2023.findings-acl.773.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*, 3–71. http://dx.doi.org/10.1016/0010-0277(88)90031-5, ISSN: 1873-7838(Electronic), 0010-0277(Print).

Frank, M. C. (2023). Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*, [ISSN: 1364-6613] *27*(11), 990–992. http://dx.doi.org/10.1016/j.tics.2023.08.007, URL https://www.sciencedirect.com/science/article/pii/S1364661323002036.

Garnsey, S. M., Pearlmutter, N. J., Myers, E., & Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language*, [ISSN: 0749-596X] *37*(1), 58–93. http://dx.doi.org/10.1006/jmla.1997.2512, URL https://www.sciencedirect.com/science/article/pii/S0749596X97925122.

Gendler, T., & Hawthorne, J. (Eds.), (2002). *Conceivability and possibility*. New York: Oxford University Press.

Gordon, J., & Van Durme, B. (2013). Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on automated knowledge base construction* (pp. 25–30). New York, NY, USA: Association for Computing Machinery, ISBN: 978-1-4503-2411-3.

Goulding, B. W., & Friedman, O. (2021). A similarity heuristic in children's possibility judgments. *Child Development*, [ISSN: 1467-8624] *92*(2), 662–671. http://dx.doi.org/10.1111/cdev.13534.

Goulding, B. W., & Friedman, O. (2023). Perceived similarity explains beliefs about possibility. *Journal of Experimental Psychology: General*.

Goulding, B. W., Khan, F., Fukuda, K., Lane, J. D., & Ronfard, S. (2023). The development of modal intuitions: A test of two accounts. *Journal of Experimental Psychology: General*, ISSN: 1939-2222(Electronic), 0096-3445(Print).

Griffiths, T. L. (2015). Revealing ontological commitments by magic. *Cognition*, [ISSN: 0010-0277] *136*, 43–48.

Groeneveld, D., Beltagy, I., Walsh, E., Bhagia, A., Kinney, R., Tafjord, O., Jha, A., Ivison, H., Magnusson, I., Wang, Y., Arora, S., Atkinson, D., Authur, R., Chandu, K., Cohan, A., Dumas, J., Elazar, Y., Gu, Y., Hessel, J., ..., Khot, T. et al. (2024). OLMo: Accelerating the science of language models. In L.-W. Ku, A. Martins, V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 15789–15809). Bangkok, Thailand: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.acl-long.841, URL https://aclanthology.org/2024.acl-long.841.

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science*, *304*(5669), 438–441. http://dx.doi.org/10.1126/science.1095455.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north American chapter of the association for computational linguistics*.

Harris, P. L. (2021). Early constraints on the imagination: The realism of young children. *Child Development*, *92*(2), 466–483.

Hofmann, M. J., Remus, S., Biemann, C., Radach, R., & Kuchinke, L. (2022). Language models explain word reading times better than empirical predictability. *Frontiers in Artificial Intelligence*, [ISSN: 2624-8212] *4*, URL https://www.frontiersin.org/articles/10.3389/frai.2021.730570.

Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. (2020). A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 1725–1744). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.acl-main.158, Online, URL https://aclanthology.org/2020.acl-main.158.

Juola, P. (2022). Google books ngrams. In L. A. Schintler, & C. L. McNeely (Eds.), *Encyclopedia of big data* (pp. 517–521). Cham: Springer International Publishing, ISBN: 978-3-319-32010-6, http://dx.doi.org/10.1007/978-3-319-32010-6_359.

Kahneman, D., & Tversky, A. (1981). *The simulation heuristic*. National Technical Information Service.

Kamide, Y., Altmann, G. T., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, [ISSN: 0749-596X] *49*(1), 133–156. http://dx.doi.org/10.1016/S0749-596X(03)00023-8, URL https://www.sciencedirect.com/science/article/pii/S0749596X03000238.

Katz, J. J., & Fodor, J. A. (1963). The structure of a semantic theory. *Language*, *39*(2), 170–210.

Kauf, C., Chersoni, E., Lenci, A., Fedorenko, E., & Ivanova, A. A. (2024). Comparing plausibility estimates in base and instruction-tuned large language models. URL https://arxiv.org/abs/2403.14859.

Kauf, C., Ivanova, A. A., Rambelli, G., Chersoni, E., She, J. S., Chowdhury, Z., Fedorenko, E., & Lenci, A. (2023). Event knowledge in large language models: The gap between the impossible and the unlikely. *Cognitive Science*, *47*(11), Article e13386.

Komatsu, L. K., & Galotti, K. M. (1986). Children's reasoning about social, physical, and logical regularities: A look at two worlds. *Child Development*, [ISSN: 00093920] *57*(2), 413–420. http://dx.doi.org/10.2307/1130597, URL http://www.jstor.org/stable/1130597.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, [ISSN: 1545-2085] *62*(Volume 62, 2011), 621–647. http://dx.doi.org/10.1146/annurev.psych.093008.131123, URL https://www.annualreviews.org/content/journals/10.1146/annurev.psych.093008.131123.

Lane, J. D., Ronfard, S., Francioli, S. P., & Harris, P. L. (2016). Children's imagination and belief: Prone to flights of fancy or grounded in reality? *Cognition*, [ISSN: 1873-7838] *152*, 127–140. http://dx.doi.org/10.1016/j.cognition.2016.03.022.

Leahy, B. P., & Carey, S. E. (2020). The acquisition of modal concepts. *Trends in Cognitive Sciences*, *24*(1), 65–78.

Leahy, B., Huemer, M., Steele, M., Alderete, S., & Carey, S. (2022). Minimal representations of possibility at age 3. *Proceedings of the National Academy of Sciences*, *119*(52), Article e2207499119.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, [ISSN: 0010-0277] *106*(3), 1126–1177.

Lewis, D. K. (1986). *On the plurality of worlds*. Oxford: Blackwell.

Lewry, C., Curtis, K., Vasilyeva, N., Xu, F., & Griffiths, T. L. (2021). Intuitions about magic track the development of intuitive physics. *Cognition*, *214*.

Li, X. L., Kuncoro, A., Hoffmann, J., de Masson d'Autume, C., Blunsom, P., & Nematzadeh, A. (2022). A systematic investigation of commonsense knowledge in large language models. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 11838–11855). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.emnlp-main.812, URL https://aclanthology.org/2022.emnlp-main.812.

Liu, J., Min, S., Zettlemoyer, L., Choi, Y., & Hajishirzi, H. (2024). Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. In *First conference on language modeling*. URL https://openreview.net/forum?id=u2vAyMeLMm.

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, *3*(2), 273–302. http://dx.doi.org/10.1111/j.1756-8765.2010.01106.x, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1756-8765.2010.01106.x.

Louwerse, M. M. (2018). Knowing the meaning of a word by the linguistic and perceptual company it keeps. *Topics in Cognitive Science*, *10*(3), 573–589. http://dx.doi.org/10.1111/tops.12349, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/tops.12349.

Luo, Z. (2012). Formal semantics in modern type theories with coercive subtyping. *Linguistics and Philosophy*, [ISSN: 1573-0549] *35*(6), 491–513. http://dx.doi.org/10.1007/s10988-013-9126-4.

Magidor, O. (2009). Category mistakes are meaningful. *Linguistics and Philosophy*, [ISSN: 1573-0549] *32*(6), 553–581.

Magidor, O. (2013). *Category mistakes*. Oxford University Press, ISBN: 978-0-19-957297-7, http://dx.doi.org/10.1093/acprof:oso/9780199572977.001.0001.

Mankowitz, P. (2023). Category mistakes electrified. *Review of Philosophy and Psychology*, [ISSN: 1878-5166] http://dx.doi.org/10.1007/s13164-023-00684-y.

Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 1192–1202). Brussels, Belgium: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D18-1151, URL https://aclanthology.org/D18-1151.

Matsuki, K., Chow, T., Hare, M., Elman, J. L., Scheepers, C., & McRae, K. (2011). Event-based plausibility immediately influences on-line language comprehension. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, [ISSN: 1939-1285] *37*(4), 913–934. http://dx.doi.org/10.1037/a0022964.

McCoy, J., & Ullman, T. (2019). Judgments of effort for magical violations of intuitive physics. *PLOS ONE*, *14*(5).

McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D., & Griffiths, T. L. (2024). Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, *121*(41), Article e2322420121. http://dx.doi.org/10.1073/pnas.2322420121.

McRae, K., Ferretti, & Liane Amyote, T. R. (1997). Thematic roles as verb-specific concepts. *Language and Cognitive Processes*, [ISSN: 0169-0965] *12*(2–3), 137–176. http://dx.doi.org/10.1080/016909697386835.

McRae, K., & Matsuki, K. (2009). People use their knowledge of common events to understand language, and do so as quickly as possible. *Language and Linguistics Compass*, [ISSN: 1749-818X] *3*(6), 1417–1429. http://dx.doi.org/10.1111/j.1749-818X.2009.00174.x.

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, [ISSN: 0749-596X] *38*(3), 283–312. http://dx.doi.org/10.1006/jmla.1997.2543, URL https://www.sciencedirect.com/science/article/pii/S0749596X97925432.

Merullo, J., Castricato, L., Eickhoff, C., & Pavlick, E. (2023). Linearly mapping from image to text space. In *The eleventh international conference on learning representations*. URL https://openreview.net/forum?id=8tYRqb05pVn.

Montague, R. (1973). The proper treatment of quantification in ordinary english. In K. J. J. Hintikka, J. M. E. Moravcsik, & P. Suppes (Eds.), *Approaches to natural language: proceedings of the 1970 stanford workshop on grammar and semantics* (pp. 221–242). Dordrecht: Springer Netherlands, ISBN: 978-94-010-2506-5, http://dx.doi.org/10.1007/978-94-010-2506-5_10.

Morales, L. E. (2018). On the representation and learning of concepts: programs, types, and Bayes (Ph.D. thesis), Massachusetts Institute of Technology.

Myers, E. B., & Blumstein, S. E. (2005). Selectional restriction and semantic priming effects in normals and Broca's aphasics. *Journal of Neurolinguistics*, [ISSN: 0911-6044] *18*(3), 277–296. http://dx.doi.org/10.1016/j.jneuroling.2004.05.001, URL https://www.sciencedirect.com/science/article/pii/S0911604405000084.

Paczynski, M., & Kuperberg, G. R. (2012). Multiple influences of semantic memory on sentence processing: Distinct effects of semantic relatedness on violations of real-world event/state knowledge and animacy selection restrictions. *Journal of Memory and Language*, [ISSN: 0749-596X] *67*(4), 426–448. http://dx.doi.org/10.1016/j.jml.2012.07.003, URL https://www.sciencedirect.com/science/article/pii/S0749596X12000745.

Patel, R., & Pavlick, E. (2022). Mapping language models to grounded conceptual spaces. In *International conference on learning representations*. URL https://openreview.net/forum?id=gJcEM8sxHK.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, [ISSN: 0010-0277] *28*(1), 73–193. http://dx.doi.org/10.1016/0010-0277(88)90032-7, URL https://www.sciencedirect.com/science/article/pii/0010027788900327.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. URL https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

Ramsey, W. (1997). Do connectionist representations earn their explanatory keep? *Mind & Language*, [ISSN: 0268-1064] *12*(1), 34–66. http://dx.doi.org/10.1111/j.1468-0017.1997.tb00061.x.

Rayner, K., Warren, T., Juhasz, B. J., & Liversedge, S. P. (2004). The effect of plausibility on eye movements in reading. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *30*(6), 1290–1301. http://dx.doi.org/10.1037/0278-7393.30.6.1290, ISSN: 1939-1285(Electronic), 0278-7393(Print).

Resnik, P. (1993). Semantic classes and syntactic ambiguity. In *Human language technology: proceedings of a workshop held at plainsboro, New Jersey, March 21-24, 1993*. URL https://aclanthology.org/H93-1054.

Resnik, P. (1996). Selectional constraints: an information-theoretic model and its computational realization. *Cognition*, *61*, 127–159.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. *vol. 2*, In *Parallel distributed processing: explorations in the microstructures of cognition* (pp. 216–271). MIT Press, URL https://web.stanford.edu/~jlmcc/papers/PDP/Chapter18.pdf.

Ryle, G. (1949). *The concept of mind*. University of Chicago Press.

Sathe, A., Fedorenko, E., & Zaslavsky, N. (2024). Language use is only sparsely compositional: The case of English adjective-noun phrases in humans and large language models. *Vol. 46*, In *Proceedings of the annual meeting of the cognitive science society*. URL https://escholarship.org/uc/item/0qd3662b.

Shain, C., Meister, C., Pimentel, T., Cotterell, R., & Levy, R. (2024). Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, *121*(10), Article e2307876121. http://dx.doi.org/10.1073/pnas.2307876121, URL https://doi.org/10.1073/pnas.2307876121.

Shtulman, A. (2009). The development of possibility judgment within and across domains. *Cognitive Development*, *24*, 293–309. http://dx.doi.org/10.1016/j.cogdev.2008.12.006.

Shtulman, A., & Carey, S. (2007). Improbable or impossible? How children reason about the possibility of extraordinary events. *Child Development*, *78*(3), 1015–1032.

Shtulman, A., & Morgan, C. (2017). The explanatory structure of unexplainable events: Causal constraints on magical reasoning. *Psychonomic Bulletin & Review*, [ISSN: 1531-5320] *24*(5), 1573–1585.

Shtulman, A., & Phillips, J. (2018). Differentiating "could" from "should": Developmental changes in modal cognition. *The Early Development of the Normative Mind*, [ISSN: 0022-0965] *165*, 161–182. http://dx.doi.org/10.1016/j.jecp.2017.05.012, URL https://www.sciencedirect.com/science/article/pii/S0022096517303429.

Shwartz, V., & Choi, Y. (2020). Do neural language models overcome reporting bias? In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th international conference on computational linguistics* (pp. 6863–6870). International Committee on Computational Linguistics.

Sitnikova, T., Holcomb, P. J., Kiyonaga, K. A., & Kuperberg, G. R. (2008). Two neurocognitive mechanisms of semantic integration during the comprehension of visual real-world events. *Journal of Cognitive Neuroscience*, [ISSN: 0898-929X] *20*(11), 2037–2057. http://dx.doi.org/10.1162/jocn.2008.20143.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, [ISSN: 0010-0277] *128*(3), 302–319, URL http://www.sciencedirect.com/science/article/pii/S0010027713000413.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, [ISSN: 0140-525X] *11*(1), 1–23. http://dx.doi.org/10.1017/S0140525X00052432, URL https://www.cambridge.org/core/product/4B8871A82A932DB96D183AAC9C0CF037.

Soldaini, L., Kinney, R., Bhagia, A., Schwenk, D., Atkinson, D., Authur, R., Bogin, B., Chandu, K., Dumas, J., Elazar, Y., Hofmann, V., Jha, A., Kumar, S., Lucy, L., Lyu, X., Lambert, N., Magnusson, I., Morrison, J., Muennighoff, N., ..., Naik, A. et al. (2024). Dolma: an open corpus of three trillion tokens for language model pretraining research. In L.-W. Ku, A. Martins, V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 15725–15788). Bangkok, Thailand: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.acl-long.840, URL https://aclanthology.org/2024.acl-long.840.

Sorower, M., Doppa, J., Orr, W., Tadepalli, P., Dietterich, T., & Fern, X. (2011). Inverting grice's maxims to learn rules from natural language extractions. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.), *Vol. 24*, *Advances in neural information processing systems*. Curran Associates, Inc., URL https://proceedings.neurips.cc/paper/2011/hash/8c7bbbba95c1025975e548cee86dfadc-Abstract.html.

Sosa, F. A., & Ullman, T. (2022). Type theory in human-like learning and inference. In *Beyond Bayes: paths towards universal reasoning systems workshop at the international conference on machine learning*. URL https://arxiv.org/abs/2210.01634.

Sutton, P. R. (2024). Types and type theories in natural language analysis. *Annual Review of Linguistics*, [ISSN: 2333-9691] *10*(Volume 10, 2024), 107–126. http://dx.doi.org/10.1146/annurev-linguistics-031422-113929, URL https://www.annualreviews.org/content/journals/10.1146/annurev-linguistics-031422-113929.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., ..., Fuller, B. et al. (2023). Llama 2: Open foundation and fine-tuned chat models. URL https://arxiv.org/abs/2307.09288.

Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from ERPs and reading times. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, [ISSN: 0278-7393] *31*(3), 443–467. http://dx.doi.org/10.1037/0278-7393.31.3.443.

van Gompel, R. P., Pickering, M. J., Pearson, J., & Liversedge, S. P. (2005). Evidence against competition during syntactic ambiguity resolution. *Journal of Memory and Language*, [ISSN: 0749-596X] *52*(2), 284–307. http://dx.doi.org/10.1016/j.jml.2004.11.003, URL https://www.sciencedirect.com/science/article/pii/S0749596X04001238.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Vol. 30*, *Advances in neural information processing systems*. Curran Associates, Inc., URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wang, S., Durrett, G., & Erk, K. (2018). Modeling semantic plausibility by injecting world knowledge. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics:*

human language technologies, volume 2 (short papers) (pp. 303–308). New Orleans, Louisiana: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N18-2049, URL https://aclanthology.org/N18-2049.

Warren, T., & McConnell, K. (2007). Investigating effects of selectional restriction violations and plausibility violation severity on eye-movements in reading. *Psychonomic Bulletin & Review*, [ISSN: 1069-9384] *14*(4), 770–775. http://dx.doi.org/10.3758/bf03196835.

Warren, T., Milburn, E., Patson, N. D., & Dickey, M. W. (2015). Comprehending the impossible: What role do selectional restriction violations play? *Language, Cognition and Neuroscience*, [ISSN: 2327-3798] *30*(8), 932–939. http://dx.doi.org/10.1080/23273798.2015.1047458.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2020). BLiMP: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, *8*, http://dx.doi.org/10.1162/tacl_a_00321.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., ..., Drame, M. et al. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 38–45). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.emnlp-demos.6, Online, URL https://aclanthology.org/2020.emnlp-demos.6.

Wu, Z., Qiu, L., Ross, A., Akyürek, E., Chen, B., Wang, B., Kim, N., Andreas, J., & Kim, Y. (2024). Reasoning or reciting? Exploring the capabilities and limitations of language models through counterfactual tasks. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 conference of the north American chapter of the association for computational linguistics: human language technologies (volume 1: long papers)* (pp. 1819–1862). Mexico City, Mexico: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.naacl-long.102, URL https://aclanthology.org/2024.naacl-long.102.

Yildirim, I., & Paul, L. (2024). From task structures to world models: What do LLMs know? *Trends in Cognitive Sciences*, [ISSN: 1364-6613] http://dx.doi.org/10.1016/j.tics.2024.02.008.

Zhang, Y., Warstadt, A., Li, X., & Bowman, S. R. (2021). When do you need billions of words of pretraining data? In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)* (pp. 1112–1125). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.acl-long.90, Online, URL https://aclanthology.org/2021.acl-long.90.

Zhou, X., Zhang, Y., Cui, L., & Huang, D. (2020). Evaluating commonsense in pre-trained language models. *Vol. 34*, In *Proceedings of the AAAI Conference on Artificial Intelligence* (05), (pp. 9733–9740). http://dx.doi.org/10.1609/aaai.v34i05.6523, URL https://ojs.aaai.org/index.php/AAAI/article/view/6523.