COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 47 (2023) e13386

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13386

# Event Knowledge in Large Language Models: The Gap Between the Impossible and the Unlikely

Carina Kauf,[a,b,#] Anna A. Ivanova,[a,b,c,#] Giulia Rambelli,[d] Emmanuele Chersoni,[e] Jingyuan Selena She,[a,b] Zawad Chowdhury,[f] Evelina Fedorenko,[a,b] Alessandro Lenci[g]

[a]*Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology*
[b]*McGovern Institute for Brain Research, Massachusetts Institute of Technology*
[c]*Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology*
[d]*Department of Modern Languages, Literatures and Cultures, University of Bologna*
[e]*Department of Chinese and Bilingual Studies, Hong Kong Polytechnic University*
[f]*Department of Mathematics, University of Washington*
[g]*Department of Philology, Literature, and Linguistics, University of Pisa*

Received 7 December 2022; received in revised form 27 October 2023; accepted 4 November 2023

**Abstract**

Word co-occurrence patterns in language corpora contain a surprising amount of conceptual knowledge. Large language models (LLMs), trained to predict words in context, leverage these patterns to achieve impressive performance on diverse semantic tasks requiring world knowledge. An important but understudied question about LLMs' semantic abilities is whether they acquire generalized knowledge of common events. Here, we test whether five pretrained LLMs (from 2018's BERT to 2023's MPT) assign a higher likelihood to plausible descriptions of agent−patient interactions than to minimally different implausible versions of the same event. Using three curated sets of minimal sentence pairs (total $n = 1215$), we found that pretrained LLMs possess substantial event knowledge, outperforming other distributional language models. In particular, they almost always assign a higher likelihood to possible versus impossible events (*The teacher bought the laptop* vs. *The laptop bought the teacher*). However, LLMs show less consistent preferences for likely versus unlikely events (*The nanny tutored the boy* vs. *The boy tutored the nanny*). In follow-up analyses, we show that (i) LLM scores are

[#]The two lead authors contributed equally to this work.
Correspondence should be sent to Carina Kauf, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 43 Vassar St., Cambridge, MA 02139, USA. E-mail: ckauf@mit.edu

driven by both plausibility and surface-level sentence features, (ii) LLM scores generalize well across syntactic variants (active vs. passive constructions) but less well across semantic variants (synonymous sentences), (iii) some LLM errors mirror human judgment ambiguity, and (iv) sentence plausibility serves as an organizing dimension in internal LLM representations. Overall, our results show that important aspects of event knowledge naturally emerge from distributional linguistic patterns, but also highlight a gap between representations of possible/impossible and likely/unlikely events.

*Keywords:* Generalized event knowledge; World knowledge; Plausibility; Typicality; Artificial neural networks; Language models; Syntax; Semantics

## 1. Introduction

### *1.1. Language and event knowledge*

A vital component of human intelligence is our ability to learn, store, and flexibly use rich, structured knowledge about the world. World knowledge spans different domains (from physical properties to social conventions) and covers different types of information, including knowledge of objects, agents, actions, and ideas. One important component of world knowledge is our *generalized event knowledge (GEK)*—templates of common events observed in the world (e.g., McRae & Matsuki, 2009). Humans acquire GEK both through sensorimotor experiences (i.e., from participating in and observing events in the world) and through linguistic experiences (i.e., from event descriptions generated by other people) (Dove, 2020; Dove, 2023; Günther et al., 2020). Here, we ask: To which extent can GEK be learned simply by tracking distributional properties of linguistic input?

On the one hand, positing that GEK can be learned from language alone appears to contradict the fact that in humans, much of conceptual knowledge is innate (e.g., Spelke & Kinzler, 2007) or learned through direct experience (Meteyard & Vigliocco, 2008). On the other hand, co-occurrence patterns learned from language exhibit a remarkable degree of correspondence with distributional spaces learned through other modalities, like vision (Abdou et al., 2021; Lewis, Zettersten, & Lupyan, 2019; Patel & Pavlick, 2021; Roads & Love, 2020; Sorscher et al., 2021). This alignment suggests that language-based distributional information might be able to replace other modalities as a source of world knowledge (Louwerse, 2011). Indeed, knowledge of events is abundantly represented in language corpora, presumably because humans typically communicate events that are, were, or will be happening in the world (e.g., McRae & Matsuki, 2009). Consequently, the GEK that can be learned from distributional linguistic knowledge might faithfully reflect the GEK that people typically acquire multimodally.

### *1.2. LLMs as models of semantic knowledge*

To disentangle contributions of distributional linguistic knowledge from other sources of information—a feat that is difficult to accomplish in humans (e.g., Kim, Elli, & Bedny, 2019; Lewis et al., 2019; Ostarek, Van Paridon, & Montero-Melis, 2019)—we turn to large language models (LLMs). LLMs are the latest generation of distributional semantic models

(Lenci & Sahlgren, 2023), which learn rich semantic representations through tracking word co-occurrence patterns in text in service of their training objective, that is, predicting the next/a missing word from a given linguistic context. A wealth of research has demonstrated that distributional semantic models can explain a broad range of phenomena in human cognition, for example, synonym judgment (Landauer & Dumais, 1997; Levy, Bullinaria, & McCormick, 2017), similarity judgments (Hill, Reichart, & Korhonen, 2015), semantic priming effects in word naming and lexical decision tasks (Mandera, Keuleers, & Brysbaert, 2017). This makes them a useful tool for probing language representations in the human mind and for understanding what kind of information can be learned from text alone.

We focus on LLMs that have been trained on large, general text corpora with a word-in-context prediction objective, often referred to as "pretrained" LLMs. The word-prediction objective enables these models to learn rich amounts of knowledge without being constrained by specific task demands; moreover, this objective closely parallels the next-word-prediction behavior observed in humans (e.g., Altmann & Kamide, 1999; Kuperberg & Jaeger, 2016; Kutas & Federmeier, 2011; Levy, 2008; Mani & Huettig, 2012; McRae, Spivey-Knowlton, & Tanenhaus, 1998; Shain, Blank, van Schijndel, Schuler, & Fedorenko, 2020; Smith & Levy, 2013; Traxler, Morris, & Seely, 2002), making it a cognitively plausible training function for distributional language models (e.g., Goldstein et al., 2022; Hosseini et al., 2022; Schrimpf et al., 2021). Due to the focus on models that capture the task-agnostic distributional language spaces, fine-tuned LLMs are beyond the scope of this paper.

LLMs today generate grammatically correct, syntactically varied, and semantically relevant texts, indicating that these models have essentially mastered *formal* linguistic competence, that is, knowledge of the rules and patterns that govern natural language (Contreras Kallens, Kristensen-McLachlan, & Christiansen, 2023; Mahowald et al., 2023; see also Piantadosi, 2023). However, their *functional* linguistic competence, that is, their general knowledge and reasoning skills as expressed through language, remain highly debated (e.g., Bender & Koller, 2020; Mahowald, Diachek, Gibson, Fedorenko, & Futrell, 2023; Marcus, 2020). Despite their seemingly remarkable success across a variety of tasks, such as generating syntactically and semantically coherent paragraphs of text (Brown et al., 2020), sentiment analysis and logical inference (e.g., Devlin et al., 2018; Liu et al., 2019; Radford et al., 2019; Yang et al., 2019), closed-book question answering (QA) (Roberts, Raffel, & Shazeer, 2020), theory of mind (Kosinski, 2023; Shapira et al., 2023; Trott, Jones, Chang, Michaelov, & Bergen, 2023), and certain aspects of commonsense reasoning (e.g., Zellers et al., 2018), a closer examination of LLM performance reveals that they frequently rely on low-level word-co-occurrence patterns, which, when removed, drastically affect LLM performance (e.g., She et al., 2023; Ullman, 2023). This performance pattern stands in contrast to human performance, which is typically robust to such low-level variations (although see Dasgupta et al., 2022; Lampinen, 2022 for calls to not overestimate human performance).

Studies of world knowledge in LLMs have likewise produced mixed results. On the one hand, even non-fine-tuned LLMs perform well on multiple tasks designed to probe world knowledge, such as the Winograd Schema Challenge (WSC; Levesque, Davis, & Morgenstern, 2012), the Story Cloze Test (SWAG; Zellers et al., 2018), and the Choice of Plausible Alternatives Test (COPA; Roemmele, Bejan, & Gordon, 2011), so much so that some authors

have proposed and evaluated their use as off-the-shelf knowledge base models (Kassner, Dufter, & Schütze, 2021; Petroni et al., 2019; Roberts et al., 2020; Tamborrino, Pellicanò, Pannier, Voitot, & Naudin, 2020). On the other hand, studies using more fine-grained tests have shown that world knowledge in contemporary LLMs is often brittle and depends strongly on the specific way the problem is stated (Elazar et al., 2021a; 2021b; Ettinger, 2020; Kassner & Schütze, 2020; McCoy, Pavlick, & Linzen, 2019; Niven & Kao, 2019; Pedinotti et al., 2021; Ravichander, Hovy, Suleman, Trischler, & Cheung, 2020; Ribeiro, Wu, Guestrin, & Singh, 2020). For example, some authors have noted that, when low-level co-occurrence statistics are properly controlled for, LLMs that were considered to have high accuracy on world knowledge tasks start to perform randomly (Elazar, Zhang, Goldberg, & Roth, 2021b; Sakaguchi, Bras, Bhagavatula, & Choi, 2021), highlighting the potential discrepancy between the word-in-context prediction objective (which benefits from tracking surface-level statistics) and world knowledge acquisition (which should be invariant to surface-level statistics).

### *1.3. LLMs as models of GEK*

In principle, LLMs should be well-posed to acquire GEK. First, significant subparts of GEK are readily available in language co-occurrence statistics. This is evidenced by the success of relatively small distributional semantic models, such as distributional selectional preference models (Erk, 2007; Padó, Crocker, & Keller, 2006; Padó, Padó, & Erk, 2007), or a more recent Structured Distributional Model (Chersoni et al., 2019), which explicitly represents GEK as a distributional event graph (DEG) of syntagmatic relations extracted from dependency-parsed corpora (see also Santus, Chersoni, Lenci, & Blache, 2017; Sayeed, Shkadzko, & Demberg, 2015) on thematic fit modeling tasks (Vassallo et al., 2018). Furthermore, Elman and McRae (2019) show that a small recurrent network trained with a string prediction objective is able to extract GEK about certain events from a small set of curated training set sentences.

Second, the increased scale of LLMs in comparison to earlier generations of distributional semantic models of GEK (such as SDM)—for example, in terms of their numbers of parameters, training data size, or context window size—should be conducive for learning much richer patterns of event knowledge than traditional distributional methods (e.g., Erk, 2012). They should, therefore, be able to generalize their representations more easily to unseen event descriptions. Moreover, the size of recent LLMs allows for the verbatim memorization of a large number of text sequences (e.g., Carlini et al., 2021, 2022), which necessarily contain event descriptions.

Third, the word-in-context prediction objective that LLMs are trained with is closely tied to GEK in humans. A range of psycholinguistic studies shows that humans continuously predict upcoming words in service of efficient and resource-optimal language comprehension (e.g., Altmann & Kamide, 1999; Kuperberg & Jaeger, 2016; Kutas & Federmeier, 2011; Levy, 2008; Mani & Huettig, 2012; McRae et al., 1998; Shain et al., 2020; Smith & Levy, 2013; Traxler et al., 2002) and that, in doing so, they rely extensively on their GEK to dynamically update their expectations (Bicknell, Elman, Hare, McRae, & Kutas, 2010; Ferretti, McRae, & Hatherell, 2001; Hare, Jones, Thomson, Kelly, & McRae, 2009; Matsuki et al., 2011; McRae

& Matsuki, 2009; McRae, Hare, Elman, & Ferretti, 2005). Knowledge of events is helpful for predicting words from context since it helps restrict the range of possible, plausible continuations to those that are compatible with the event interpretation triggered by the linguistic context. For example, in the context *Donna used the shampoo to wash her filthy* ___, the integration of the lexical items *wash* and *shampoo* triggers a washing subevent that renders the mention of *hair* unsurprising relative to other possible patients of a generalized washing event, such as *car* (Matsuki et al., 2011). Given the utility of deploying GEK for predicting words from context in humans, a possible strategy for LLMs to succeed in word-in-context prediction is to likewise construct rich internal, generalizable representations of event knowledge from distributional linguistic information.

Nonetheless, the complexity and scale of GEK makes modeling it a challenging target for any model relying on distributional semantic knowledge. First, language is sparse and the possible combinatorial space of events and their arguments is vast, even in the relatively small domain of transitive agent−patient interaction events that we focus on here. To assess the plausibility of an arbitrary event, a successful model of GEK must, therefore, acquire robust, generalizable representations of a vast number of actions and their associated restrictions on event participants. Many traditional and current distributional models have been argued to lack the representations of these building blocks for more complex semantic structures (Lenci, 2023; Lenci & Sahlgren, 2023; Pedinotti et al., 2021; Zhu, Li, & De Melo, 2018). The acquisition of GEK is complicated even more because the frequency with which events are reported in the pragmatically influenced texts available in the world is not a robust indicator of the frequency with which they occur in the real world (Gordon & Van Durme, 2013; see also Section 4.3). Thus, it remains unclear whether the latest generation of distributional semantic models acquire human-like robust, generalizable GEK from text co-occurrence statistics.

### *1.4. This study*

In this work, we test whether pretrained LLMs encode human-like generalized world knowledge in the domain of events. The term "event" has different meanings across disciplines and can encompass both an individual action or a sequence of several actions (Zacks, 2020; Zacks, Speer, Swallow, Braver, & Reynolds, 2007; see Kuperberg, 2021 for discussion). Some research on event knowledge in LLMs, for example, asks whether LLMs trained on word-in-context prediction encode human-like knowledge of event boundaries, investigating their capacity to replicate a fundamental aspect of human cognitive processing related to understanding sequential events in narratives (Kumar et al., 2022; Michelmann, Kumar, Norman, & Toneva, 2023; Wang, Jafarpour, & Sap, 2022).

Here, we define an event in the linguistic tradition, as a singular action along with the entities that participate in that action in a particular role (e.g., Dowty, 1989; Fillmore, 1967; Jackendoff, 1987). We focus on transitive two-participant events: agent−patient interactions, such as *The teacher bought the laptop*. Our goal here is to explore implicit knowledge of events in LLMs, operationalized as a systematic preference for generating descriptions of plausible over implausible events. We investigate five open-source LLMs: MPT, GPT-J, GPT-2, RoBERTa, and BERT, as well as a range of non-LLM distributional models.

We hypothesize that, if GEK emerges naturally from the word-in-context prediction objective, pretrained LLMs should treat plausible sentences as more likely than implausible sentences. If, on the other hand, distributional knowledge in pretrained LLMs does not consistently reflect event knowledge, their event representations would fail to systematically align with GEK.

To minimize the effect of confounding factors, we use highly controlled, syntactically simple minimal sentence pairs drawn from three datasets. In two datasets (Datasets 1 and 3), plausibility is manipulated via swapping the agent and patient of the sentence (e.g., *The teacher bought the laptop* vs. *The laptop bought the teacher*). This manipulation ensures identical word-level content within a sentence pair, such that the plausibility inference requires identifying the role played by each participant (e.g., *teacher* = agent, *laptop* = patient). In Dataset 2, plausibility is manipulated by replacing the patient of the event (e.g., *The actor won the award/battle*). The three datasets were selected to span event descriptions across a range of event participant compositions (interactions between two animate or one animate and one inanimate event participant) as well as varying degrees of semantic incongruence of the manipulated sentence (ranging from impossible to moderately implausible events). We focus on our largest dataset (Dataset 1, see Methods) for most analyses but show in Supporting Information that the findings extend to other datasets. We restrict ourselves to simple event descriptions in English, with the caveat that our results might not generalize to other languages (Atari et al., 2023; Blasi, Henrich, Adamou, Kemmerer, & Majid, 2022).

In Sections 3.1 and 3.2, we ask whether LLMs and humans assign higher likelihood scores to descriptions of plausible events compared to their implausible counterparts. In Sections 3.3 and 3.4, we investigate the degree to which these scores are *generalized*, that is, abstracted away from the surface-level properties of the input. Finally, we conduct detailed analyses of LLM performance by studying their error patterns (Section 3.5) and the nature of their internal representations of event plausibility (Section 3.6).

To foreshadow our key results, we find that LLMs possess substantial implicit event knowledge and outperform strong baseline models. In particular, they systematically prefer events that are possible (e.g., *The teacher bought the laptop*) to events that are, in the absence of contextual information, impossible (e.g., *The laptop bought the teacher*). However, LLMs are less consistent when distinguishing events that are likely (e.g., *The nanny tutored the boy*) from events that are unlikely but not impossible (e.g., *The boy tutored the nanny*), although their performance is still significantly above chance. Thus, we conclude that possible and impossible events naturally segregate in the distributional linguistic space, whereas likely and unlikely events segregate to a lesser extent, suggesting that *some but not all kinds* of event knowledge can be naturally learned by tracking distributional linguistic knowledge.

## 2. Methods

### 2.1. Sentence sets

We compare event plausibility scores in humans and language models using three sentence sets adapted from previous cognitive science and neuroscience studies (see Tables 1 and 2 for a summary).

Table 1
Sentence manipulations in Dataset 1

| Item type | Plausible? | Possible? | Sentence |
|---|---|---|---|
| animate-inanimate (AI) | Yes | Yes | The teacher bought the laptop. |
| | No | No | The laptop bought the teacher. |
| animate-animate (AA) | Yes | Yes | The nanny tutored the boy. |
| | No | Yes | The boy tutored the nanny. |

Table 2
Sentence manipulations across the three datasets

| Sentence set | Plausible? | Voice | Synonym # | Sentence |
|---|---|---|---|---|
| **Dataset 1** | Yes | Active | 1 | The teacher bought the laptop. |
| (Fedorenko et al., 2020) | | | 2 | The instructor purchased the computer. |
| | | Passive | 1 | The laptop was bought by the teacher. |
| | | | 2 | The computer was purchased by the instructor. |
| | No | Active | 1 | The laptop bought the teacher. |
| | | | 2 | The computer purchased the instructor. |
| | | Passive | 1 | The teacher was bought by the laptop. |
| | | | 2 | The instructor was purchased by the computer. |
| **Dataset 2** | Yes | Active | – | The actor won the award. |
| (Vassallo et al., 2018) | No | Active | – | The actor won the battle. |
| **Dataset 3** | Yes | Active | – | The cop is arresting the criminal. |
| (Ivanova et al., 2021) | No | Active | – | The criminal is arresting the cop. |

*2.1.1. Dataset 1—Main (based on Fedorenko, Blank, Siegelman, & Mineroff, 2020)*

This sentence set contains 391 items, each of which includes (i) a plausible active sentence that describes a transitive event in the past tense (e.g., *The teacher bought the laptop*) and (ii) the implausible version of the same sentence, constructed by swapping the noun phrases (NPs) (*The laptop bought the teacher*). The dataset also includes passive voice versions of the same sentences (*The laptop was bought by the teacher* and *The teacher was bought by the laptop*). Further, 249 of the 391 items are grouped into pairs where the sentences consist of words with synonymous, or closely related, meanings (e.g., *The teacher bought the laptop* and *The instructor purchased the computer*). For simplicity, we call those sentences "synonymous" throughout the paper.

The items are split into two types: (1) animate-inanimate (AI) items (e.g., *The teacher bought the laptop* vs. *The laptop bought the teacher; n* = 128; 76 with synonyms); (2) animate-animate (AA) items (e.g., *The nanny tutored the boy* vs. *The boy tutored the nanny; n* = 129; 82 with synonyms). Due to the animacy differences, the role reversal manipulation on AI sentences often violates the animacy selectional restrictions on the verb, making the sentence mostly semantically impossible, whereas the plausibility violations in AA sentences are more graded. Finally, the dataset includes a set of animate-animate, reversible (AA-control)

items ($n = 134$; 78 with synonyms), where both event participants are animate and both agent-patient combinations are plausible (e.g., *The cheerleader kissed the quarterback* vs. *The quarterback kissed the cheerleader*) and that we used as control in some of the analyses.

#### *2.1.2. Dataset 2 (DTFit; based on Vassallo et al., 2018)*

This sentence set contains 395 items, each of which includes (i) a plausible active sentence that describes a transitive event in the past tense, where the animate agent entity is interacting with an inanimate patient entity that is prototypical/canonical for the agent (e.g., *The actor won the award*), and (ii) the less plausible version of the same sentence, constructed by varying the inanimate patient entity (*The actor won the battle*). Plausibility depends on the entire <agent, verb, patient> triple rather than just on the <agent, verb> or <verb, patient> combination. All sentence pairs in this dataset describe interactions between an animate agent and an inanimate patient, making them most comparable to the AI sentence pairs from Dataset 1. However, unlike in Dataset 1, word content and not word order distinguishes between plausible and implausible sentences within a pair. Note further that the plausibility manipulation in this sentence set is graded: the events can be described as typical/atypical rather than possible/impossible.

#### *2.1.3. Dataset 3 (based on Ivanova et al., 2021)*

This sentence set contains 38 items, each of which includes (i) a plausible active sentence that describes a transitive event in the present tense (e.g., *The cop is arresting the criminal*), and (ii) the implausible version of the same sentence, constructed by swapping the NPs (*The criminal is arresting the cop*). All sentence pairs in this dataset describe nonreversible interactions between two animate entities, making them comparable to the AA sentence pairs from Dataset 1. As in Dataset 1, only word order but not word content distinguishes between plausible and implausible sentences within a pair.

The majority of the sentences in Datasets 1 and 3 and all sentences in Dataset 2 use single nouns as subjects and objects; a small subset of sentences in Datasets 1 and 3 use multi-word NPs (e.g., social worker). All active voice sentences in Datasets 1 and 2 and most sentences in Dataset 3 use the structure "Subject-Verb-Direct Object"; a small subset of sentences in Dataset 3 also contain indirect objects (*A doctor is using a stethoscope on the patient*). All datasets can be found at https://github.com/carina-kauf/lm-event-knowledge.

### *2.2. Human data collection*

For all three sentence sets, we compared language model predictions with human plausibility judgments. Human judgments for Dataset 2 had been previously collected by Vassallo et al. (2018) on Prolific, a web-based platform for collecting behavioral data. Participants in this experiment answered questions of the form "*How common is it for an actor to win an award*?" on a Likert scale from 1 (very atypical) to 7 (very typical). Human judgments for Datasets 1 and 3 were collected on Amazon Mechanical Turk, another web-based platform. Here, participants evaluated the extent to which each sentence was "plausible, i.e., likely to occur in the real world" on a Likert scale from 1 (completely implausible) to

7 (completely plausible). The protocol for the study was approved by MIT's Committee on the Use of Humans as Experimental Subjects (COUHES). All participants gave written informed consent in accordance with protocol requirements.

For Dataset 1 (our main dataset), we recruited 966 participants, restricting our task to participants with IP addresses in the United States. The sentences were divided into 32 experimental lists such that each of the items occurred only in one of its versions in any given list. The median response time was 20.6 min. Each participant completed between 1 and 3 lists (mean = 1.1).

Participants were included in the analyses if they satisfied all the following criteria: (i) self-reported location ("USA"), (ii) native English proficiency (evaluated via self-report and two-sentence completion trials), (iii) fewer than 20% of blank responses, and (iv) accurate responses to attention checks ("Please select the leftmost/rightmost option"). We additionally filtered participants based on their responses to the AI items (*The teacher bought the laptop* vs. *The laptop bought the teacher*), retaining participants with a minimum plausibility difference of 1 point (out of 7) between plausible and implausible items in this condition. These criteria left data from 658 participants for analysis. Each sentence had a minimum of 18 ratings (average: 22.9 ratings; maximum: 27 ratings). Participants were paid $4.25 (estimated completion time was 25 min), with payment contingent only on the attention-check questions and excessive blank responses (>30%).

For Dataset 3, we recruited 100 participants, restricting our task to participants with IP addresses in the United States. The sentences were divided into two experimental lists and each of the items occurred only in one of its versions in any given list. The median response time was 15.7 min. Each participant completed one list. We filtered the data using the same criteria as for Dataset 1, except for the sentence completion trials for assessing English proficiency (which were not included) and the minimum plausibility difference criterion. The inclusion/exclusion criteria left data from 96 participants for analysis (48 ratings per sentence). Participants were paid $2.70, with payment contingent only on the attention-check questions and excessive blank responses (>30%).

### 2.3. Model description and score estimation

#### 2.3.1. Large language models

We tested five attention-based Transformer (Vaswani et al., 2017) language models: MPT (The MosaicML NLP Team, 2023), GPT-J (Wang & Komatsuzaki, 2021), GPT-2 (Radford et al., 2019), RoBERTa (Liu et al., 2019), and BERT (Devlin et al., 2018). GPT-2, GPT-J, and MPT are unidirectional (aka autoregressive or causal) models, trained to predict upcoming words based only on left context (e.g., *The teacher bought the <MASK>*). BERT and RoBERTa are bidirectional models; their primary training task is predicting masked words in the input based both on left and right context (e.g., *The <MASK> bought the laptop*). For all Transformer models, we used pretrained implementations available via the HuggingFace Transformers library (Wolf et al., 2020). Specifically, we investigated the following model instantiations: *mpt-30b* (Number of layers, L = 48; Hidden size, H = 4096), *gpt-j-6B* (L = 28, H = 4096), *gpt2-xl* (L = 28, H = 4096), *roberta-large* (L = 24, H = 1024), *bert-large-*

*cased* (L = 24, H = 1024), that is, the largest pretrained version per model available via HuggingFace. See Table S1 for more information about the LLMs' architecture and training.

For the unidirectional LLMs, we define the sentence score as the sum of the log-probabilities of each token $w_i$ in the sequence, conditioned on the preceding sentence tokens $w_{<i}$.

For the bidirectional LLMs, we define the sentence score as a variant of the sentence's pseudo-log-likelihood score (PLL). The original PLL scoring method defines a sentence's score as the sum of the log-probabilities of each token given all other tokens (Salazar, Liang, Nguyen, & Kirchhoff, 2020; Wang & Cho, 2019). This method, however, yields inflated scores for multitoken words (Kauf & Ivanova, 2023). Here, we use the improved PLL scoring method introduced by Kauf and Ivanova (2023), which avoids this bias by masking not only the target token, but also all within-word tokens to the right of the target during inference. We show in Figs. S11 and S12 that sentence generation likelihood is a more robust indicator of event knowledge in bidirectional LLMs than other prediction-based metrics, such as last-word prediction probability or verb prediction probability for our datasets.

To encourage transparency in the NLP community, we do not report results from closed models, such as GPT-3. We also do not report results from models that have been fine-tuned on additional objectives, such as reinforcement learning from human feedback: our goal is to specifically test world knowledge encoded in the distributional patterns learned via word-in-context prediction.

### *2.3.2. Baseline models*

To investigate whether knowledge of event plausibility depends on specific linguistic patterns, we additionally compared the performance of the LLMs against four baseline models. This comparison allows us to evaluate the added value of LLMs in comparison to more "traditional" but less complex distributional semantics models, typically trained on a much smaller amount of data (Lenci & Sahlgren, 2023).

TinyLSTM is a two-layer LSTM recurrent neural network trained with a next-word prediction objective on the string data from the 1-million-word English Penn Treebank §2-21 (Marcus, Santorini, & Marcinkiewicz, 1993). Like for unidirectional LLMs, a sentence score for TinyLSTM is estimated as the sum of negative log probabilities of each token conditioned on the preceding tokens. The model is available through the LM Zoo library (Gauthier et al., 2020).

Thematic fit models the degree of semantic compatibility between an event's "prototype" verb argument, calculated from distributional text information (McRae et al., 1998), and the role filler proposed by the sentence. We follow the approach for calculating prototypical argument representations by Lenci (2011) and compute a prototype representation for the event patient slot as the centroid vector representations from the most associated entities with the predicate and agent in the sentence. However, instead of computing updates to the prototype using Distributional Memory vectors (as in Lenci, 2011), we here do the same computations using FastText (Bojanowski, Grave, Joulin, & Mikolov, 2017) static embeddings (see also Rambelli, Chersoni, Lenci, Blache, & Huang, 2020). A sentence's plausibility score is

computed as the cosine similarity between the FastText embedding of the proposed patient and the relevant prototype vector.

The Structured Distributional Model (SDM; Chersoni et al., 2019) is a model of thematic fit that computes both a *context-independent* and a *context-dependent* representation of the prototype role filler based on the current linguistic context. The context-independent representation is obtained via summing the FastText embeddings of all lexical items in the current linguistic context. The context-dependent representation is derived based on a dynamic representation of the context: given the lexical items in the current context and the syntactic function of the next word to be predicted, SDM queries a DEG to retrieve the words with the strongest statistical associations with those items for the target function (the DEG was extracted from a large number of dependency-parsed corpora: words are linked with their syntactic collocates and the links weighted with mutual information scores). It then computes the centroid of the FastText embeddings associated with the highest-ranked lexical entities according to DEG. Finally, a sentence's plausibility score is calculated as the sum of the SDM thematic fit scores for each verb argument (in our case: agent and patient), whereby each score is derived as the average cosine similarity of the argument filler's representation with the context-dependent and context-independent prototype representations of the role.

Lastly, the PPMI-syntax model quantifies the statistical association between verbs and their dependents (marked for syntactic role, i.e., $PPMI(arrest, cop_{subj}) \neq PPMI(arrest, cop_{obj})$) in terms of Positive Pointwise Mutual Information (PPMI). It is trained on the same dependency-parsed corpus as SDM. We apply Laplace smoothing and compute the plausibility score of a sentence as the PPMI score between the verb and the subject plus the PPMI score between the verb and the object.

See Supporting Information 2 for additional baseline model description details.

### 2.4. Binary accuracy estimation

To assess GEK in language models and in humans, we present them with minimally different plausible versus implausible event descriptions (Section 2.1). We evaluate their ability to assign a higher score to the plausible event description than the implausible one (Sections 2.2 and 2.3). Human scores were averaged to obtain a single score for each sentence. For each sentence pair, we assigned a score of 1 if the model/human subject pool succeeded on this task, that is, if a higher score was assigned to the plausible version of the sentence and 0 otherwise.

### 2.5. Word frequency estimation

To account for potential effects of word frequency, we estimated the average frequency of the word/phrase denoting the agent, patient, and verb of each sentence, as well as the average frequency of all words in the sentences. Frequency was operationalized as the log of the number of occurrences of the word/phrase in the 2012 Google NGram corpus. Laplace smoothing was applied prior to taking the log.

### 2.6. Probing analysis

To investigate the emergence of explicit plausibility information in LLMs, we trained a decoding probe to distinguish plausible and implausible sentences from their embeddings at different LLM layers. Separate logistic regression classifiers were trained for each model layer and the static word embedding space of the models. For each sentence, the input was the model-specific sequence summary token; the output was a binary plausibility label. The choice of model-specific sequence summary tokens followed the default settings from Huggingface Transformers: for the bidirectional LLMs, BERT and RoBERTa, we used the representation of the special token [CLS], which was prepended to each stimulus and was designed and trained specifically for sequence classification tasks. For the unidirectional LLMs, GPT-J and GPT-2, we prepared the stimulus by adding the [EOS] token to the beginning and end of the sequence and used the representation of the final token as the sequence's summary representation. For all analyses, probes were trained using 10-fold cross-validation, ensuring that plausible and implausible versions of the same sentence remain in the same split (train or test). To estimate the best-case model performance, we computed empirical ceiling values by training probes on the average human plausibility ratings for each sentence. The probe setup and the cross-validation procedure for ceiling probes were the same as for LLM probes.

To probe the generalization ability of the LLMs, we trained the classifiers on just one type of sentence (either on specific animacy combinations, AI or AA, or specific voice, active or passive) and evaluated the performance on the held-out type.

We used sklearn's (Pedregosa et al., 2011) Logistic Regression module with a liblinear solver for all probing analyses.

### 2.7. Statistical analyses

#### 2.7.1. Binary accuracy

Binary accuracy results were compared to chance performance of 0.5 using a binomial test. Tests of equal proportion were used to compare model performance to human performance, as well as AI sentence accuracy to AA sentence accuracy within each metric.

#### 2.7.2. Correlations

All reported correlations are Pearson correlations. Correlation significance was assessed using the test for correlation for paired samples (cor.test in R). Model correlation was compared to human correlation using the *cocor* package's (Diedenhofen & Musch, 2015) implementation of (Raghunathan, Rosenthal, & Rubin, 1996) test for nonoverlapping correlations based on dependent groups.

#### 2.7.3. Mixed effects modeling

We fitted separate linear mixed effects models to human ratings and each language model's scores. The key predictors for Dataset 1 were plausibility, item type (AI vs. AA vs. AA-control), and voice (active vs. passive), as well as interactions between them. We also included agent, patient, verb, and average sentence frequencies, sentence length in tokens (for LLMs) or words (for humans and baseline models). Random effects included the item

number intercept and item number by plausibility slope. For Datasets 2 and 3, the formula was simplified to account for dataset structure (i.e., no item type or voice predictors).

Continuous variables were normalized before fitting. We used dummy coding for plausibility, with "plausible" as the reference level, dummy coding for item type, with "AA" as the reference level, and sum coding for voice. The analysis was conducted using the *lme4* R package (Bates et al., 2014).

#### *2.7.4. Probing analyses*

To compare the performance of probing classifiers across LLM layers, we divided LLM layers into three same-sized groups: early, middle, and late. Within each layer group, we compared average probe performance to the ceiling value (probe trained on human ratings; see Section 2.4), as well as the linear trend within each layer group (i.e., whether classifier performance increases, decreases, or stays constant within that layer group).

In all analyses, the results were False-Discovery Rate (FDR)-corrected for the number of models within each category (humans, LLMs, and baselines). For probing analyses, the results were additionally corrected for the number of classifiers used within each analysis (e.g., 5 for generalization across trial types; 5 classifiers $\times$ 5 LLMs $=$ 25 comparisons). Analysis code and data files can be found on GitHub: https://github.com/carina-kauf/lm-event-knowledge.

## 3. Results

We report a variety of tests to establish whether pretrained LLMs are sensitive to event plausibility. In our main test (Sections 3.1 and 3.2), we investigate whether LLMs systematically assign higher scores to the plausible sentence compared to the implausible sentence within the minimal pair. We compare LLM performance with human performance (whether crowdsourced plausibility scores are higher for plausible than for implausible sentences within each pair) and with baseline model performance. Then, we move beyond the minimal pair setup to conduct detailed analyses of all sentence scores, in order to determine the relative contributions of event plausibility and surface-level properties to LLM sentence scores (Section 3.3). We investigate whether the event knowledge acquired by LLMs is *generalized* and *systematic* (Section 3.4), conduct an error analysis of LLM performance (Section 3.5), and use a probing analysis to track the emergence of explicit event plausibility signatures across LLM layers (Section 3.6).

### *3.1. All models show a gap between impossible and unlikely events*

Our main sentence set (Dataset 1) contains two types of plausible-implausible sentence pairs: AI (animate−inanimate interactions, e.g., *The teacher bought the laptop* vs. *The laptop bought the teacher*) and AA (animate−animate interactions, e.g., *The nanny tutored the boy* vs. *The boy tutored the nanny*). In most cases, AI plausibility violations result in impossible events, whereas AA plausibility violations make the event unlikely but not impossible. We

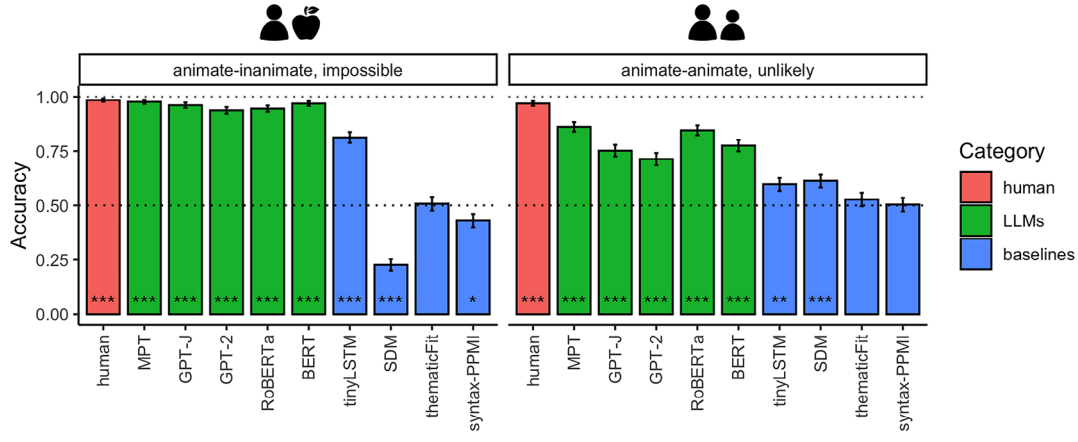


Fig. 1. Main results, Dataset 1 sentences. Human, LLM, and baseline model accuracy scores for AI (left) and AA (right) sentence pairs. Significance was established via a binomial test. Here and elsewhere, significant results are marked with asterisks ($p<.05$: *; $p<.01$: **; $p<.001$: ***). Error bars show the standard error of accuracy scores across sentence pairs.

found that all language models exhibited differential performance on these sentence sets, with substantially better results for AI than for AA sentence pairs (Fig. 1).

In the main analysis, we tested whether models systematically assign higher likelihood scores to plausible versus implausible sentences within each minimal sentence pair. For each sentence pair, a model received a score of 1 if it assigned a higher score to the plausible version of the sentence and 0 otherwise. The same procedure was performed on human plausibility ratings for each sentence pair.

### *3.1.1. AI sentence performance is high*

All models showed good performance on AI sentences (Fig. 1A, left). MPT and RoBERTa scores were not significantly different from the human accuracy of 1, and other LLMs also had high performance, although slightly lower than humans (MPT: accuracy 0.97, $\chi 2 = 2.29$, $p = .145$; GPT-J: accuracy 0.93, $\chi 2 = 7.37$, $p = .011$; GPT-2: accuracy 0.95, $\chi 2 = 4.27$, $p = .049$; RoBERTa: accuracy 0.98, $\chi 2 = 1.35$, $p = .245$; BERT: accuracy 0.95, $\chi 2 = 4.27$, $p = .044$). Baseline model performance was above chance, although not as high as that of LLMs and significantly lower than human performance; the best-performing baseline model was SDM, which was designed specifically to capture thematic fit for agent-verb-patient triplets (tinyLSTM: accuracy 0.80, $\chi 2 = 25.53$, $p<.001$; SDM: accuracy 0.90, $\chi 2 = 11.66$, $p<.001$; thematicFit: accuracy 0.73, $\chi 2 = 36.93$, $p<.001$; syntax-PPMI: accuracy 0.66, $\chi 2 = 50.74$, $p<.001$).

### *3.1.2. AA sentence performance is moderate*

On AA sentences, all LLMs still performed above chance (Fig. 1A, right) but their performance was significantly below the human accuracy of 0.95 (MPT: 0.84, $\chi 2 = 13.57$; GPT-J:

Table 3
Difference in performance between AI and AA sentence pairs

| Category | Metric | Difference | $\chi 2$ | $p$-value |
|---|---|---|---|---|
| Human | Human | 0.05 | 5.24 | .022* |
| LLMs | MPT | 0.13 | 11.21 | <.001*** |
| | GPT-J | 0.18 | 13.84 | <.001*** |
| | GPT-2 | 0.22 | 21.34 | <.001*** |
| | RoBERTa | 0.19 | 20.92 | <.001*** |
| | BERT | 0.19 | 16.88 | <.001*** |
| Baselines | tinyLSTM | 0.3 | 24.37 | <.001*** |
| | SDM | 0.41 | 48.84 | <.001*** |
| | thematicFit | 0.11 | 3.33 | .091 |
| | syntax-PPMI | 0.1 | 2.2 | .138 |

0.75, $\chi 2 = 27.12$; GPT-2: 0.74, $\chi 2 = 29.73$; RoBERTa: 0.78, $\chi 2 = 22.04$; BERT: 0.77, $\chi 2 = 24.56$; all $p < .001$). All baseline models performed at chance except for thematicFit (accuracy 0.62), indicating that information about AA event plausibility is more difficult to extract from subject-verb-object co-occurrence patterns in natural language than information about AI event plausibility.

### *3.1.3. The gap between AI and AA sentences is significant*

As shown in Table 3, humans, LLMs, and two of the baseline models all show a performance gap between AI and AA sentence sets. However, the size of the gap for the models (average 0.18 for LLMs, 0.23 for baseline models) is much larger than the one in humans (0.05), a result we discuss further in Section 4.5. LLMs and most baseline models show comparable performance on the passive voice versions of AI and AA sentences (Fig. S4).

For completeness, we also test the models on a set of AA-control items from Dataset 1, for which both sentences in a pair describe a plausible event (e.g., *The cheerleader kissed the quarterback vs. The quarterback kissed the cheerleader*). As expected, in that case, the models produced comparable scores for the two events within each pair, as did humans (Figs. S5 and S6).

### *3.1.4. Model-human score correlations also reflect the AI-AA gap*

We directly correlate model scores with human ratings (Fig. S7) and show that the correlation is moderate for AI sentences (mean LLM $r = .59$) and poor for AA sentences (mean LLM $r = .17$). Note, however, that we would not necessarily expect LLM scores to fully align with human plausibility judgments, given that the models' task is word-in-context prediction, not plausibility evaluation per se. Nevertheless, this analysis helps reveal dissociable contributions of plausibility and language-specific features on LLM sentence likelihood scores, which we explore further in Section 3.3.

#### 3.1.5. Scaling helps to partially bridge the AI-AA gap

To investigate the effect of LLM size on performance in more detail, we tested an extended set of seven unidirectional models (with MPT being the largest) on AI and AA minimal pair performance (Table S5 and Fig. S8). We found consistently high performance on AI sentences across all tested models (even the smallest, DistilGPT-2 and the base GPT-2). AA sentence performance increased steadily with model size, although, as noted above, the gap was not fully bridged even for the 30-billion-parameter MPT model.

#### 3.1.6. Quantitative analysis confirms the validity of the binary labels

To ensure that our binary accuracy results reflect meaningful plausibility differences in human ratings, we compute the average difference between plausible and implausible sentence scores within each pair. This value can range from −1 to 1 (with 1 reflecting a situation where people rated all plausible sentences as completely plausible and all implausible sentences as completely implausible). The mean difference was 0.78 ($SD = .18$) for AI sentences and 0.38 ($SD = .24$) or AA sentences, confirming the validity of our binary labels (see Section 3.5 for more details).

### 3.2. The gap in model performance between implausible and impossible events is not fully explainable by animacy or lexical variables

The gap between model performance on AI and AA sentences from Dataset 1 could be explained by several factors. First, implausible AI sentences in Dataset 1 mostly described impossible events (*The laptop bought the teacher*), whereas implausible AA sentences were often unlikely rather than impossible (*The boy tutored the nanny*), which resulted in a wider distribution of plausibility scores in humans (Fig. 3B). Second, as follows from their name, AI sentences described animate−inanimate interactions, such that switching the agent and the patient typically violated the animacy selectional restriction on the verb; in contrast, AA sentences described animate−animate interactions, so our plausibility manipulation did not violate the animacy restriction. Finally, the AA sentences were more difficult overall (human accuracy 0.95 vs. 1 for AI sentences), possibly because AA sentences had a lower average word frequency (Google Ngram log frequency of 10.8 for AA vs. 11.1 for AI). To determine whether the latter two factors might explain differential model performance, we compared model and human performance on two additional sentence sets.

#### 3.2.1. Dataset 2 (based on Vassallo et al., 2018)

This sentence set describes animate−inanimate (AI) interactions; plausibility is manipulated by varying the patient (e.g., *The actor won the award* vs. *The actor won the battle*; Table 2). Unlike AI sentences in Dataset 1, implausible sentences here are simply unlikely rather than impossible. This difference is reflected in the distribution of human judgments for this sentence set, which are less polarized than for AI sentences from Dataset 1 (mean difference 0.55; see Fig. S9 for details). If argument animacy determines model performance, their accuracy on Dataset 2 should be similarly high to that for AI sentences from Dataset 1. If, on

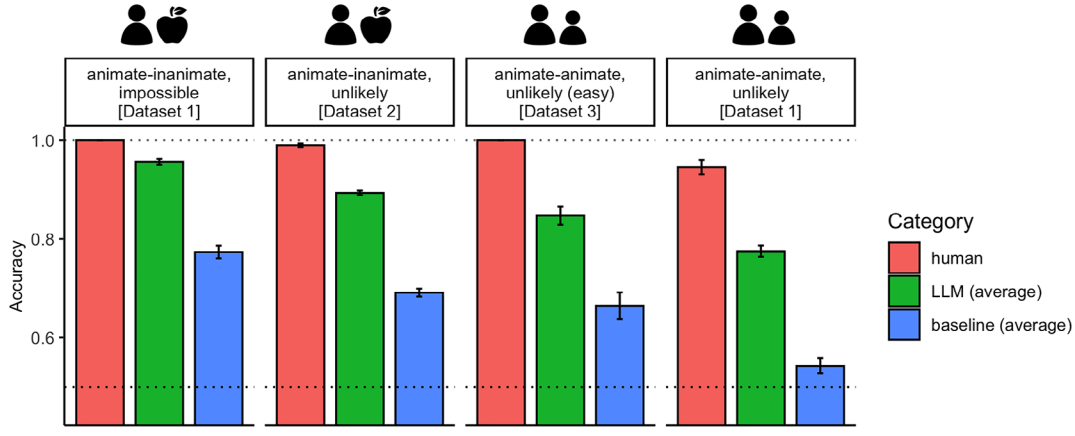


Fig. 2. Human and model performance patterns on Dataset 1 (the first and last set of bars; same data as in Fig. 1), as well as Datasets 2 and 3 (the second and third set of bars); results ordered by LLM performance. Dotted lines indicate chance-level performance.

the other hand, unlikely events are more challenging for the models to evaluate compared to impossible events, then models should perform better on AI sentences from Dataset 1.

All models scored above chance but significantly below human performance of 0.99 (MPT: 0.93, $\chi 2 = 20.8$; GPT-J: 0.89, $\chi 2 = 40.5$; GPT-2: 0.88, $\chi 2 = 46.1$; RoBERTa: 0.91, $\chi 2 = 29.5$; BERT: 0.86, $\chi 2 = 55.3$; all $p < .001$). Average LLM performance on this sentence set (0.89) is higher than on AA sentence pairs from Dataset 1 (0.78) but is lower than on possible-impossible AI sentence pairs from Dataset 1 (0.96) (Fig. 2). We, therefore, conclude that distinguishing likely and unlikely events remains a nontrivial challenge for LLMs even for AI sentences.

Further, the words in Dataset 2 are on average more frequent (log word frequency for Dataset 2: 11.5; log word frequency for AI sentences in Dataset 1: 11.1). We conclude that word frequency cannot fully account for the performance gap either.

### *3.2.2. Dataset 3 (based on Ivanova et al., 2021)*

Dataset 3 is a small sentence set from a neuroimaging study by Ivanova et al. (2021) with the same manipulation as in Dataset 1: implausible sentences are generated by switching the agent and the patient (*The cop arrested the criminal* vs. *The criminal arrested the cop*; Table 2). Both agents and patients are animate. Average word frequency is higher than in Dataset 1 sentences (Google Ngram log frequency of 11.9), and human ratings are more polarized than those of AA sentences from Dataset 1 (mean difference for Dataset 3 = 0.76). Human accuracy for distinguishing plausible and implausible sentences in this dataset was 1, meaning that the plausibility judgments for this dataset were easy and unambiguous.

All models performed above chance but below human performance, who had perfect accuracy on this task, although the difference was nonsignificant for MPT and BERT (MPT: 0.89, $\chi 2 = 2.37$, n.s.; GPT-J: 0.82, $\chi 2 = 5.66$, $p = .023$; GPT-2: 0.84, $\chi 2 = 4.52$, $p = .038$;

RoBERTa: 0.79, $\chi 2 = 6.85$, $p = .014$; BERT: 0.89, $\chi 2 = 2.37$, n.s.). Similar to Dataset 2, average LLM performance on this sentence set (0.85) falls between performance on AI sentences from Dataset 1 (0.96) and on AA sentences from Dataset 1 (0.78) (Fig. 2). Although this dataset is too small to draw definitive conclusions, the results suggest that the performance gap between impossible and unlikely events in Dataset 1 cannot simply be explained by the fact that likely-unlikely sentence pairs were more challenging.

Together, the results from Sections 3.2.1 and 3.2.2 suggest that although event participant animacy and word frequency contribute to model performance, they do not fully explain performance patterns. In particular, unlikely sentences (across animacy configurations) pose challenges for LLMs despite being easy for humans.

Baseline model performance on Datasets 2 and 3 follows similar patterns to LLMs (Fig. 2). In the remainder of the paper, we focus on LLM performance; detailed analyses of baseline model performance can be found in Supporting Information 3.

### *3.3. LLM scores are strongly influenced by surface-level sentence properties*

So far, we have focused on comparing model scores within minimal pairs. Now we ask: to what extent do model scores dissociate for all plausible and implausible sentences in our datasets?

Under the view of LLMs as knowledge bases (Petroni et al., 2019), one might expect LLM scores to strongly track real-world plausibility, such that plausibility would be the main contributing factor to the probability of a sentence being generated. However, LLM outputs are known to be sensitive to diverse surface-level factors, most notably word frequency (e.g., Gong et al., 2018), which might overwhelm plausibility in determining the overall LLM score. Thus, we conduct a series of analyses to examine the relative contributions of plausibility and surface-level factors to the overall LLM score. As a control, we use human plausibility scores, which we expect to be primarily determined by plausibility and not surface-level properties of the stimulus.

#### *3.3.1. Plausible and implausible score distributions in language models show substantial overlap*

As shown in Fig. 3A, human plausibility rating distributions for plausible and implausible sentences in Dataset 1 show little overlap (mean difference for AI sentences = 0.78, AA sentences = 0.38). In contrast, the distributions of likelihood scores assigned to plausible versus implausible sentences under language models show significant overlap (mean difference for LLMs: AI = 0.19, AA: 0.06; for baseline models: AI = 0.09, AA: 0.01). This suggests that language model scores are determined predominantly by factors other than plausibility, such as word frequency and sentence length.

#### *3.3.2. Switching the agent and the patient strongly influences human plausibility judgments but not LLM scores*

Our plausibility manipulation (switching the agent and patient in a sentence) was specifically designed to alter the plausibility of the described event while preserving the identities of

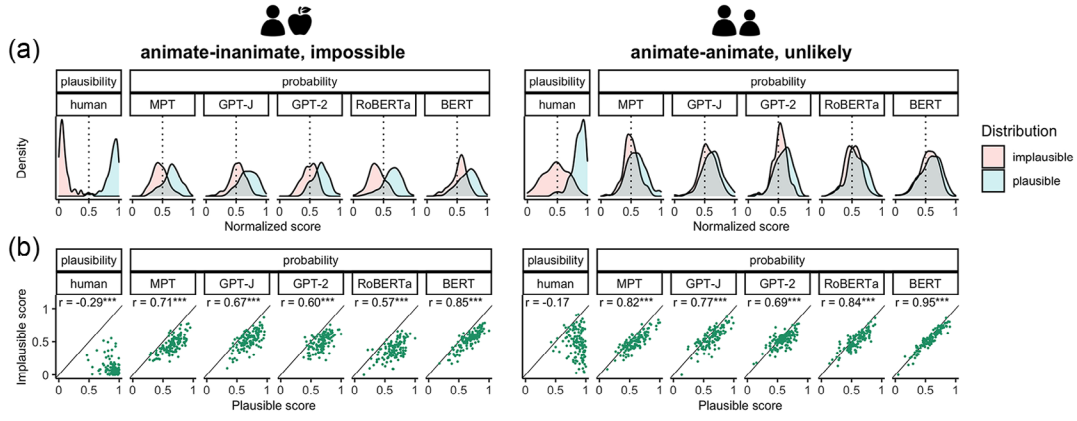


Fig. 3. Human plausibility rating and model probability score distributions. (A) Density plots for plausible and implausible sentences. The dotted line shows the midpoint on the normalized score scale (0.5). (B) Correlation plots for plausible and implausible sentences. Each dot represents a sentence score. The diagonal is an identity line. Annotations show Pearson *r* correlation values and significance levels. See Fig. S1 for detailed analyses of the score distributions for the baseline models.

individual words. If LLM scores depend on word-level properties (such as word frequency), the correlation between the scores for the two versions should be positive.

As shown in Fig. 3B, human plausibility judgments show a negative correlation for plausible and implausible versions of the same AI sentence ($r = -.29$, $p < .001$) and a nonsignificant correlation for AA sentences ($r = -.17$, $p = .06$), indicating that word-level properties do not influence the average plausibility rating of a sentence pair. Conversely, under all LLMs, sentence likelihood scores exhibit a strong positive correlation (ranging from 0.57 for RoBERTa on AI sentences to 0.95 for BERT on AA sentences), indicating that LLM scores are largely driven by individual word features, rather than by event plausibility. This trend is more pronounced for AA than AI sentences, presumably due to a smaller relative contribution of plausibility—a hypothesis we explore next.

### *3.3.3. Both plausibility and surface-level features predict LLM scores: Mixed effects modeling*

To systematically test how different factors contribute to individual sentence scores, we fitted mixed effects models to likelihood scores from each model and to human plausibility judgments (Table 4; see Methods for model and contrast definition). Because we normalize the scores for each metric (humans and models), the resulting coefficients can be interpreted as effect sizes.

As expected, human plausibility ratings are primarily driven by plausibility manipulations. For plausibility, we consider two main contrasts: (a) implausible AI sentences (*The laptop bought the teacher*) versus implausible AA sentences (*The boy tutored the nanny*) and (b) implausible AA sentences versus plausible AA sentences (*The nanny tutored the boy.*). In

Table 4
Mixed effects modeling results

| | | Plausibility | Probability | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Human | MPT | GPT-J | GPT-2 | RoBERTa | BERT | Mean across LLMs |
| **Core effects** | **Implausible AA > Plausible AA** | **−0.38*****| **−0.08*****| **−0.07*****| **−0.06*****| **−0.07*****| **−0.04*****| **−0.07** |
| | **Implausible AI > Implausible AA** | **−0.37*****| **−0.11*****| **−0.11*****| **−0.11*****| **−0.2*****| **−0.12*****| **−0.13** |
| Surface-level effects | Voice | | | | | −0.06*** | −0.13*** | −0.04 |
| | Agent frequency | | 0.03*** | 0.03*** | 0.02*** | | −0.01* | 0.01 |
| | Patient frequency | | 0.03*** | 0.03*** | 0.02*** | | −0.01* | 0.02 |
| | Verb frequency | | | | | | | 0 |
| | Avg. word frequency | | | | | 0.03** | | 0.01 |
| | Sentence length | | −0.02*** | −0.02*** | −0.02*** | −0.03*** | −0.07*** | −0.03 |
| | Voice × Sentence (AA>control) | | | | | | | 0.01 |
| | **Voice × Sentence (AI>AA)** | **0.03**** | | | **0.03**** | **0.04***** | | **0.02** |
| | Plausibility × Voice × Sentence (AA>control) | | | | | | | 0 |
| | **Plausibility × Voice × Sentence (AI>AA)** | **−0.07***** | | | | | **0.04***** | **0.01** |

*Note*. Variables that have a significant effect on human plausibility judgments are highlighted in bold. Sentence length is measured in tokens for LLMs, in words for humans. See Table S2 for baseline model results. See Tables S6 and S7 for the same analysis for Datasets 2 and 3.

humans, the effect of AI versus AA plausibility violation (−.37) is as strong as the implausibility effect for AA sentences (−.38).

All LLMs are also sensitive to both plausibility effects when assigning string likelihood scores; however, these effects are much weaker than the effects in human plausibility judgments, and the implausible AI>implausible AA effect (−.13) is larger than the implausible AA>plausible AA effect (−.07), consistent with the performance gap that we observed for AI and AA sentences.

In addition, model probabilities but not human plausibility judgments are sensitive to the main effects of surface-level sentence properties. Each LLM's performance on the critical task is affected by at least three of the following factors: voice, agent frequency, patient frequency, average word frequency, and sentence length, whereas human plausibility judgments are not affected by any of these features.

Finally, even in humans, the AI implausibility effect is modulated by some surface-level properties. Compared to AA sentences, humans are likely to assign more polarized scores to AI sentences presented in active voice than in passive voice (higher for plausible, lower for implausible). GPT-2 and RoBERTa likelihood scores partially capture this effect, and BERT shows an effect in the opposite direction, penalizing passive implausible AI sentences more harshly. The best-performing LLM, MPT, fails to capture the fine-grained effects of surface-level properties on human judgments.

Overall, the mixed-effects model analysis is consistent with other analyses. All LLMs show a significant effect of plausibility on resulting sentence likelihood scores, indicating that they are sensitive to generalized event knowledge. Yet, we still observe a performance gap between AA and AI sentences and a strong effect of surface-level linguistic properties on LLM sentence scores that diverge from those of human plausibility judgments, indicating that raw probability of an event description cannot be used directly as an indicator of its plausibility.

### *3.4. LLMs generalize well across syntactic sentence variants, but only partially across semantic sentence variants*

In the previous section, we demonstrated that LLM sentence scores are strongly influenced by surface-level sentence features, a factor that might negatively affect these models' ability to generalize. In this section, we directly tested how well LLM scores generalize across different forms of a sentence. To do so, we evaluated the extent to which model scores generalize across sentence voice (active vs. passive) and across sentences with synonymous, or closely related, meanings.

#### *3.4.1. LLMs generalize across active and passive sentences*

To test invariance to sentence syntax, we calculated the Pearson correlations between the active and passive voice versions of the same sentence (*The teacher bought the laptop* vs. *The laptop was bought by the teacher*; Fig. 4A). Human scores were highly correlated ($r = .96$, $p<.001$), indicating that human plausibility ratings are indeed invariant to sentence voice. LLM likelihood scores were also strongly correlated (max: BERT, $r = .93$; min:

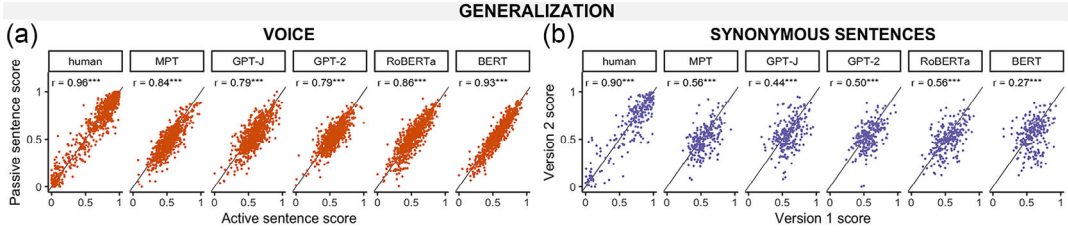


Fig. 4. Generalization results. (A) Human and LLM scores for active voice and passive voice versions of the same sentence (e.g., "The author finished the novel." vs. "The novel was finished by the author."). (B) Human and LLM scores for sentences that have synonymous, or closely related, meanings (e.g., "The author finished the novel." vs. "The writer completed the book."). Each dot represents a sentence score. The diagonal is an identity line. See Fig. S2 for baseline model results.

GPT-J/GPT-2, $r = .79$; all $p<.001$), indicating that LLMs can successfully generalize across active and passive voice forms of the same sentence.

### *3.4.2. LLMs show some generalization across synonymous sentences*

To test invariance to specific lexical forms, we compared scores for sentence pairs where subject, verb, and object words were synonymous, or closely related in meaning (*The teacher bought the laptop* vs. *The instructor purchased the computer*; Fig. 4B). Human judgments were highly correlated across synonymous sentence pairs ($r = .90$, $p<.001$), indicating that they are largely invariant to specific word identity. LLMs showed some generalization (max: MPT and RoBERTa, $r = .56$; min: BERT, $r = .27$; all $p<.001$), indicating that these models are somewhat consistent in assigning scores to synonymous utterances, but this relationship is far weaker than that observed in humans or than LLMs' syntactic generalization capabilities. This result is consistent with the results in Section 3.3.3, which showed that the models are sensitive to lexical-item-level properties, such as word frequency, and presents a potential challenge for robust representations of generalized event knowledge in LLMs.

## *3.5. LLM deviations from ground-truth labels are partially, but not fully explained by plausibility violation strength*

To understand the nature and severity of LLM errors, we conducted a quantitative and a qualitative analysis of the sentence pairs that most LLMs got wrong.

We first tested whether the severity of the plausibility violation correlates with model performance. To do so, we correlated the violation magnitude in each sentence pair (operationalized as the difference between human scores for plausible and implausible sentence versions) and the number of LLMs (0 through 5) that correctly evaluated that sentence pair. For both AI and AA sentences, we observed a moderate positive correlation, suggesting that sentence pairs that are more difficult for humans to decide are also more challenging for LLMs.

Then, we conducted a qualitative analysis of sentence pairs that most LLMs got wrong (Table 5). We found that these include several sentence pairs where human judgments actually deviated from ground truth labels (e.g., *The orderly assisted the dentist* vs. *The dentist*

Table 5
All sentence pairs (out of 391) that were evaluated correctly by at most 1 LLM, ordered by human score difference from largest to smallest

| | Trial type | #LLMs correct (of 5) | Human score difference | Plausible sentence | Implausible sentence |
|---|---|---|---|---|---|
| 1 | AA | 2 | 0.61 | The principal scolded the child. | The child scolded the principal. |
| 2 | AA | 1 | 0.54 | The craftsman taught the trainee. | The trainee taught the craftsman. |
| 3 | AA | 1 | 0.54 | The lion chased the tour-guide. | The tour-guide chased the lion. |
| 4 | AA | 2 | 0.49 | The brunette tipped the busboy. | The busboy tipped the brunette. |
| 5 | AI | 2 | 0.45 | The milliner adorned the fedora. | The fedora adorned the milliner. |
| 6 | AA | 2 | 0.26 | The impersonator conned the inspector. | The inspector conned the impersonator. |
| 7 | AA | 1 | 0.24 | The vagabond revered the priest. | The priest revered the vagabond. |
| 8 | AA | 2 | 0.21 | The deceiver imitated the conqueror. | The conqueror imitated the deceiver. |
| 9 | AA | 0 | 0.16 | The environmentalist cautioned the tobacconist. | The tobacconist cautioned the environmentalist. |
| 10 | AA | 1 | 0.12 | The warmonger terrorized the gunsmith. | The gunsmith terrorized the warmonger. |
| 11 | AA | 2 | 0.12 | The biker defied the trainer. | The trainer defied the biker. |
| 12 | AA | 1 | 0.11 | The nomad cherished the clergyman. | The clergyman cherished the nomad. |
| 13 | AA | 2 | 0.1 | The genius shocked the cousin. | The cousin shocked the genius. |
| 14 | AA | 1 | 0.06 | The prodigy surprised the relative. | The relative surprised the prodigy. |
| 15 | AA | 1 | 0.03 | The neuroscientist overwhelmed the lab assistant. | The lab assistant overwhelmed the neuroscientist. |
| *16* | *AA* | *1* | *−0.01* | *The liar emulated the victor.* | *The victor emulated the liar.* |
| *17* | *AA* | *2* | *−0.04* | *The reviewer criticized the right-winger.* | *The right-winger criticized the reviewer.* |
| *18* | *AA* | *1* | *−0.07* | *The pixie mesmerized the ogre.* | *The ogre mesmerized the pixie.* |
| *19* | *AA* | *2* | *−0.11* | *The orderly assisted the dentist.* | *The dentist assisted the orderly.* |

*Note.* Sentences where the human ratings also deviated from the ground truth labels are grayed out. See Table S3 for baseline model results. See Tables S8 and S9 for the same analysis for Datasets 2 and 3.

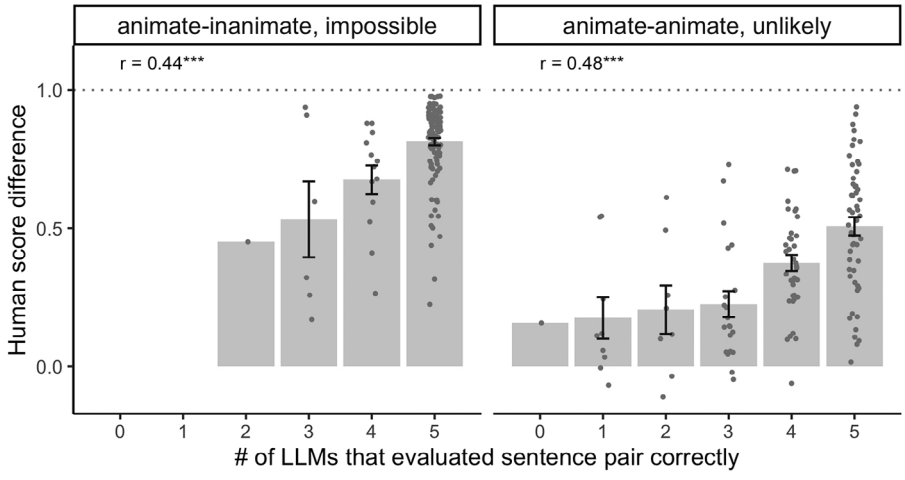


Fig. 5. Error analysis for Dataset 1. The number of LLMs (out of 5) which evaluated a given sentence pair correctly correlates with the magnitude of the human score difference between plausible and implausible versions of that sentence (the higher the difference, the better humans are at distinguishing plausible and implausible sentence versions). Each dot is a minimal sentence pair; error bars denote standard errors of the mean. See Fig. S3 for baseline model results. See Fig. S10 for the same analysis for Datasets 2 and 3.

*assisted the orderly*; see Table S4), but in two-thirds of the cases, there was at least a 0.1 difference between plausible and implausible sentence ratings in humans. Some errors might be explained by low-level features of the input such as nonstandard spelling (e.g., *tour-guide* instead of *tour guide*); some might be caused by low-frequency words (e.g., *milliner*) that were underrepresented in the models' training data; and some might reflect a failure to identify typical agent/patient roles (e.g., most LLMs fail to identify *trainee* as a typical patient for the verb *taught*, even though human judgments in this example are rather unambiguous). Overall, we conclude that (1) many of the models' errors are "reasonable," being caused by ambiguous event plausibility labels and nonstandard spelling; (2) the knowledge gap for unlikely (AA) sentences cannot be fully explained by such "reasonable" errors.

### *3.6. Internal representations of event plausibility generalize across sentences*

The previous sections have investigated the behavioral performance of LLMs in distinguishing plausible and implausible events. Here, we ask: is the distinction between plausible and implausible events encoded in the LLMs' representational spaces? Can a linear classifier trained to distinguish plausible and implausible events generalize to new sentences? If so, an LLM may have learned to represent plausible and implausible events in systematically different ways, a strategy that might help it to generalize in spite of its sensitivity to surface-level properties.

We find that sentence plausibility is indeed linearly decodable from internal LLM representations (Fig. 5). Consistent with our main results (Section 3.1), impossible AI sentences have a much stronger plausibility signature than unlikely AA sentences, with AI-to-AI classifier

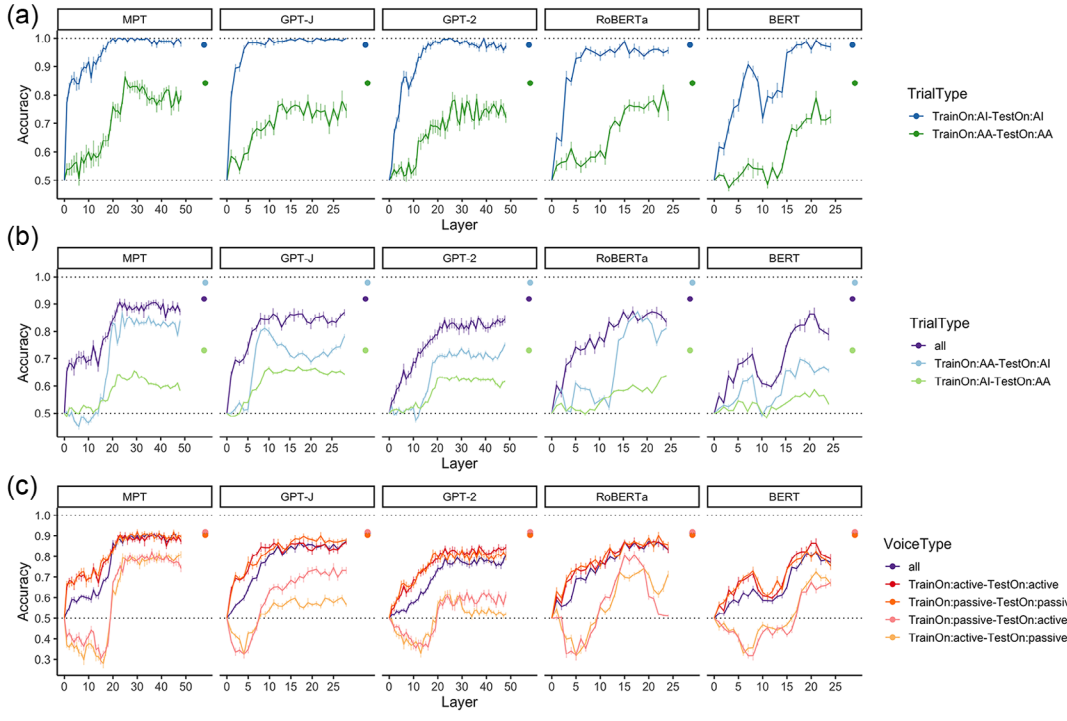


Fig. 6. Classification accuracies for linear probes trained to differentiate plausible from implausible event descriptions in model embeddings. (A) Performance within an item type (animate-inanimate, AI, vs. animate-animate, AA), active voice. (B) Generalization across item type, active voice. (C) Generalization across active and passive voice. The dots denote classification accuracy of probes trained on human scores (which can serve as an empirical ceiling value). Dotted lines indicate chance-level and ideal performance. Error bars show the standard error of the mean across the 10 cross-validation folds.

performance reaching ceiling for late and, in some cases, middle model layers, whereas AA-to-AA classifier performance on most LLM representations reaches above-chances levels later and (except for certain middle layers in MPT) generally falls short of the ceiling (Fig. 6A). The successful performance of both classifiers indicates that plausibility is one of the organizing dimensions of the underlying distributional spaces for middle and late layers.

The fact that sentence representations of later model layers are more suitable for decoding plausibility than those of earlier layers is consistent with previous results showing that semantic information tends to be encoded more strongly in later layers (Belinkov et al., 2017; Papadimitriou, Futrell, & Mahowald, 2022; Tenney, Das, & Pavlick, 2019). The trend we observed in one of the models, MPT, where mid-layer performance exceeded late-layer performance, should be examined further in other large (multi-billion parameter) models.

Next, we examined how well the plausibility signature generalizes across different sentence types (AI vs. AA) and across different surface-level forms (active vs. passive).

Generalizing across sentence types is only partially successful. Generalizing to AA sentences from AI sentences leads to a drop in classifier accuracy compared to testing an

AI-trained classifier on AI sentences (Fig. 6B). In contrast, probes that are trained to distinguish plausible versus implausible AA sentences have similar performance on AI and AA test sets, although they fall short of the probes trained and evaluated on sentence representations from both sentence sets (labeled "all" in the figure).

Generalizing across sentence voice is possible given the right training regime. A classifier trained and evaluated on all sentences (active and passive) performs as well as classifiers trained and evaluated on sentences in only one voice (Fig. 6C), a result consistent with good voice generalization performance in Section 4.4. However, active-to-passive and passive-to-active classifiers perform substantially worse, with below-chance accuracy in early layers and sometimes at-chance accuracy in late layers, indicating that these classifiers leverage surface-level plausibility signatures that do not generalize across surface forms. Thus, LLM sentence embeddings contain both syntax-specific and syntax-invariant plausibility information.

Overall, from the probing results, we conclude that event plausibility is linearly decodable from LLM sentence embeddings. This plausibility information becomes salient in middle LLM layers and remains high thereafter, making it possible to use this information during output generation. Similar to the behavioral results, there is a performance gap between AI and AA sentences, and the plausibility signature only partially generalizes across AI and AA event types (although it can generalize across sentence voice).

For detailed statistical comparisons, see Tables S10 and S11. For extended probing results, see Figs. S13 and S14; Table S12.

## 4. Discussion

Can generalized event knowledge emerge from distributional information encoded in language? To find out, we compared the likelihood scores that pretrained LLMs assigned to plausible versus implausible event descriptions. To minimize the putative influence of confounding factors, we used syntactically simple, tightly controlled minimal pair sentences. We demonstrate that LLMs acquire substantial event knowledge and improve over strong baseline distributional models, especially when it comes to distinguishing possible and impossible events (*The teacher bought the laptop* vs. *The laptop bought the teacher*). However, they are less consistent when assigning probabilities to likely events versus events that are unlikely but not impossible (*The nanny tutored the boy* vs. *The boy tutored the nanny*). Using three different sentence sets, we demonstrated that this gap in performance cannot be fully explained by the animacy of the event participants or word frequency.

We further conducted a rigorous set of analyses to elucidate the relationship between an LLM sentence score (which reflects its generation probability) and plausibility, showing that LLM scores depend both on sentence plausibility and surface-level sentence properties. In generalization analyses, we found that both LLM and human scores are consistent for active and passive voice versions of the same sentence, but LLMs are less consistent than humans for synonymous sentence forms. Lastly, we found that sentence plausibility is linearly decodable from internal LLM representations, with the same gap between impossible and unlikely event performance as that observed in behavioral tests. We conclude that sentence plausibility is a

major contributor to sentence generation probability, although this relationship is less clear-cut for likely versus unlikely event descriptions.

### 4.1. In case of impossible events, LLMs might leverage selectional restrictions

LLMs in our study were close to ceiling when distinguishing possible and impossible events. A notable feature of the impossible event descriptions in our datasets is the violation of selectional restrictions on the verb, that is, the set of semantic features that a verb requires of its arguments (e.g., requiring an agent to be animate) (Chomsky, 1965; Katz & Fodor, 1963; Levin, 1993). When plausibility violations were not driven by selectional restrictions (as in the "unlikely" sentence sets), model performance dropped.

Our findings suggest that selectional restrictions are a linguistic property that is learnable from corpus data (as also confirmed by the large number of experiments with computational methods for selectional restriction acquisition from texts; e.g., Erk, 2007; Thrush, Wilcox, & Levy, 2020) and whose violations are meaningfully distinct from violations of graded world knowledge (Warren & McConnell, 2007; cf. Matsuki et al., 2011). Computational evidence suggests that BERT models are able to generalize their knowledge of selectional restrictions in novel word-learning paradigms (Thrush et al., 2020) and can partially rely on the semantics of the head predicate to predict upcoming event participants (Metheniti, Van de Cruys, & Hathout, 2020). The asymmetry in performance on possible/impossible versus likely/unlikely events was independent from the specifics of LLM architecture and training and was additionally present, in an even more marked way, in our baseline models. Furthermore, a classifier probe trained on possible versus impossible sentence embeddings performed almost perfectly on other sentences from the same category but failed to generalize to likely versus unlikely events, indicating that selectional restrictions have a distinct representational signature. These results are consistent with psycholinguistic evidence from reading times and Electroencephalography (EEG) indicating that violations of selectional restrictions and violations of world knowledge evoke distinct processing signatures (e.g., Paczynski & Kuperberg, 2012; Sitnikova, Holcomb, Kiyonaga, & Kuperberg, 2008; Warren, Milburn, Patson, & Dickey, 2015; cf. Hagoort, Hald, Bastiaansen, & Petersson, 2004).

The fact that selectional restrictions are easier to learn from distributional linguistic data than graded event likelihood is an important distinction, as both of these factors affect plausibility judgments in humans (e.g., Hagoort et al., 2004; Warren et al., 2015). To verify and extend our findings, future work should test LLMs' knowledge of selectional restrictions on features other than animacy, such as the physical constraints that a predicate places on its patients (Wang, Durrett, & Erk, 2018), evaluate their performance on impossible events that do not violate selectional restrictions per se (e.g., *She gave birth to her mother, The man was killed twice*, or *After 10 coin tosses, she got 12 heads.*), and conduct more targeted tests of agent-verb and patient-verb plausibility (Metheniti et al., 2020).

### 4.2. LLMs can infer thematic roles

The stimuli in Datasets 1 and 3 are constructed such that the model has to leverage word order information to successfully determine event plausibility. LLMs successfully accomplish

this task for most possible versus impossible events and for a number of likely versus unlikely events. Furthermore, they produce highly correlated scores for active and passive versions of the same sentence, suggesting that thematic role information generalizes beyond a specific word order.

The probing results produce additional insight into the emergence of thematic role information in the LLMs (Fig. 5B). A probe that is trained on a mix of active and passive sentences performs as successfully as the probe trained and tested on only one voice type, suggesting that plausible and implausible sentence embeddings in late LLM layers are linearly separable by the same hyperplane across syntactic structures. This finding aligns with recent computational work showing that even though most sentences in the language input describe prototypical events (Mahowald et al., 2023), LLMs are able to correctly represent the argument structure of nonprototypical event descriptions in late layers (Papadimitriou et al., 2022).

Despite LLMs' general success in thematic role inference, some confusion about thematic role assignment might remain for unlikely events in our main dataset. These events describe animate−animate (AA) interactions. Animate direct objects are treated as a special case in many languages. In Spanish, for example, they are differentially marked with a preposition (e.g., Aissen, 2003; Bossong, 1991). Even though English (the language that we test here) does not overtly mark any direct objects, it could be that the correct thematic role assignment remains ambiguous for AA sentences in a way that it does not for AI sentences. In humans, this ambiguity can lead to a reinterpretation of the sentence as plausible even when the word order indicates an implausible interpretation (e.g., Gibson, Bergen, & Piantadosi, 2013); a fine-tuned conversational LLM, ChatGPT-3.5, exhibited a similar bias (Cai et al., 2023), suggesting that in LLMs, like in humans, plausibility priors might overrule thematic role assignment.

### *4.3. The "reporting bias" in language corpora makes it harder to distinguish likely and unlikely events*

A core challenge for modeling plausibility based on linguistic input is the fact that the frequency with which events are described in the language is not a reliable predictor of the frequency with which events occur in the real world. Because much of our world knowledge is shared across individuals (e.g., McRae et al., 2005) and human communication is shaped by efficiency (Gibson et al., 2019) and cooperation (Grice, 1975), language is biased toward reporting extraordinary facts and events rather than the trivial (Gordon & Van Durme, 2013). Many commonsense facts about the world are thus presupposed rather than stated explicitly; in contrast, unusual events are discussed extensively. As a result, likely events are underrepresented in linguistic corpora, whereas unlikely events are overrepresented.

The "reporting bias" of rare and newsworthy events in language corpora has traditionally provided difficulty for modeling semantic knowledge via text mining (e.g., Lucy & Gauthier, 2017). Recent studies probing world knowledge in LLMs show that although the generalization capabilities of these models are able to overcome the reporting bias to some extent (Shwartz & Choi, 2020; Weir, Poliak, & Van Durme, 2020), they still tend to reflect biases that exist in their training corpus (Shwartz & Choi, 2020; Vig et al., 2020; Zmigrod, Mielke,

Wallach, & Cotterell, 2019). As a result, one explanation of the performance gap that we observe for likely versus unlikely events in LLMs could be that unlikely events are overrepresented in the corpus, leading the models to predict them as frequently as likely events. In contrast, impossible events are nearly absent from the training data, and so the models correctly assign them low likelihood scores.

It is worth noting that the reporting bias present in pragmatically influenced natural language also affects concept learning in humans: blind people's beliefs about the canonical color of animals (Kim et al., 2019), for example, are consistent with the inadequate color information encoded in sighted people's linguistic productions (Ostarek et al., 2019). This, along with the successful acquisition of many other visual concepts by the blind (e.g., Landau & Gleitman, 1985; Marmor, 1978; Wang, Men, Gao, Caramazza, & Bi, 2020), implies that learning from distributional linguistic information is a likely, though not the only, strategy that humans adduce to organize facets of world knowledge, especially those to which they do not have direct sensorimotor access.

A possible solution to overcoming the reporting bias would be to adjust the event distribution via injecting manually elicited knowledge about object and entity properties into models (Wang et al., 2018; although see Porada, Suleman, Trischler, & Cheung, 2021) or via data augmentation (e.g., Zmigrod et al., 2019). Alternatively, information about event typicality might enter LLMs through input from different modalities, such as visual depictions of the world in the form of large databases of images and/or image descriptions (Bisk et al., 2020). Distributional models trained on multimodal data have indeed been shown to outperform text-only trained models in overcoming the reporting bias for visual concept knowledge (e.g., Paik et al., 2021; Zhang et al., 2022). In the future, we plan to extend our analysis of GEK to multimodal LLMs (e.g., CLIP; Radford et al., 2021) in order to investigate the role of extralinguistic evidence, which might reduce the impact of the reporting bias and better simulate the multimodal information that humans use to acquire GEK. Finally, a training objective that emphasizes robust, generalizable event representations might lead to more robust GEK knowledge than word-in-context prediction, although what such an objective would look like remains to be discovered.

### *4.4. Distributional language models are good models of language but imperfect models of world knowledge*

We have shown that the probability for generating a particular sentence under a given LLM depends both on plausibility and on surface-level features of that sentence, such as word frequency. This result is largely expected, because distributional models are naturally geared toward producing more frequent tokens more often. However, it does result in a high overlap between the score distributions we observe for plausible and implausible sentences, meaning that many implausible sentences have higher likelihood generation simply because they contain frequent words.

The fact that LLMs are sensitive to both sentence plausibility and surface-level features makes them good candidate models of human language processing. On the one hand, sentence plausibility substantially facilitates language processing in humans (e.g., Bicknell et al., 2010;

Federmeier & Kutas, 1999; Kutas & Hillyard, 1984; McRae & Matsuki, 2009). On the other hand, humans are also sensitive to lexical frequency effects when processing linguistic inputs (e.g., Broadbent, 1967; Goodkind & Bicknell, 2021; Haeuser & Kray, 2022; Rayner & Duffy, 1986) and can use both linguistic knowledge and event knowledge in real time depending on task demands (Willits, Amato, & MacDonald, 2015). As a result, LLM scores are a good predictor of human reading times (Oh & Schuler, 2023; Oh, Clark, & Schuler, 2022; Shain et al., 2022), neural predictability signatures like N400 (Michaelov, Bardolph, Van Petten, Bergen, & Coulson, 2023; Szewczyk & Federmeier, 2022), and brain response patterns to individual sentences (e.g., Caucheteux & King, 2022; Schrimpf et al., 2021; Tuckute et al., 2023).

However, sensitivity to surface-level features of the input can make LLMs unreliable as knowledge bases. Due to this sensitivity, they produce inconsistent results when the same description is phrased differently (Elazar et al., 2021a; Ravichander et al., 2020; Ribeiro et al., 2020), produce unsystematic judgments (Talmor, Elazar, Goldberg, & Berant, 2020), hallucinate facts (Ji et al., 2023; Liu et al., 2022), fail to learn commonsense event schemas (Pedinotti et al., 2021), and generalize only weakly across synonymous descriptions of the same event (Section 4.4). The ability to abstract away from specific inputs is a key feature of GEK; thus, the ability of future language-based models to acquire robust, flexible event schemas will depend crucially on their ability to generalize beyond corpus statistics.

Even though world knowledge and language processing behavior are closely linked in humans, world knowledge and language are two fundamentally different capabilities that have been shown to dissociate in humans (e.g., Caramazza, Berndt, & Brownell, 1982; Lambon Ralph, Jefferies, Patterson, & Rogers, 2017; Patterson, Nestor, & Rogers, 2007), including in a study that specifically evaluated event plausibility (Ivanova et al., 2021). We, therefore, speculate that the acquisition of robust, statistics-invariant world knowledge representations would require a different objective function from that required for acquiring linguistic proficiency (Mahowald et al., 2023). The word-in-context prediction objective, which enables LLMs to excel at acquiring formal linguistic competence, encourages pretrained LLMs to organize their semantic spaces mainly by relatively simple features such as similarity and association (Lenci, 2023). This organization principle, however, does not always lead to robust concepts and relations, which are useful for natural language understanding tasks and serve as important units for developing more complex semantic structures (Lenci, 2023; Lenci & Sahlgren, 2023).

Based on our results and on studies from the literature, we conclude that the word-in-context prediction objective alone is suitable for acquiring a wealth of event knowledge but cannot ensure the consistency of these representations. Thus, in both humans and models, distributional linguistic knowledge is not a replacement for GEK but rather a useful foundation for further enrichment and fine-tuning of generalized semantic representations.

### *4.5. Generating descriptions of unlikely events: A feature rather than a flaw?*

The fact that LLMs' distributional linguistic knowledge does not limit them to the realm of plausible events could be considered a feature rather than a flaw. The power of language is not

only in its ability to convey factual knowledge: language allows humans to brainstorm, fantasize, discuss counterfactuals, speculate, and dream. With enough backstory, even an impossible event like *The laptop bought the teacher* can be rendered plausible, eliminating the processing difficulty in humans (e.g., Jouravlev et al., 2019; Nieuwland & Van Berkum, 2006; Warren, McConnell, & Rayner, 2008) and in LLMs (Michaelov et al., 2022). Thus, restricting the models to the realm of a priori plausible events would handicap their potential as models of human language. Of course, in the absence of contextual information (as is the case in our study), we would still expect LLMs to generate plausible event descriptions more often than implausible ones. However, an overly strong alignment between an LLM and a knowledge base will likely be counterproductive for its linguistic fluency.

Finally, a naïve approach to pretrained LLMs as knowledge bases overlooks their core design feature: they are prediction machines that aim to faithfully mimic all properties of the input, not simply semantic plausibility. As seen in Section 3.3, LLM scores are sensitive to a variety of surface-level properties of the stimulus that need to be factored out to receive a more faithful estimate of plausibility. Therefore, LLMs should be regarded at most only as partial models of human semantic plausibility. In turn, if the goal is to directly compare LLM scores with human scores, one should consider a human metric that is more appropriate, such as reading times (Oh et al., 2022).

Prediction-based LLMs are an important tool for investigating which cognitive capacities can, in principle, rely on distributional linguistic knowledge. Contemporary LLMs show that large amounts of world knowledge can be learned from language alone with a simple word-in-context prediction objective, yet controlled, targeted manipulations like the ones used in this study can also highlight areas of knowledge where LLM behavior is not yet fully aligned with human behavior. Future work should explore the extent to which LLMs master other types of event knowledge, such as knowledge of typical/possible event sequences, knowledge of impossible events that do not violate selectional restrictions per se, and the extent of their sensitivity to selectional restrictions other than animacy. Furthermore, the fact that LLMs in our study sometimes perform below humans even on syntactically simple sentences (*The X Ved the Y*) suggests that testing them on longer sequences of text might uncover even larger deviations from GEK. Overall, detailed investigations of world knowledge in distributional language models are a valuable source of evidence for clarifying the relationship between language and broader cognition.

## Acknowledgments

We thank Josh Tenenbaum, Roger Levy, Jacob Andreas, and HuthLab members for helpful comments. This collaborative work was made possible thanks to the MIT-UNIPI Project, a grant from the MISTI Global Seed Fund (the MIT-Italy program). CK was supported by the K. Lisa Yang Integrative Computational Neuroscience (ICoN) Center at MIT. AI was supported by the Whitaker Health Sciences Fund Fellowship from MIT and by MIT Quest for Intelligence. GR contributed to this work during her PhD granted by the University of Pisa. EC was supported by the General Research Fund "Modeling Generalized Event

Knowledge for Noun Compound Interpretation and Prediction with Vector Spaces and Transformers" (B-Q0AH). EF was supported by NIH awards R01-DC016607, R01-DC016950, and U01-NS121471, as well as by research funds from the McGovern Institute for Brain Research, the Brain and Cognitive Sciences Department, the Simons Center for the Social Brain, and the Middleton Professorship. This research was also partly funded by PNRR—M4C2—Investimento 1.3, Partenariato Esteso PE00000013—"FAIR—Future Artificial Intelligence Research"—Spoke 1 "Human-centered AI," funded by the European Commission under the NextGeneration EU programme. For the specific concerns of the Italian academic attribution system, GR is responsible for Section 2.3.2.

## Open Research Badges

This article has earned Open Data and Open Materials badges. Data and materials are available at https://github.com/carina-kauf/lm-event-knowledge.

## 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).

Aissen, J. (2003). Differential object marking: Iconicity vs. economy. *Natural Language & Linguistic Theory*, *21*(3), 435–483.

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264.

Atari, M., Xue, M. J., Park, P. S., Blasi, D., & Henrich, J. (2023). *Which humans?* https://doi.org/10.31234/osf.io/5b26t

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv Preprint arXiv:1406.5823*.

Belinkov, Y., Màrquez, L., Sajjad, H., Durrani, N., Dalvi, F., & Glass, J. (2017). Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1–10).

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198).

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.

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A., & others. (2020). Experience grounds language. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 8718–8735).

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences*, *26*(12), 1153–1170.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Bossong, G. (1991). Differential object marking in Romance and beyond. In D. Wanner & D. Kibbee (Eds.), *New analyses in Romance linguistics* (pp. 143–170).

Broadbent, D. E. (1967). Word-frequency effect and response bias. *Psychological Review*, *74*(1), 1.

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

Cai, Z. G., Haslett, D. A., Duan, X., Wang, S., & Pickering, M. J. (2023). Does ChatGPT resemble humans in language use? *arXiv Preprint arXiv:2303.08014*.

Caramazza, A., Berndt, R. S., & Brownell, H. H. (1982). The semantic deficit hypothesis: Perceptual parsing and object classification by aphasic patients. *Brain and Language*, *15*(1), 161–189.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., & Zhang, C. (2022). Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U. & others. (2021). Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, 2633–2650.

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*(1), 1–10.

Chersoni, E., Santus, E., Pannitto, L., Lenci, A., Blache, P., & Huang, C.-R. (2019). A structured distributional model of sentence meaning and processing. *Natural Language Engineering*, *25*(4), 483–502.

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

Contreras Kallens, P., Kristensen-McLachlan, R. D., & Christiansen, M. H. (2023). Large language models demonstrate the potential of statistical learning in language. *Cognitive Science*, *47*(3), e13256.

Dasgupta, I., Lampinen, A. K., Chan, S. C., Creswell, A., Kumaran, D., McClelland, J. L., & Hill, F. (2022). Language models show human-like content effects on reasoning. *arXiv Preprint arXiv:2207.07051*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. https://doi.org/10.48550/arXiv.1810.04805

Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS One*, *10*(4), e0121945.

Dove, G. (2020). More than a scaffold: Language is a neuroenhancement. *Cognitive Neuropsychology*, *37*(5–6), 288–311.

Dove, G. O. (2023). Rethinking the role of language in embodied cognition. *Philosophical Transactions of the Royal Society B*, *378*(1870), 20210375.

Dowty, D. R. (1989). On the semantic content of the notion of 'thematic role'. In G. Chierchia, B. H. Partee & R. Turner (Eds.), *Properties, types and meaning: Volume II: Semantic issues* (pp. 69–129). Springer.

Elazar, Y., Kassner, N., Ravfogel, S., Ravichander, A., Hovy, E., Schütze, H., & Goldberg, Y. (2021a). Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, *9*, 1012–1031.

Elazar, Y., Zhang, H., Goldberg, Y., & Roth, D. (2021b). Back to square one: Artifact detection, training and commonsense disentanglement in the Winograd schema. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 10486–10500).

Elman, J. L., & McRae, K. (2019). A model of event knowledge. *Psychological Review*, *126*(2), 252.

Erk, K. (2007). A simple, similarity-based model for selectional preferences. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* pp. 216–223.

Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, *6*(10), 635–653.

Ettinger, A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, *8*, 34–48.

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, *41*(4), 469–495.

Fedorenko, E., Blank, I. A., Siegelman, M., & Mineroff, Z. (2020). Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition*, *203*, 104348.

Ferretti, T. R., McRae, K., & Hatherell, A. (2001). Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language*, *44*(4), 516–547.

Fillmore, C. J. (1967). *The case for case*.
Gauthier, J., Hu, J., Wilcox, E., Qian, P., & Levy, R. (2020). *SyntaxGym: An online platform for targeted evaluation of language models*.
Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*(20), 8051–8056.
Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, *23*(5), 389–407.
Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., & Cohen, A. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), 369–380.
Gong, C., He, D., Tan, X., Qin, T., Wang, L., & Liu, T.-Y. (2018). Frage: Frequency-agnostic word representation. *Advances in Neural Information Processing Systems*, *31*, 1341–1352.
Goodkind, A., & Bicknell, K. (2021). Local word statistics affect reading times independently of surprisal. *arXiv Preprint arXiv:2103.04469*.
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).
Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Speech acts* (pp. 41–58). Brill.
Günther, F., Nguyen, T., Chen, L., Dudschig, C., Kaup, B., & Glenberg, A. M. (2020). Immediate sensorimotor grounding of novel concepts learned from language alone. *Journal of Memory and Language*, *115*, 104172.
Haeuser, K. I., & Kray, J. (2022). How odd: Diverging effects of predictability and plausibility violations on sentence reading and word memory. *Applied Psycholinguistics*, *43*(5), 1193–1220.
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.
Hare, M., Jones, M., Thomson, C., Kelly, S., & McRae, K. (2009). Activating event knowledge. *Cognition*, *111*(2), 151–167.
Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665–695.
Hosseini, E. A., Schrimpf, M. A., Zhang, Y., Bowman, S., Zaslavsky, N., & Fedorenko, E. (2022). Artificial neural network language models align neurally and behaviorally with humans even after a developmentally realistic amount of training. https://doi.org/10.1101/2022.10.04.510681
Ivanova, A. A., Mineroff, Z., Zimmerer, V., Kanwisher, N., Varley, R., & Fedorenko, E. (2021). The language network is recruited but not required for nonverbal event semantics. *Neurobiology of Language*, *2*(2), 176–201.
Jackendoff, R. (1987). The status of thematic relations in linguistic theory. *Linguistic Inquiry*, *18*(3), 369–411.
Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1–38.
Jouravlev, O., Schwartz, R., Ayyash, D., Mineroff, Z., Gibson, E., & Fedorenko, E. (2019). Tracking colisteners' knowledge states during language comprehension. *Psychological Science*, *30*(1), 3–19.
Kassner, N., Dufter, P., & Schütze, H. (2021). Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 3250–3258).
Kassner, N., & Schütze, H. (2020). Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7811–7818).
Katz, J. J., & Fodor, J. A. (1963). The structure of a semantic theory. *Language*, *39*(2), 170–210.
Kauf, C., & Ivanova, A. A. (2023). A better way to do masked language model scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 925–935).
Kim, J. S., Elli, G. V., & Bedny, M. (2019). Knowledge of animal appearance among sighted and blind adults. *Proceedings of the National Academy of Sciences*, *116*(23), 11213–11222.
Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. *arXiv Preprint arXiv:2302.02083*.

Kumar, S., Sumers, T. R., Yamakoshi, T., Goldstein, A., Hasson, U., Norman, K. A., Griffiths, T. L., Hawkins, R. D., & Nastase, S. A. (2022). Reconstructing the cascade of language processing in the brain using the internal computations of a transformer-based language model. https://doi.org/10.1101/2022.06.08.495348

Kuperberg, G. R. (2021). Tea with milk? A hierarchical generative framework of sequential event comprehension. *Topics in Cognitive Science*, *13*(1), 256–298.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32–59.

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*, *62*, 621.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163.

Lambon Ralph, M. A., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, *18*(1), 42–55.

Lampinen, A. K. (2022). Can language models handle recursively nested grammatical structures? A case study on comparing models and humans. *arXiv Preprint arXiv:2210.15303*.

Landau, B., & Gleitman, L. R. (1985). *Language and experience: Evidence from the blind child*.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211.

Lenci, A. (2011). Composing and updating verb argument expectations: A distributional semantic model. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics* (pp. 58–66).

Lenci, A. (2023). Understanding natural language understanding systems. A critical analysis. *arXiv Preprint arXiv:2303.04229*.

Lenci, A., & Sahlgren, M. (2023). *Distributional semantics*. Cambridge University Press.

Levesque, H., Davis, E., & Morgenstern, L. (2012). The Winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago Press.

Levy, J. P., Bullinaria, J., & McCormick, S. (2017). Semantic vector evaluation and human performance on a new vocabulary MCQ test. In *Proceedings of the Annual Conference of the Cognitive Science Society: CogSci 2017 London: "Computational Foundations of Cognition"*.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Lewis, M., Zettersten, M., & Lupyan, G. (2019). Distributional semantics as a source of visual knowledge. *Proceedings of the National Academy of Sciences*, *116*(39), 19237–19238.

Liu, T., Zhang, Y., Brockett, C., Mao, Y., Sui, Z., Chen, W., & Dolan, B. (2022). A token-level reference-free hallucination detection benchmark for free-form text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 6723–6737).

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

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, *3*(2), 273–302.

Lucy, L., & Gauthier, J. (2017). Are distributional representations ready for the real world? Evaluating word vectors for grounded perceptual meaning. In *Proceedings of the 1st Workshop on Language Grounding for Robotics* (pp. 76–85).

Mahowald, K., Diachek, E., Gibson, E., Fedorenko, E., & Futrell, R. (2023). Grammatical cues to subjecthood are redundant in a majority of simple clauses across languages. *Cognition*, *241*, 105543.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). Dissociating language and thought in large language models: A cognitive perspective. *arXiv Preprint arXiv:2301.06627*.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78.

Mani, N., & Huettig, F. (2012). Prediction during language processing is a piece of cake—But only for skilled producers. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(4), 843.

Marcus, G. (2020). The next decade in AI: Four steps towards robust artificial intelligence. *arXiv Preprint arXiv:2002.06177*.

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330.

Marmor, G. S. (1978). Age at onset of blindness and the development of the semantics of color names. *Journal of Experimental Child Psychology*, *25*(2), 267–278.

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*, *37*(4), 913.

McCoy, T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3428–3448).

McRae, K., Hare, M., Elman, J. L., & Ferretti, T. (2005). A basis for generating expectancies for verbs from nouns. *Memory & Cognition*, *33*(7), 1174–1184.

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*, *3*(6), 1417–1429.

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*, *38*(3), 283–312.

Meteyard, L., & Vigliocco, G. (2008). The role of sensory and motor information in semantic representation: A review. In P. Calvo & T. Gomila (Eds.), *Handbook of cognitive science* (pp. 291–312).

Metheniti, E., Van de Cruys, T., & Hathout, N. (2020). How relevant are selectional preferences for transformer-based language models? In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 1266–1278).

Michaelov, J. A., Bardolph, M. D., Van Petten, C. K., Bergen, B. K., & Coulson, S. (2023). Strong prediction: Language model surprisal explains multiple N400 effects. *Neurobiology of language*, 1–71.

Michaelov, J. A., Coulson, S., & Bergen, B. K. (2022). *Do we need situation models? Distributional semantics can explain how peanuts fall in love* [Poster]. HSP 2022, UC San Diego (virtual).

Michelmann, S., Kumar, M., Norman, K. A., & Toneva, M. (2023). Large language models can segment narrative events similarly to humans. *arXiv Preprint arXiv:2301.10297*.

Nieuwland, M. S., & Van Berkum, J. J. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience*, *18*(7), 1098–1111.

Niven, T., & Kao, H.-Y. (2019). Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4658–4664).

Oh, B.-D., Clark, C., & Schuler, W. (2022). Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, *5*, 777963.

Oh, B.-D., & Schuler, W. (2023). Transformer-based LM surprisal predicts human reading times best with about two billion training tokens. *arXiv Preprint arXiv:2304.11389*.

Ostarek, M., Van Paridon, J., & Montero-Melis, G. (2019). Sighted people's language is not helpful for blind individuals' acquisition of typical animal colors. *Proceedings of the National Academy of Sciences*, *116*(44), 21972–21973.

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*, *67*(4), 426–448.

Padó, S., Padó, U., & Erk, K. (2007). Flexible, corpus-based modelling of human plausibility judgements. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (pp. 400–409).

Padó, U., Crocker, M., & Keller, F. (2006). Modelling semantic role plausibility in human sentence processing. In *11th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 345–352).

Paik, C., Aroca-Ouellette, S., Roncone, A., & Kann, K. (2021). The World of an Octopus: How reporting bias influences a language model's perception of color. *arXiv Preprint arXiv:2110.08182*.

Papadimitriou, I., Futrell, R., & Mahowald, K. (2022). When classifying arguments, BERT doesn't care about word order… except when it matters. *Proceedings of the Society for Computation in Linguistics*, *5*(1), 203–205.

Patel, R., & Pavlick, E. (2021). Mapping language models to grounded conceptual spaces. In *International Conference on Learning Representations*.

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, *8*(12), 976–987.

Pedinotti, P., Rambelli, G., Chersoni, E., Santus, E., Lenci, A., & Blache, P. (2021). Did the cat drink the coffee? Challenging transformers with generalized event knowledge. In *Proceedings of* SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. & others. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., & Miller, A. (2019). Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 2463–2473).

Piantadosi, S. (2023). Modern language models refute Chomsky's approach to language. *Lingbuzz Preprint*.

Porada, I., Suleman, K., Trischler, A., & Cheung, J. C. K. (2021). Modeling event plausibility with consistent conceptual abstraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1732–1743).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., & Clark, J. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* (pp. 8748–8763).

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

Raghunathan, T. E., Rosenthal, R., & Rubin, D. B. (1996). Comparing correlated but nonoverlapping correlations. *Psychological Methods*, *1*(2), 178.

Rambelli, G., Chersoni, E., Lenci, A., Blache, P., & Huang, C.-R. (2020). Comparing probabilistic, distributional and transformer-based models on logical metonymy interpretation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (AACL-IJCNLP)*.

Ravichander, A., Hovy, E., Suleman, K., Trischler, A., & Cheung, J. C. K. (2020). On the systematicity of probing contextualized word representations: The case of hypernymy in BERT. In *Proceedings of the 9th Joint Conference on Lexical and Computational Semantics* (pp. 88–102).

Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, *14*(3), 191–201.

Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020). Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4902–4912).

Roads, B. D., & Love, B. C. (2020). Learning as the unsupervised alignment of conceptual systems. *Nature Machine Intelligence*, *2*(1), 76–82.

Roberts, A., Raffel, C., & Shazeer, N. (2020). How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 5418–5426).

Roemmele, M., Bejan, C. A., & Gordon, A. S. (2011). Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning* (pp. 90–95).

Sakaguchi, K., Bras, R. L., Bhagavatula, C., & Choi, Y. (2021). Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, *64*(9), 99–106.

Salazar, J., Liang, D., Nguyen, T. Q., & Kirchhoff, K. (2020). Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 2699–2712).

Santus, E., Chersoni, E., Lenci, A., & Blache, P. (2017). Measuring thematic fit with distributional feature overlap. In *2017 Conference on Empirical Methods in Natural Language Processing* (pp. 648–658).

Sayeed, A., Shkadzko, P., & Demberg, V. (2015). An exploration of semantic features in an unsupervised thematic fit evaluation framework. *Italian Journal of Computational Linguistics*, *1*(1–1), 31–46.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), 45.

Shain, C., Blank, I. A., van Schijndel, M., Schuler, W., & Fedorenko, E. (2020). fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, *138*, 107307.

Shain, C., Meister, C., Pimentel, T., Cotterell, R., & Levy, R. P. (2022). *Large-scale evidence for logarithmic effects of word predictability on reading time*.

Shapira, N., Levy, M., Alavi, S. H., Zhou, X., Choi, Y., Goldberg, Y., Sap, M., & Shwartz, V. (2023). Clever hans or neural theory of mind? Stress testing social reasoning in large language models. *arXiv Preprint arXiv:2305.14763*.

She, J. S., Potts, C., Bowman, S. R., & Geiger, A. (2023). ScoNe: Benchmarking negation reasoning in language models with fine-tuning and in-context learning. *arXiv Preprint arXiv:2305.19426*.

Shwartz, V., & Choi, Y. (2020). Do neural language models overcome reporting bias? In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 6863–6870).

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*, *20*(11), 2037–2057.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 3.

Sorscher, B., Ganguli, S., & Sompolinsky, H. (2021). The geometry of concept learning. https://doi.org/10.1101/2021.03.21.436284

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, *10*(1), 89–96.

Szewczyk, J. M., & Federmeier, K. D. (2022). Context-based facilitation of semantic access follows both logarithmic and linear functions of stimulus probability. *Journal of Memory and Language*, *123*, 104311.

Talmor, A., Elazar, Y., Goldberg, Y., & Berant, J. (2020). oLMpics—On what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, *8*, 743–758.

Tamborrino, A., Pellicanò, N., Pannier, B., Voitot, P., & Naudin, L. (2020). Pre-training is (almost) all you need: An application to commonsense reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 3878–3887).

Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4593–4601).

The MosaicML NLP Team. (2023). *MPT-30B: Raising the bar for open-source foundation models*. Retrieved from https://www.mosaicml.com/blog/mpt-30b

Thrush, T., Wilcox, E., & Levy, R. (2020). Investigating novel verb learning in BERT: Selectional preference classes and alternation-based syntactic generalization. In *Proceedings of the 3rd BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP* (pp. 265–275).

Traxler, M. J., Morris, R. K., & Seely, R. E. (2002). Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language*, *47*(1), 69–90.

Trott, S., Jones, C., Chang, T., Michaelov, J., & Bergen, B. (2023). Do Large Language Models know what humans know? *Cognitive Science*, *47*(7), e13309.

Tuckute, G., Sathe, A., Srikant, S., Taliaferro, M., Wang, M., Schrimpf, M., Kay, K., & Fedorenko, E. (2023). Driving and suppressing the human language network using large language models. https://doi.org/10.1101/2023.04.16.537080

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

Vassallo, P., Chersoni, E., Santus, E., Lenci, A., & Blache, P. (2018). Event knowledge in sentence processing: A new dataset for the evaluation of argument typicality. *LREC 2018 Workshop on Linguistic and Neurocognitive Resources (LiNCR)*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998–6008.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., & Shieber, S. (2020). Investigating gender bias in language models using causal mediation analysis. *Advances in Neural Information Processing Systems*, *33*, 12388–12401.

Wang, A., & Cho, K. (2019). BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation* (pp. 30–36).

Wang, B., & Komatsuzaki, A. (2021). *GPT-J-6B: A 6 billion parameter autoregressive language model*. Retrieved from https://github.com/kingoflolz/mesh-transformer-jax

Wang, S., Durrett, G., & Erk, K. (2018). Modeling semantic plausibility by injecting world knowledge. In *Proceedings of NAACL-HLT* (pp. 303–308).

Wang, X., Men, W., Gao, J., Caramazza, A., & Bi, Y. (2020). Two forms of knowledge representations in the human brain. *Neuron*, *107*(2), 383–393.

Wang, Z., Jafarpour, A., & Sap, M. (2022). Uncovering surprising event boundaries in narratives. In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*.

Warren, T., & McConnell, K. (2007). Investigating effects of selectional restriction violations and plausibility violation severity on eye-movements in reading. *Psychonomic Bulletin & Review*, *14*(4), 770–775.

Warren, T., McConnell, K., & Rayner, K. (2008). Effects of context on eye movements when reading about possible and impossible events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(4), 1001.

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*, *30*(8), 932–939.

Weir, N., Poliak, A., & Van Durme, B. (2020). Probing neural language models for human tacit assumptions. In *42nd Annual Virtual Meeting of the Cognitive Science Society (CogSci)*.

Willits, J. A., Amato, M. S., & MacDonald, M. C. (2015). Language knowledge and event knowledge in language use. *Cognitive Psychology*, *78*, 1–27.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M. & others. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45).

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, *32*, 5753–5763.

Zacks, J. M. (2020). Event perception and memory. *Annual Review of Psychology*, *71*, 165–191.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind-brain perspective. *Psychological Bulletin*, *133*(2), 273.

Zellers, R., Bisk, Y., Schwartz, R., & Choi, Y. (2018). SWAG: A large-scale adversarial dataset for grounded commonsense inference. *EMNLP*.

Zhang, C., Van Durme, B., Li, Z., & Stengel-Eskin, E. (2022). Visual commonsense in pretrained unimodal and multimodal models. *arXiv Preprint arXiv:2205.01850*.

Zhu, X., Li, T., & De Melo, G. (2018). Exploring semantic properties of sentence embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 632–637).

Zmigrod, R., Mielke, S. J., Wallach, H., & Cotterell, R. (2019). Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1651–1661).

## Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supplemental Information