American Psychological Association

Psychological Review

ISSN: 0033-295X

https://doi.org/10.1037/rev0000319

Transformer Networks of Human Conceptual Knowledge

Sudeep Bhatia and Russell Richie
Department of Psychology, University of Pennsylvania

We present a computational model capable of simulating aspects of human knowledge for thousands of real-world concepts. Our approach involves a pretrained transformer network that is further fine-tuned on large data sets of participant-generated feature norms. We show that such a model can successfully extrapolate from its training data, and predict human knowledge for new concepts and features. We apply our model to stimuli from 25 previous experiments in semantic cognition research and show that it reproduces many findings on semantic verification, concept typicality, feature distribution, and semantic similarity. We also compare our model against several variants, and by doing so, establish the model properties that are necessary for good prediction. The success of our approach shows how a combination of language data and (laboratory-based) psychological data can be used to build models with rich world knowledge. Such models can be used in the service of new psychological applications, such as the modeling of naturalistic semantic verification and knowledge retrieval, as well as the modeling of real-world categorization, decision-making, and reasoning.

Keywords: conceptual knowledge, semantic cognition, distributional semantics, connectionist modeling, transformer networks

Knowledge of concepts and their features is one of the fundamental topics of inquiry in cognitive science (Murphy, 2004; Rips et al., 2012). Understanding the content, structure, and representational format of conceptual knowledge is of great intrinsic interest but is also critical as conceptual knowledge is a central component of much of high-level cognition. Unfortunately, it has traditionally been difficult to model what people actually know about real-world concepts, owing to their complexity and richness of knowledge acquired over many years of learning. Thus, even though there are many influential computational theories of conceptual knowledge (Collins & Loftus, 1975; Rogers & McClelland, 2004; Smith et al., 1974), most such theories cannot make a priori predictions about how participants will judge the truth or falsehood of a proposition involving an arbitrary concept–feature pair. This is also why quantitative cognitive process models (Busemeyer & Diederich, 2010; Lewandowsky & Farrell, 2010) common in other areas of cognitive psychology are seldom applied to naturalistic semantic cognition. For example, evidence accumulation models have had great success in capturing response time distributions in perceptual and preferential choice (Brown & Heathcote, 2008; Ratcliff, 1978; Usher & McClelland, 2001), but do not extend simplistically to semantic verification tasks involving everyday concepts and features, such as those in Collins and Quillian (1969). Likewise, computational memory models accurately predict word list recall data (Polyn et al., 2009; Raaijmakers & Shiffrin, 1981) but are not used to model sequences of features recalled by participants in feature norm studies, such as those in McRae et al. (2005).

Our inability to predict knowledge of real-world concepts also makes it difficult for us to build models that use human knowledge of naturally occurring entities in the service of high-level cognition. There are many successful theories of categorization, judgment, decision-making, induction, and reasoning, but these are typically tested using only abstracted stimuli involving small sets of experimenter-defined features and relations. In order to quantitatively predict semantic verification, response time, and recall probability in everyday semantic cognition, and to build models capable of naturalistic high-level cognition, we first need a way to proxy human knowledge for thousands of concepts and an almost infinite set of possible features. This is not currently possible.

Fortunately, the field of distributional semantics has made great progress in automatically extracting knowledge representations for real-world concepts from large-scale natural language data. These advances rely on the insight that much human knowledge is reflected in the patterns of co-occurrence between words, which can be exploited to derive representations for words as vectors in high-dimensional semantic spaces or in terms of a set of probabilistic semantic topics (Griffiths et al., 2007; Jones & Mewhort, 2007; Landauer & Dumais, 1997; Mikolov et al., 2013; Pennington et al., 2014; for reviews, see Bhatia et al., 2019; Lenci, 2018; or Günther et al., 2019). However, even though distributional semantic representations predict the semantic relatedness of concepts with reasonable accuracy and can be applied to tens of thousands of common concepts, they do not possess complex featural and relational representations for concepts. In response, researchers have begun to combine distributional semantic representations with survey and experimental data, such as feature norms and participant ratings (Andrews et al., 2009; Derby et al., 2019; Lu et al., 2019; Richie et al., 2019). In most such applications, "pretrained" distributional semantic representations for concepts serve as inputs into more

Sudeep Bhatia https://orcid.org/0000-0001-6068-684X
Russell Richie https://orcid.org/0000-0002-9686-8407

Funding was received from the National Science Foundation grant SES-1847794. Data and code for our model is available at https://osf.io/fr4cz/.

Correspondence concerning this article should be addressed to Sudeep Bhatia, Department of Psychology, University of Pennsylvania, D22 Solomon Labs, 3720 Walnut St. Philadelphia, PA 19104, United States. Email: bhatiasu@sas.upenn.edu

complex models that are "fine-tuned" on participant data and are thus able to proxy the (often structured) knowledge at play in the participants' responses. Such extensions of the distributional semantics framework are promising. However, they seldom predict human knowledge for out-of-sample features (features on which the models have not been fine-tuned) and are thus unable to capture the richness of human knowledge for naturalistic concepts.

In this article, we extend the distributional semantics approach using contemporary implementations of classical connectionist models of semantic cognition (Hinton, 1981; Rogers & McClelland, 2004). We rely, in particular, on transformer networks (Devlin et al., 2018; Vaswani et al., 2017), which represent sentences using high-dimensional vectors and manipulate these vectors in multiple interconnected neural network layers based on an attention mechanism. These networks are pretrained on billions of words of natural language and have rich representations for thousands of words, as well as the capacity to process aspects of linguistic structure (Linzen & Baroni, 2020; Manning et al., 2020; McClelland et al., 2020). We fine-tune transformer networks on sentences generated from existing feature and category norm studies in psychology (Devereux et al., 2014; McRae et al., 2005; Van Overschelde et al., 2004), and in doing so, teach them to use their rich world knowledge in a manner that proxies the representations and processes at play in simple semantic judgment.

The power and generality of transformer networks mean that they can be used to make predictions for a very large set of naturalistic concepts and features, including concepts and features for which human data are not available. We use cross-validation analysis to test the accuracy of such predictions—our models are trained on a subset of category and feature norms data and then tested on out-of-sample concepts, out-of-sample features, out-of-sample concept–feature combinations, and even out-of-sample domains. We also evaluate our models in terms of their ability to reproduce 17 distinct behavioral patterns from 25 experiments involving response times in semantic verification tasks, typicality ratings for concepts and categories, the distribution of features across concepts, and similarity ratings between pairs of concepts. We do this with stimuli from classic as well as more recently published articles. We are the first to make quantitative predictions for most of these stimuli sets, and the success or failure of our tests has significant implications for the psychological validity of transformer networks. If these networks succeed at predicting participant-generated feature norms, as well as classical effects in semantic cognition, we can consider them to be valuable tools for studying human conceptual knowledge. They can be used to understand the computational mechanisms that give rise to observed behavioral patterns, and the predictions of these networks can be used in conjunction with existing theories for more sophisticated models of semantic verification and naturalistic high-level cognition.

Theoretical Background

Connectionist Models

The work in this article draws from two traditions in semantic cognition research. The first is connectionist modeling (McClelland & Rumelhart, 1986; Rumelhart & McClelland, 1986), which proposes that "semantic abilities arise from the flow of activation among simple, neuron-like processing units, as governed by the strengths of interconnecting weights" (Rogers & McClelland, 2004, p. 689). This approach has had a long history in cognitive science, with the earliest model in this theoretical framework being Hinton (1981, 1986) distributed network for propositional knowledge. Hinton's model assumes separate vector representations for each part of a proposition. For simple propositions involving a head, a tail, and a relation word (e.g., *cat has fur*), this model involves three base layers, each corresponding to one word in the proposition (e.g., *cat*, *has*, and *fur*). Each of the word layers in Hinton's model are recurrently connected to a proposition layer, as well as to themselves. Thus, feeding the words in a proposition as inputs into the model leads to changing activation patterns on both the proposition layer and the word layers, culminating in a stable state in which activation patterns in the proposition layer reflect the emergent meaning of the proposition. See Figure 1A, for an illustration.

Hinton's model has several properties that make it a useful model of semantic cognition, including the ability to use the proposition layer as an input into additional cognitive processes, the ability to store a large set of propositions in memory using a small set of weights, the ability to generalize knowledge to novel concepts and features, and the ability to encode sentence structure by allowing for different weights between each pair of layers (see Rogers & McClelland, 2004, for a summary). As we discuss below, Hinton's model has architectural similarities to the transformer networks used in the current article, which also involve vector representations for the individual words in a proposition, as well as the interactive transformation of these representations in hidden layers to generate emergent vector representations for propositions.

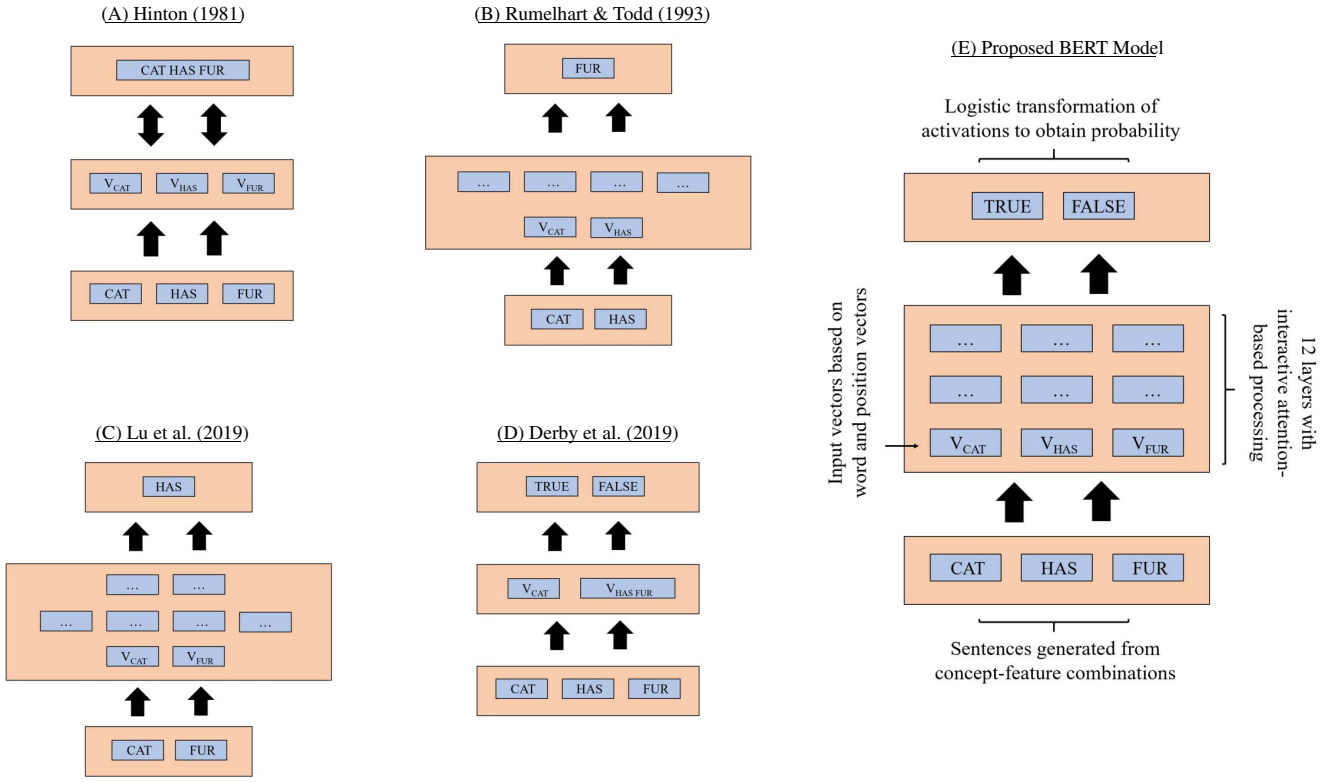
Since Hinton's foundational work, many other researchers have developed computational connectionist models capable of semantic representations and judgment. Perhaps the most popular of these is Rumelhart and Todd (1993) network. This network has a feed-forward structure that transforms vector inputs of entities and relations (e.g., *cat* and *has*), using hidden layers, to produce predictions for words and features that match the inputs (e.g., *fur*). See Figure 1B, for an illustration. Rogers and McClelland (2004) have trained this model on a taxonomic structure adapted from Collins and Quillian (1969) to study concept organization and explain how concept organization changes over the course of development, as well as with expertise and dementia (see also Saxe et al., 2019, for further analytical results). There are of course many other connectionist models and hybrid symbolic-connectionist models of semantics (as well as connectionist networks of distributional semantics discussed below). These models typically involve distributed (vector) representations for the words and concepts that make up propositions, and the transformation of these representations in neural network layers. The richness of this work reflects the importance of the connectionist paradigm to the study of conceptual knowledge. Connectionist models do have limitations, which we will discuss in detail at the end of this article. However, they also provide a powerful approach for understanding the dynamics of concept learning, as well as semantic reasoning and judgment, within a well-specified (and in some cases, neurobiologically feasible) computational framework.

Distributional Semantics

Our article also builds off a second area of research in semantic cognition: Distributional semantics. Models within this tradition use

Figure 1

Previous Neural Network and Distributed Semantics Models of Conceptual Knowledge (A–D) as Well as Proposed BERT Model (E)



Note. Boxes in the above diagram correspond to sets of nodes and layers, and V describes a vectorized representation of the word in the subscript. As with prior models, our proposed model represents concept and feature inputs using vector representations, which are transformed and manipulated through hidden layers. BERT = Bidirectional encoder representations from transformers. See the online article for the color version of this figure.

ONLINE COLOR

word distribution statistics in large-scale natural language data to derive representations for words that preserve word meaning. For example, the classic method, latent semantic analysis (LSA), performs a singular value decomposition on word–context co-occurrence matrices, which assigns each word in the corpus a vector representation in a high-dimensional semantic space (Landauer & Dumais, 1997). As words that occur in similar contexts in natural language have similar vectors in this space, vector similarity measures like cosine are a good proxy for word association and word relatedness.

Since Landauer and Dumais (1997) introduced LSA, a host of models exploiting lexical distribution statistics in the text have emerged. One such model is GloVe (Pennington et al., 2014), which tabulates a word-word co-occurrence matrix and then learns vectors such that the dot product of two-word vectors approximates the logarithm of the words' probability of co-occurrence. Below we will be using vectors derived from a GloVe model as a baseline for model comparison. Other related models include BEAGLE (Jones & Mewhort, 2007), topic models (Griffiths et al., 2007), and Word2Vec (Mikolov et al., 2013).

Word vector representations learned through these methods have been enormously successful in psychological applications (for reviews, see Bhatia et al., 2019; Günther et al., 2019; Jones et al., 2015; Lenci, 2018; Mandera et al., 2017). For example, cosine similarity between word vectors correlates with Likert-scale judgments of words' similarity and relatedness (Hill et al., 2015; Richie & Bhatia, in press), strength of semantic priming in, for example, the lexical decision task, as measured by reaction times (Jones et al., 2006; Mandera et al., 2017), and even with probability of recall given a cue in free association, or given an earlier recalled item in list recall (although there are often better ways to use word vectors for such tasks rather than simply computing cosine; see Nematzadeh et al., 2017 and Jones et al., 2018). Semantic judgments about words (e.g., the size of an animal) can also be approximated by calculating the relative vector similarity of a judgment target (e.g., *tiger*) to words high (e.g., *large*, *big*) and low (e.g., *small*, *tiny*) on a judgment dimension (Grand et al., 2018). Most relevant to this article, Bhatia (2017) has found that cosine similarity between word vectors provides a good measure of association in simple judgment tasks, and can thus predict association-based biases in probability judgment, factual judgment, and forecasting (Evans, 2008; Kahneman, 2011; Sloman, 1996). Finally, on the basis of these and related findings, some scholars have suggested that word distribution statistics also play a critical role in human semantic development (see Unger & Fisher, 2021, for a recent review).

Fine-Tuning Word Representations

Although the distributional semantics approach has been very successful at quantifying degree of similarity, relatedness, or

association among pairs of words, or modeling simple semantic judgments of single words, it is, in its simplest form, unable to capture the types of word relationships that are necessary for structured knowledge representation (Lake & Murphy, in press; Lenci, 2018; Lu et al., 2019). For example, these models may be able to predict that *cat* and *fur* are closely associated (e.g., by observing that semantic vector representations or topic distributions for these words are very similar), but these models cannot evaluate whether *cats have fur*, *cats are a type of fur*, or *fur has cats*.

In response, recent work uses representations derived from the above approaches in more sophisticated models that are fine-tuned on task-specific human data. For example, the Bayesian analogy with relational transformations (BART) model (Lu et al., 2012, 2019) transforms Word2Vec representations for the entities in a proposition (e.g., *cat* and *fur*) using subtraction and ranking operations to obtain vector representations for the entity pair. These representations are then mapped onto a continuous variable that predicts whether a given relation (e.g., *has*) holds between the entities. The mapping itself is trained on positive and negative examples of different relations obtained from human subjects. The resulting model makes accurate predictions for the relations between novel entity pairs and is subsequently able to model analogy and relationship similarity judgments between pairs of words. See Figure 1C, for an illustration.

The partial least squares regression (PLSR; Făgărăşan et al., 2015) and Feature2Vec (Derby et al., 2019) models provide another example of this approach. Feature2Vec takes GloVe representations of concepts as inputs, and using subject-generated feature norms for concepts, attempts to derive representations for features as points in the original GloVe vector space. The trained model evaluates whether a proposition (e.g., *cat has fur*) is true by measuring the vector similarity of its initial representation for the concept (e.g., *cat*) and its learnt representation for the feature (e.g., *has fur*; see Figure 1D, for an illustration). The PSLR model is similarly fine-tuned on feature norms but learns concept–feature relationships using regression. Note that there are also closely related approaches that use feature norms data to augment word representations extracted from the text (e.g., Andrews et al., 2009; Steyvers, 2010). These approaches rely on the insight that participant-generated feature norms reflect experiences with the world and capture sensory–motor data not present in spoken and written language, resulting in higher quality word representations.

Finally, in our previous work (Richie et al., 2019; also see Bhatia, 2019; Bhatia et al., in press; Zou & Bhatia, 2021), we have used fine-tuning to predict human ratings of concepts on psychological properties (e.g., the femininity of traits, the warmth and competence of people, the riskiness of activities, and the nutrition of foods, etc.). We take Word2Vec and GloVe representations of concepts as inputs, and, using human ratings of concepts, train a regularized regression model that maps these vector representations onto a scalar prediction for the human rating. Mathematically, this regression technique resembles PSLR and Feature2Vec and can be seen as deriving representations of the rating dimension (e.g., femininity) as a point in the semantic space. Our regression technique is able to accurately predict human ratings of the psychological properties of concepts on which the model has not been fine-tuned. Importantly, such predictions are more accurate than those obtained by the relative similarity of the judgment target to words high and low in the judgment dimension, that is, the Grand et al. (2018) approach discussed above.

Fine-tuning methods have been shown to be successful at describing certain types of structured knowledge. That said, they are still severely limited by the human data that are used to train the models. For example, PSLR and Feature2Vec are unable to make predictions for novel features that are not present in their training data. Likewise, our own regression approaches are applicable only to the psychological dimensions for which we elicit human ratings, and cannot extrapolate to novel psychological dimensions. The approaches discussed here are also limited by the vocabularies of the semantic vector models (e.g., GloVe) that they take as inputs. These semantic vector models typically do not have representations for noun phrases (e.g., *house cat*, *black fur*), and for this reason, cannot be applied to more complex concepts and features known to humans, including concepts and features that have been used in earlier research on semantic cognition.

Transformer Networks

Recent years have seen technological breakthroughs in computational linguistics: Computer models are now able to achieve unprecedented levels of performance in question answering, semantic entailment, machine translation, sentiment analysis, and other natural language processing tasks. Perhaps the most impressive advances have come from a new type of deep, feed-forward neural network known as the transformer (Brown et al., 2020; Cer et al., 2018; Dai et al., 2019; Devlin et al., 2018; Radford et al., 2018; Vaswani et al., 2017). There are now numerous variants of transformers, but we will focus on a particularly prominent transformer, the bidirectional encoder representations from transformers (BERT) model (Devlin et al., 2018; Rogers et al., 2020).

BERT is a stack of layers. The first layer receives vector representations of each of the individual words that make up the sentence and transforms these representations into a vector output. This output is fed into the next layer, whose output is fed into the subsequent layer, and so on. Each layer transforms its input using a feed-forward neural network and a self-attention mechanism. Self-attention enables the model to look at other positions in the input sentence for information about how to best process a word. This makes it sensitive to word order and allows it to solve problems related to word-sense disambiguation (e.g., knowing that *fur* in *cats have fur* refers to the hairy coat on their skin rather than the garment) and coreference resolution (e.g., knowing that *cats* and *their* refers to the same entity in *cats like to lick their fur*). Recent work has also shown that this mechanism allows BERT to process certain types of linguistic structure, including syntactic dependencies (Jawahar et al., 2019; Linzen & Baroni, 2020; Manning et al., 2020; McClelland et al., 2020; Tenney et al., 2019).

BERT is "pretrained" on tasks such as masked word prediction and next sentence prediction on large amounts of text data. The pretrained BERT model can be subsequently fine-tuned on additional data (e.g., pairs of questions and answers, or sentiment ratings for sentences) to learn how to apply its learnt representations to specific natural language processing tasks. This "pretrain, then fine-tune" paradigm, known as "transfer learning," is the key to BERT's state-of-the-art achievements in natural language processing.

Building off these successes, our proposal is to fine-tune BERT to perform semantic verification (true/false judgments) for sentences

generated from feature and category production norms. In our tests, BERT learns to predict if a sentence is composed of a concept and feature pair generated by subjects (i.e., is true) or if it is composed of a concept and feature pair not generated by subjects (i.e., is false). Pretraining equips BERT with world knowledge, as well as the ability to process simple linguistic structure. By fine-tuning BERT on feature and category norms data obtained from psychological studies, we teach it facts about the world, and thus give it the ability to generalize its semantic and linguistic knowledge to evaluate new facts. In this way, BERT develops human-like semantic capabilities, at least for the types of propositions on which it is trained and tested. To better understand how the various properties of the BERT model contribute to its ability to mimic human conceptual knowledge, we also compare this model with variants that are not fine-tuned on subject-generated norms, variants that are not pretrained on large amounts of text data, as well as with variants that use alternate BERT outputs to make predictions. See Figure 1E, for an illustration of the BERT model.

We are not the first to apply BERT to structured semantic cognition in this manner. Some recent work has used a similar approach to augment knowledge bases with graphs consisting of {head, relation, tail} triples (e.g., {*Steve Jobs*, *founded*, *Apple*}; e.g., Bosselut et al., 2019; Bouraoui et al., 2020; Petroni et al., 2019; Yao et al., 2019. However, transformer networks have been underutilized for psychological applications of interest. For example, our fine-tuned BERT model generates continuous predictions of truth and falseness for natural language sentences generated from concept and feature pairs. We suggest that it may be possible to use these continuous predictions to predict classic findings in semantic verification tasks including choice probability, response time, and typicality effects (e.g., Collins & Quillian, 1969; Rosch, 1975). Since BERT can specify the relationship between thousands of concepts and features, it can also be used to predict how correlated different features are with each other (e.g., Malt & Smith, 1984; McRae et al., 1997). BERT's sensitivity to word ordering in sentences may also allow it to capture asymmetry effects in similarity judgment, thereby outperforming classical distributional semantics models (whose spatial structure imposes strong symmetry assumptions on similarity, at least when using symmetric measures like cosine or dot product; e.g., Tversky, 1977; Whitten et al., 1979).

Transformer networks that produce representations for sentences are natural successors to earlier distributional semantics models that produce vectors (or topic distributions) for words. Both are based on word distribution statistics in natural language, and both generate representations for linguistic units (words or sentences) that preserve semantic similarity. Likewise, the "pretrain, then fine-tune" paradigm is identical to the BART, Feature2Vec, and regression applications discussed in the preceding section, in that vector representations for linguistic units are transformed in order to better capture the task to which the models are applied. Indeed, our use of feature norms data to fine-tune BERT is inspired by the successes of PSLR and Feature2Vec, which have already demonstrated the utility of these data for modeling feature-based knowledge representations in distributional semantics.

As outlined above, transformer networks can also be seen as modern counterparts to earlier connectionist models of semantic cognition. In our case, BERT's architecture is related to that of Hinton (1981) connectionist network. Both take as inputs vector representations for the individual words that make up the sentence, and both transform these inputs into more refined representations using neural network operations. Of course, Hinton's network implements these operations in recurrent layers, whereas BERT implements them in large numbers of feed-forward layers. However, it is well known that series of operations in feed-forward layers can approximate the outputs of recurrent networks. Indeed, BERT's transformations of concatenated word vectors in successive layers (each of which gives a vector output of the same dimensionality as the input) is similar to how recurrent networks successively manipulate a set of vectors using a single set of operations. For this reason, our BERT model and Hinton's model both have similar properties regarding prediction, generalization, and representational parsimony (see Rogers & McClelland, 2004, for a discussion). Of course, classic networks relied on handcrafted data sets of only a handful of concepts and their properties and relations. The promise of BERT is that, having been pretrained on large language data sets and fine-tuned on large (laboratory-based) psychological data sets, it can capture a much more realistic range of human knowledge.

Implementational Details

Norms Data Sets

We fine-tuned our models using three existing data sets of subject-generated norms. The first data set comes from Van Overschelde et al. (2004), who collected category norms for 70 different categories. Participants were asked to generate as many concepts in the categories as possible, and concepts were ranked based on their production frequencies. The data set offered by Van Overschelde et al. has a total of 1,621 distinct concepts. The second data set comes from McRae et al. (2005), who collected feature norms for 541 concepts. Participants in their study were asked to list the features of each of the concepts and were encouraged to list different types of features (e.g., perceptual and functional properties, concept categories, and encyclopedic facts). McRae et al. processed their data to exclude infrequently listed features and pool synonymous features (e.g., *has fur*, *is fluffy*). The data set released by McRae et al. has a total of 2,526 distinct experimenter-processed features. Our third data set comes from Devereux et al. (2014), who collected feature norms for 638 concepts. Participants were given similar instructions as in McRae et al., and the final data set contains a total of 26,075 unique participant-generated features. Devereux et al. also pooled synonymous features, which yielded 5,926 semantically distinct experimenter-processed features. The concepts and categories used in the above data sets were taken from prior research on semantic cognition and included animals, foods, tools, clothes, geographic locations, occupations, and other real-world entities.

True Sentences

We trained our models using "true" and "false" sentences generated from these three norms data sets. True sentences were simply natural language propositions obtained by combining concepts and their corresponding categories or features. In the case of Van Overschelde et al.'s (2004) data, concepts (e.g., *cat*) were combined with their corresponding superordinate category (e.g., *a four-footed animal*) using an "is a" category membership relation, yielding natural language sentences of the form [CONCEPT] *is a* [SUPERODINATE CATEGORY] (e.g., *cat is a four-footed animal*). In contrast, concepts in McRae et al. (2005; e.g., *cat*) were simplistically combined with

their corresponding features (e.g., *has fur*), yielding natural language sentences of the form [CONCEPT] [FEATURE] (e.g., *cat has fur*; this was done as most features in McRae et al. were verb phrases that specified the relation with the target concept). We used a nearly identical approach for concepts in Devereux et al. (2014) except that we applied it to the raw subject-generated features rather than experimenter-processed features. This gave us separate sentences for semantically similar features (e.g., *cat has fur* and *cat is fluffy*). True sentences generated using the above methods were combined with true "tautology" sentences made using the Van Overschelde et al. concepts. This involved merely combining the target concept with "is a" in the formula [CONCEPT] *is a* [CONCEPT] (e.g., *cat is a cat*). We decided to artificially generate such sentences as they are within the theoretical scope of most theories of semantic cognition, but are not present in our feature norms data.

Note that we did have to manually process some of the McRae et al. features, such as those prefixed by "beh" and "inbeh," in order to generate reasonable natural language sentences (these prefixes were used by McRae et al. to organize and code the subject-generated features—"Beh" is an abbreviation of behavior and "Inbeh" is an abbreviation for something an inanimate object does seemingly on its own). We also excluded features prefixed by "e.g." as these involved category examples that could not be simplistically made into natural language sentences with our approach (i.e., sentences like *cat e.g., tabby* were removed). Additionally, we removed "does" from the Devereux et al., features in which "does" was followed by a verb, and also modified the verb with a 3rd person singular inflection (e.g., sentences like *cat does like milk* were replaced with *cat likes milk*). For both the McRae et al. and Devereux et al. data sets, we manually added "is" to the features where necessary (e.g., replacing *cat made of bones* with *cat is made of bones*).

Finally, our data set repeated true sentences in proportion to the production frequencies for the features that made up the sentences (i.e., the production frequencies of the concept–feature pairs). For example, 13 participants in the Devereux et al. data set generated the feature *has fur* for the concept *cat*. Thus, the true sentence *cat has fur* was repeated 13 times in our final data set. In the case of Van Overschelde et al., who obtained hundreds of repetitions for concepts, we scaled the repetitions to be between 1 and 10. For example, 97% of participants gave *cat* as concept in response to the category *a four-footed animal*. This number was divided by 10 and rounded, to generate 10 repetitions for *cat is a four-footed animal* in the final data set. In contrast, only 23% of participants gave *mouse* as a response to this category, yielding only two repetitions for *mouse is a four-footed animal* in our data set. Each of the tautology sentences (e.g., *cat is a cat*) were repeated 10 times (corresponding to a hypothetical production frequency of 10) in our data set.

Overall, there were a total of 245,642 true sentences, including repetitions, in our final data set. These were composed of 58,385 unique sentences. 4,438 true sentences were generated using the Van Overschelde et al. data set, 75,365 were generated using the McRae et al. data set, 146,929 were generated using the Devereux et al. data set, and 18,910 were tautology sentences made up of concepts from the Van Overschelde et al. data set.

False Sentences

Our false sentences were obtained by modifying the true sentences using one of two methods. The first method, which we refer to as "concept replacement," involved replacing the target concept with a randomly selected concept from the same data set to which the feature had not been applied (e.g., the true sentence *cat has fur* could be used to generate the false sentence *tractor has fur*, by replacing *cat* with *tractor* and ensuring that *has fur* was not a feature listed in response to *tractor*). The second method, which we refer to "subject–object exchange" involved swapping the subjects and objects of the true sentence whose features contained only one noun (e.g., the true sentence *cat hunts mouse* could be used to generate the false sentence *mouse hunts cat*, by swapping the positions of *cat* and *mouse* and ensuring that *hunts mouse* has only a single noun). We did not apply the "subject–object exchange" method to sentences with a symmetric relation (e.g., *is similar to*, *is associated with*, *resembles*, *is close to*, *is mixed with*, etc.).

We generated one false sentence for each (potentially nonunique) true sentence in our training data, leading to 245,642 false sentences in our data set. This was done by first generating as many "subject–object exchange" sentences as possible, and then completing the false sentence data set with "concept replacement."

Note that our use of false sentences not directly generated by participants is a type of negative sampling. Negative sampling is common in classification tasks, in which items are randomly paired with incorrect labels in order to provide additional training data for the learner. Negative sampling is also used in models like Word2Vec, as well as in extensions of these models such as Feature2Vec (whose negative sampling method is identical to our "concept-replacement" method). As with most other negative sampling methods, the "concept-replacement" and the "subject–object exchange" algorithms used in this article can mistakenly assign the false label to otherwise true sentences. As we shall see below, many of the seemingly false sentences that our model thinks are true are in fact composed of reasonable concept–feature combinations that were not generated by subjects. Given the size and richness of the training data there is no way to avoid this problem, and we recognize that the accuracy rates in our tests are underestimates of the predictive power of our approach.

In summary, our final data set had a total of 491,284 observations (sentences and corresponding true or false labels). 245,642 of these observations involved true sentences, 33,587 involved false sentences generated using the "subject–object exchange" method, and 212,055 involved false sentences generated using the "concept-replacement" method. The 491,284 observations were composed of 284,530 unique sentences, 2,066 unique concepts, and 29,048 unique features. These statistics are summarized in Table 1. Note that to evaluate the sensitivity of our model's performance to the potentially nongrammatical and unusual false subject–object exchange sentences (e.g., *fur has cat*), we also tested a variant of this model that used only the concept-replacement sentences to generate negative examples.

Model Specification and Training

In the past year, many extensions of BERT have been proposed by researchers (e.g., RoBERTa, Liu et al., 2019 and SemBERT, Zhang et al., 2020) and the core BERT model itself has been pretrained on different types of corpora (e.g., scientific publications, Beltagy et al., 2019, or biomedical data, Lee et al., 2020). However, we used the original uncased BERT "base" model, released by Devlin et al. (2018). This model has 12 layers (i.e., transformer blocks and attention heads in each block) and a hidden layer size of

Table 1
Summary Statistics for Final Data Set

Type of sentence	*N*
True sentences (e.g., *cat has fur*)	245,642
False concept-replacement sentences (e.g., *tractor has fur*)	212,055
False subject–object exchange sentences (e.g., *fur has cat*)	33,587
Van Overschelde et al. (2004) sentences	46,696
McRae et al. (2005) sentences	150,730
Devereux et al. (2014) sentences	293,858
Unique sentences	284,530
Unique concepts	2,066
Unique features	29,048

768, resulting in 110M parameters. It was pretrained using masked word prediction and next sentence prediction on a books corpus (800M words) and Wikipedia data (2,500M words). The base BERT model also uses WordPiece vectors (Wu et al., 2016) to specify representations for the individual words in the sentence. These have a 30,522 size uncased vocabulary with "subword" vectors, that is, representations for components of words. The subword vectors are aggregated into word vectors, and combined with position vectors (which specify the position of the word in the input sequence), to generate 768-dimensional vector inputs. Input vectors are then transformed through 12 successive layers of interactive attention-based processing, which generate an emergent representation of the sentence in higher layers of the model. This sentence representation is specified using a special CLS token (short for "classifier") which has the same dimensionality (768) as the other input vectors.

In addition to the 12 transformer layers, we also specified an output layer with two nodes (corresponding to true and false). We subsequently fine-tuned the full model on our training data set, which consisted of sentences (concept and feature combinations) and corresponding true and false labels. We did this using the Hugging Face BertForSequenceClassification transformers package for Pytorch (Wolf et al., 2019). Our model generated a classification probability for each sentence using a logistic transformation of the difference between true and false activations in the output layer. If we write the activation value of the true node as A_T and that of the false node as A_F, this classification probability is simply $P_T = \frac{1}{(1+\exp^{-(A_T - A_F)})}$. We used binary cross-entropy to compute loss. Our model was trained using the Pytorch AdamW algorithm with default parameters, batch sizes of 32, and with a total of four epochs. Note that this fine-tuning exercise changed the weights of the full BERT model (i.e., all 110M parameters, including those preceding the final output layer). In this sense, we fine-tuned the "full model." Figure 1E provides an overview of this model.

Below we use the model's true classification probability, P_T, on our various test sentences, to measure its predictive accuracy. In some cases, we also use a variant of this output that calculates the difference between the true and false activation values on the output layer. The differences in true versus false activation, $A_T - A_F$, lie in the range $[-10, 10]$ for sentences tested in this article. To scale these differences so that they are in the [0, 1] range (comparable to the range of the model's classification probabilities), and easy to visualize, we simply use the formula $D_T = 0.5 + 0.05 \cdot (A_T - A_F)$. $A_T - A_F$ and D_T make the same binary (true/false) predictions as classification probability, P_T; indeed, the logistic function applied to $A_T - A_F$ to obtain classification probability is simply a monotonic transformation of $A_T - A_F$, and subsequently D_T. However, unlike P_T, which is typically saturated very close to 0 or 1, $A_T - A_F$ and D_T provide a more graded output, which we find is better at explaining certain behavioral patterns.

We anticipate that performance will improve with BERT's various extensions, as well as variants of BERT that are trained on even larger data sets. We were not interested in the performance of the state-of-the-art model (which changes frequently) so we limited our analysis to the simplest model that is tested in most articles in this literature, recognizing that our results provide only a lower bound on accuracy rates that are feasible using our approach.

Alternative Models

We compared the "full model," introduced above, against six alternatives (all seven models summarized in Table 2): The first model, which we refer to as the "no subject–object exchange model" or the "nSOE model," uses the above training and prediction algorithm, but excludes the false subject–object exchange sentences from the training data. Although these sentences are useful for learning asymmetries in relations (e.g., *cat hunts mouse* is true but *mouse hunts cat* is false), they can have unusual structure (e.g., *fur has cat*), which can bias our model's predictions. Thus, excluding these sentences, and training only on the false concept-replacement sentences, provides a useful robustness test of our full model's predictions.

The second model, which we refer at the "no pretraining model" or the "nPT model" modifies the full model by copying the architecture of the full model as well as the full, pretrained model's first-layer parameters, that is, the word token embeddings, position embeddings, token type or segment embeddings (although these are irrelevant for us since each input is not segmented into, e.g., a pair of sentences whose similarity is being predicted), layer norm weights, and layer norm biases. The rest of this model's parameters, however, are randomized prior to fine-tuning, removing the knowledge learnt during pretraining. The data used to fine-tune the model are identical to the full model. By excluding the main pretraining step, this model tests the extent to which language modeling contributes to performance. Good performance would imply that human-like semantic capabilities can be obtained using only word vectors trained on true and false propositions, and that additional semantic and linguistic knowledge, learnt through extensive pretraining in tasks like masked word prediction and next sentence prediction, is not necessary.

The third model, which we refer to as the "first-layer model," obtains vector representations for each sentence by just averaging the nonfine-tuned BERT input vectors for each word in the sentence. In this way, this model relies on the semantic content of the sentence (rather than on its emergent meaning based on the order and interaction of words in the sentence), and can be used to test the extent to which successful prediction in BERT depends on the interactive, attention-based processing performed by each layer. However, keep in mind that BERT's input vector for a word in a given position in a sentence is the sum of its word vector and the positional vector, which allows the first-layer model to possibly capture some information about the order (although, as we'll see, perhaps not adequately).

Table 2
Overview of the Seven Models Examined in the Current Paper

Model name	Description
Full model	Pretrained BERT model. Fine-tuned on all sentences.
No subject–object exchange (nSOE)	Same as full model, but no false subject–object exchange sentences for fine-tuning.
No pretraining (nPT)	BERT with only embedding layer from pretraining—all other parameters randomized initially. Fine-tuning on same sentences as full model.
First layer	Logistic regression on average of pretrained BERT input vectors for each word in sentence.
Last layer	Logistic regression on CLS representation from the last hidden layer of pretrained BERT.
Language model	Logistic regression on sentence's pseudolog-likelihood (PLL) given pretrained BERT.
GloVe	Logistic regression on GloVe cosine(concept, feature).

Note. BERT = Bidirectional encoder representations from transformers.

The fourth alternative model, referred to as the "last-layer model," provides vector representations for each sentence using the CLS token from the last layer of the nonfine-tuned BERT model. Recall that CLS is a special token standardly prepended to every sentence given to BERT models used for classification or regression purposes, including our full model described above. In the full model, fine-tuning changes all the parameters of the model, and thus the CLS vectors generated for a given sentence depend on the fine-tuning step. By comparing the predictions of our full model against those based on the CLS vector from the nonfine-tuned BERT model, we can evaluate the degree to which fine-tuning the full model helps in making good predictions for semantic verification tasks. The success of the last-layer model would indicate that BERT's pretraining steps are sufficient to equip it with sentence vectors that capture the sentence's truth or falsehood.

Both the first-layer and the last-layer models output 768-dimensional vector representations for sentences (recall, that 768 is the size of the hidden layer in the base BERT model). To map these representations onto true/false classifications, we trained a L2-regularized logistic regression on the output vectors, using cross-validation on the training set to find the optimal regression penalty. While the full fine-tuning procedure can adjust any of the 110M parameters of the BERT model, the logistic regression step for our first-layer and last-layer models can only fit 768 weights and an intercept.

Our fifth model, referred to as the "language model," computes a pseudolog-likelihood (PLL) for every sentence using our nonfine-tuned BERT model. This was done using the *mlm-scoring* package in Python (Salazar et al., 2020). PLLs essentially measure how probable the entire sentence is according to the model, and have been useful for predicting sentence acceptability judgments (Salazar et al., 2020; also see Padó et al., 2009; Porada et al., 2019). We, therefore, suspect that PLL's may be useful for proxying the semantic plausibility of a sentence. Like the previous comparison to the first-layer and last-layer models, comparing the full BERT model to the nonfine-tuned language model allows us to understand the relative contributions of pretraining and fine-tuning on true and false sentences.

Last, we examined a simple GloVe model (Pennington et al., 2014) which uses cosine similarity between the concept and the bag-of-words feature representation in GloVe space (i.e., the average of the vectors of the words that make up the feature) to predict whether a given sentence is true or false. The specific GloVe model used was trained on a combination of Wikipedia and Gigaword corpora (total 6B words) and has a vocabulary of 400,000 words and concepts, each with a 300-dimensional vector representation. Comparing the full BERT model to this GloVe model allows us to understand the extent to which the interactive, attention-based processing of BERT improves over mere similarity between concept and feature in semantic verification. As discussed below, similarity and semantic relatedness have been implicated in previous accounts of various behavioral findings in semantic verification (Glass et al., 1974; Rips et al., 1973; Smith et al., 1974), but generic measures of similarity and relatedness are of course insufficient for explaining semantic verification in general (i.e., *cats eat fur* is false although *cats* and *fur* are highly related or similar). The full BERT model may, therefore, go some distance toward covering this insufficiency.

The reason we used GloVe vectors rather than alternate popular vectors such as the Google News Word2Vec vectors (Mikolov et al., 2013) or even the pretrained WordPiece vectors used in the BERT model (Wu et al., 2016) is because the GloVe model has been used in competitor models such as Feature2Vec (Derby et al., 2019), which we examine below. GloVe also emerged as the best performing model in our own prior work on associative judgment (Bhatia, 2017), in which we found that similarity in GloVe space predicted responses in many judgment tasks, and accounted for classic biases documented in the judgment and decision-making literature. As the GloVe model specified above is nearly identical to the Derby et al. (2019) and Bhatia (2017) models, analyzing its performance can shed light on the improvements we are making over this previous work. Additionally, it is useful to note that the WordPiece vectors used by BERT rely on subword representations. Thus, for example, these vectors might specify a representation for a word like *strawberry* using separate vectors for *straw* and *berry*. Such assumptions may be convenient for transformer networks that modify word representations in multiple interactive neural network layers (using sentence context to constrain and guide representations) but are likely to yield inferior results for the type of simple semantic similarity analysis we wish to perform with our GloVe model.

To obtain semantic verification predictions from the language model's PLL output and the GloVe model's cosine similarity output we simply trained a (nonregularized) logistic regression on these outputs. In contrast to the regressions performed for the first-layer and last-layer models, this entails even less parameter fitting to our training data, allowing for only a bias/intercept and a single coefficient on the PLL score or cosine similarity. It, therefore, does not fit a decision function to the representation of a sentence, but rather to a single metric based on the predicted probability of the sentence or the semantic similarity of the concept and feature in the sentence.

Before we turn to describing the accuracy of these various models, we point the reader's attention to our open science framework (OSF) repository located at https://osf.io/fr4cz/, which contains data and code for the primary analyses with the full (fine-tuned) BERT model.

Predictive Accuracy

Overview of Tests

We began by testing the ability of our full BERT model and its six alternatives to predict the conceptual knowledge reflected in our training data set. We did this through cross-validation: The models were trained on various subsets of the data set and were evaluated in terms of their ability to predict held-out portions of the data set.

The first of these subsets corresponded to a random sample of 90% of the observations. Model performance was evaluated on the remaining 10% of observations. Some observations were repeated multiple times (e.g., in cases where the same feature was listed by multiple subjects), implying that the random training/test split may not have always led to strictly out-of-sample observations. In response, our second approach generated the train/test splits over the unique sentences. Thus, observations corresponding to 90% of the unique sentences were used to train the models and observations corresponding to 10% of the unique sentences were used to test the models. Note that the total number of observations in training and testing were the same as in the prior split, but now there was absolutely no overlap between the two subsets of the data.

Our third approach performed the train/test split over unique concepts. Thus, observations corresponding to 90% of the unique concepts were used to train the models and observations corresponding to 10% of the unique concepts were used to test the models. This approach tested the models on their ability to generalize their learnt information to concepts on which they had not been fine-tuned. Our fourth approach likewise performed the train/test split over unique features. This approach tested the models on their ability to generalize their learnt information to features on which they had not been fine-tuned. Our fifth approach attempted a combination of the out-of-sample concept and out-of-sample feature tests. Here, we randomly selected 5% of the concepts and 5% of the features and excluded all sentences with either these concepts or these features from our training data. This approach tested the models on their ability to generalize their learnt information to both out-of-sample concepts and out-of-sample features.

Our last set of splits examined how well the models generalized to novel domains. For this, we randomly selected three of the 30 categories (10%) in McRae et al. (2005) and excluded all concepts in these three categories from our training data. Models were tested based on their ability to predict features for concepts in these three out-of-sample categories. We repeated these tests a second time with a different set of three randomly selected test categories, due to the small sample sizes involved. The test categories in the first split were *birds*, *drinks*, and *kitchenware*, and the test categories in the second split were *clothing*, *fruits*, and *trees*. Below, we will average the results from these two splits to present a single out-of-sample domain accuracy statistic.

Most of the tests presented below utilize the models' binary predictions (true or false) obtained using a threshold of 0.5 for predicted true/false classification probability, P_T. The results are identical if we replace the full model's probability predictions with the difference in activation measure, $A_T - A_F$, introduced above (as P_T is simply a monotonic transformation of $A_T - A_F$).

Accuracy Rates

As shown in Figure 2A, our full BERT model was able to make predictions with accuracy rates of 97% for out-of-sample observations, 93% for out-of-sample sentences, 86% for out-of-sample concepts, 87% for out-of-sample features, 88% for out-of-sample concept and feature combinations, and 78% for out-of-sample domains. This model's high accuracy rate for the observations split is to be expected, as this split had overlapping test and training sentence-label pairs, yet high accuracy rates persisted even with out-of-sample sentences. The full BERT model was also able to accurately predict the features for concepts that it had not seen and the concepts for features it had not seen. Even though accuracy rates dropped for out-of-sample domains, the full model was still able to make a correct prediction in a majority of cases.

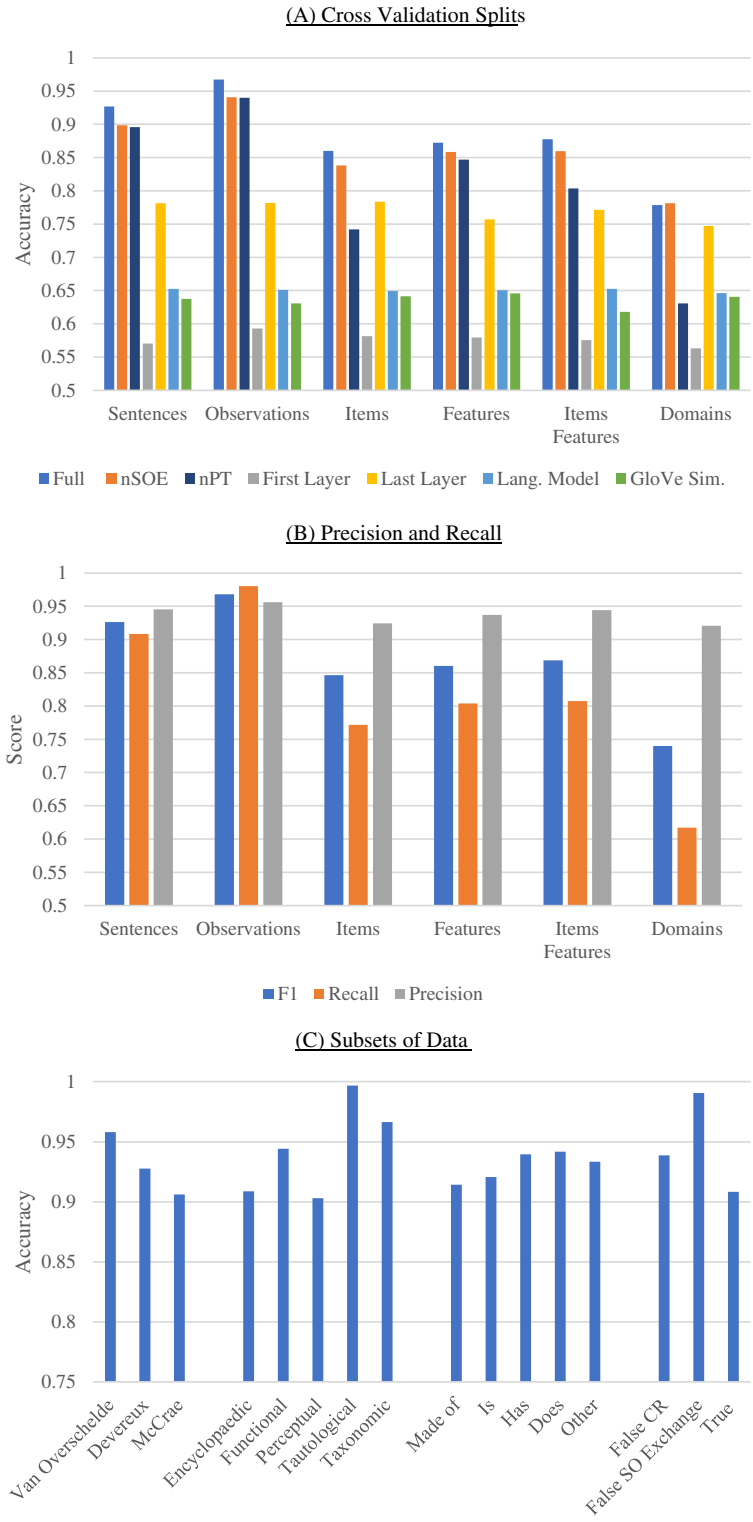
Figure 2A also illustrates performance across all splits for the six alternative models. For all splits, the full BERT model outperformed all alternative models, which generally ranked as follows (from worst to best): first-layer model (using only averaged nonfine-tuned BERT input vectors), GloVe similarity model (based on cosine similarity of the concept and feature), language model (relying on BERT-based PLLs for the sentence), last-layer model (relying on BERT's nonfine-tuned CLS representation in the last layer), nPT model (excluding BERT's pretraining step but retaining its word vectors and the fine-tuning step), and the nSOE model (identical to the full model but trained without subject–object replacement sentences). This ordering was basically consistent with our expectations, as it correlates with each model's sophistication or flexibility of sentence processing, although we expected the language model to perform better as (pseudo-)log-likelihoods of language models are sometimes thought to be capturing something about the (semantic) plausibility of a sentence (e.g., Padó et al., 2009; Porada et al., 2019). The good performance of the nSOE model indicates that subject–object exchange sentences are not strictly necessary for good predictions. Additionally, the mediocre performance of both the nPT and the last-layer models indicates that neither the pretraining nor the fine-tuning steps alone are sufficient. Here it is useful to note that although the nPT model outperformed the last-layer model for out-of-sample sentences and observations, it performed poorly for out-of-sample items and features and was especially bad for out-of-sample domains (where its performance was identical to the GloVe similarity and language models). Thus, it seems that pretraining is especially useful in generalizing beyond the (fine-tuning) training data set.

Figure 2B contains the precision, recall, and F1-scores for the full model, for all train-test splits. In our context, precision is the percentage of sentences that our model correctly judged as true that is actually true, recall is the percentage of true sentences that our model correctly judged as true, and F1 is the harmonic mean of precision and recall. That precision is higher than recall suggests that the full model is conservative about judging a sentence to be true, but is generally accurate when it does so. Interestingly, it is also mostly recall that declines with more challenging train-test splits, indicating that when true test sentences are especially unfamiliar (as in out-of-sample domains) the full BERT model is biased to say they are false. It is this bias that accounts for most of the declines in accuracy with more challenging train-test splits.

Accuracy on Subsets of the Data

We also examined the full model's accuracy rates in the sentences cross-validation split, for various subsets of our test data. Figure 2C

Figure 2
Analysis of Model Performance



Note. Model accuracy for various cross-validation splits (A), precision and recall for out-of-sample sentences for the full model (B), and accuracy of the full model for out-of-sample sentences, as a function of data set, feature type, linking relation, and sentence type (C). See the online article for the color version of this figure.

shows that the full BERT's accuracy rates in predicting out-of-sample sentences persisted for each of the data sets used in this article, although it was better at making predictions for the Van Overschelde et al. category norms than the McRae et al. and Devereux et al. feature norms. In Figure 2C, we also display accuracy as a function of the type of feature being predicted. We can see here that the full model's predictive power for out-of-sample sentences was typically higher for functional (e.g., *cat is used for playing*) and taxonomic features (e.g., *cat is an animal*), as well as for tautologies (e.g., *cat is a cat*), and slightly lower for encyclopedic (e.g., *cat does like milk*) and perceptual features (e.g., *cat is soft*). While we are unsure why encyclopedic features would be more difficult, difficulty with perceptual features is reasonable as they are less likely to be reflected in the linguistic data on which the model was pretrained and fine-tuned. Figure 2C also contains the full model's accuracy as a function of the relation linking the subject with the object of the sentence (e.g., *cat *is a* mammal* vs. *cat *has* fur* vs. *cat *is made of* bones*). Accuracy is generally similar across these linking relations, which may make sense in light of the fact that there's no clear ontological or semantic distinction between some of these relations, as the same proposition can often be expressed with different verbs, for example, *cats are furry* versus *cats have fur*.

Finally, Figure 2C displays accuracy rates for out-of-sample sentences as a function of the type of sentence: true, false with "concept replacement" or false with "subject–object exchange," for the full BERT model. This figure shows two key patterns. First, although model accuracy was high for both true and false sentences, the model was slightly better able to recognize when a given sentence was false than when it was true (which is consistent with our results on precision and recall). Second, higher accuracy rates were obtained for "subject–object exchange" sentences than for "concept-replacement" sentences. This result indicates that the full BERT model can pick up aspects of word order, allowing it to represent asymmetries in relations between concepts (i.e., differences between *cat hunts mouse* and *mouse hunts cat*).

Although not shown here, the above patterns were nearly identical for other cross-validation splits which tested the full model on out-of-sample observations, concepts, features, and domains, rather than just out-of-sample sentences. These patterns also persisted for our model variants, that is, the relative accuracies for different subsets of the test data were the same (though of course the absolute accuracy rates for these alternative models were lower). One pattern that did not persist for the nSOE and GloVe similarity models involved high accuracy on the subject–object exchange sentences. Thus, it seems that subject–object exchange training data are necessary to teach the BERT model how to handle relational asymmetry. Likewise, the GloVe model relies purely on a symmetric measure of similarity between the concept and feature representation, and was thus also not able to handle relational asymmetry. In fact, GloVe's accuracy rate for these sentences was 0.44, which is below chance.

False-Positive Bias

Having established the generally high accuracy of our model in predicting truth and falseness in out-of-sample sentences, we now turn to various tests we conducted to better understand the conditions or factors influencing our full BERT model's performance. As above, we limit our results to only our cross-validation splits involving out-of-sample sentences.

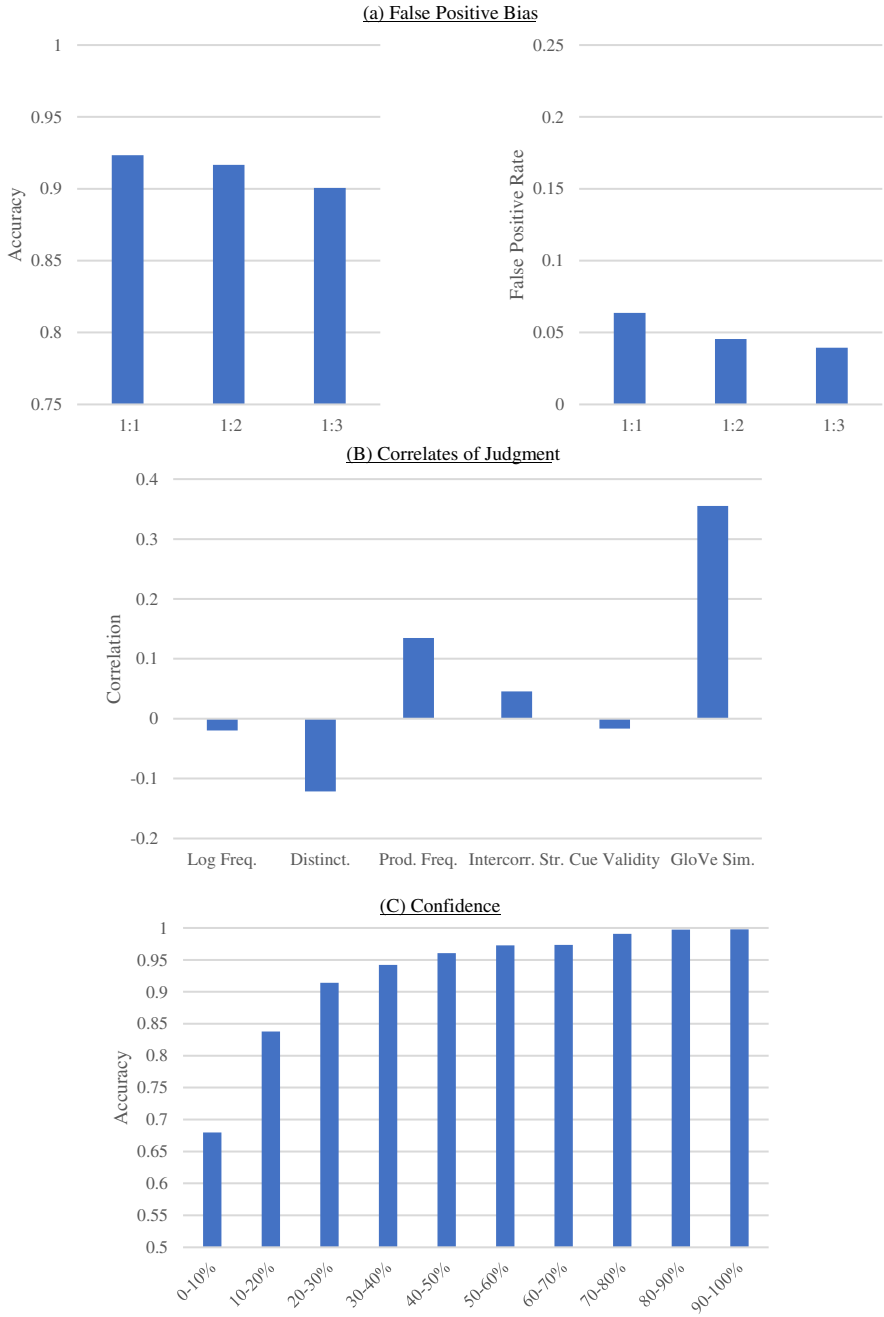
First, one might be concerned that the 1:1 ratio of true and false sentences in our train set induces a false-positive bias in the model since most possible sentences are false. To examine this possibility, we varied the ratio of true to false sentences in the train set while keeping the 1:1 ratio in the test sets. We tried training set ratios of 1:1 as before, as well as 1:2 and 1:3. The training sets used were subsets of the original training data set used above. We equated the size of the training set across all training set ratios, while maximizing the total training set available, resulting in a total of approximately 295,000 training observations in these three tests. Figure 3A shows the resulting accuracies and false-positive rates for out-of-sample sentences for the BERT model. Here, we can see that accuracy only declined slightly (no more than 4% in absolute terms) with increasing proportions of false sentences. Moreover, while the false-positive rate did decrease with lower proportions of true sentences in the training set, the decrease was modest (no more than 4% in absolute terms). Although not shown here, these patterns were nearly identical for cross-validation splits with out-of-sample observations, concepts, features, and domains.

For a second related test, we paired all the concepts used in Rosch (1975) with features that were generated at least three times by subjects in the Devereux et al. (2014) and McRae et al. (2005) data sets. This test involved 560 concepts, 4,654 features, and roughly 2.6 million sentences (concept–feature pairs). Since concepts and features are fully crossed, we would expect the vast majority of sentences generated this way to be false. Indeed, we found that our full BERT model trained on the entire training set (which contains an equal proportion of true and false sentences) classified only 5.26% of these sentences as true. Both of these tests suggest that our model does not contain a large false-positive bias, despite being trained on an equal number of true and false sentences.

Correlates of Model Judgment

Next, we examined how our full BERT model's prediction depended on variables at the word, feature, or concept–feature level. Again, we focused on our cross-validation split for out-of-sample sentences. First, we examined the influence of word frequency on model prediction, under the reasoning that BERT (like people) may be more likely to judge as true, sentences involving words it saw more often during pretraining (see, e.g., Goldstein & Gigerenzer, 2002; Hertwig et al., 2008). For each sentence, we extracted each word's estimate of log frequency from the SUBTLEX-US norms (Brysbaert & New, 2009) and averaged over every word in a sentence (excluding stop words like *the* or *a*). Sentences with one or more words missing frequency estimates were dropped from subsequent analysis. We then Spearman correlated each sentence's average word frequency with the model's predicted probability. As seen in Figure 3B, this correlation was close to zero, suggesting no effect of frequency on model prediction. One might also have expected that frequency should not make the model more likely to output "true," but should make the model more accurate, under the reasoning that BERT would be more competent at processing words it had seen more during pretraining (and hence these words would have greater influence on BERT's parameters). However, a null effect was also obtained when we correlated frequency with the model's likelihood of the correct answer. We are unsure why this null effect arises, although it may simply be that, since our sentences are generated from feature and category norms

Figure 3
Analysis of Model Properties



Note. Full model's accuracy and false-positive rates as a function of true:false ratios in the training set (A), Spearman correlations of full model's predictions with a variety of concept, feature, and sentence-level variables (B), full model's accuracy as a function of its confidence, captured by absolute activation decile (C). All analysis is done on the sentences cross-validation split. See the online article for the color version of this figure.

which generally focus on common concepts and features, there might not be big enough variation in the frequencies of the concepts and features that are in our tests.

Next, we examined the impact of variables provided by McRae et al. (2005) in their feature norms: (a) distinctiveness, which is the inverse of the number of concepts for which a feature was listed, (b) production frequency, which is the number of subjects out of a possible 30 who listed a given feature for a given concept, (c) intercorrelational strength of a given feature for a given concept, which is calculated by summing the proportion of shared variance between a feature and each of the other features of a given concept with which it is significantly correlated, and (d) cue validity, which

is the conditional probability of a concept given a feature, that is, *p(concept|feature)*. See McRae et al., for more discussion of these variables. Note that the first variable, distinctiveness, is unique to a feature, while the latter three are unique to a concept–feature pair. Therefore, the latter three variables are usually only available for concept–feature pairs actually generated by subjects, and so correlating these properties with model predictions is generally only possible for sentences labeled as true.

Figure 3B, thus, contains the Spearman correlations between these variables and the full model's probabilities for out-of-sample true sentences. Feature distinctiveness had a negative correlation, suggesting BERT is less likely to say a (concept, feature) pair is true if the feature is rare. If feature distinctiveness in the norms tracks the frequency of a feature in the world, then this could be rational to some extent, since rare features will, in general, be less likely to be true of any concept than are common features. However, Cree et al. (2006) found that, empirically, people are faster to verify distinctive than shared properties of concepts. Since we draw a correspondence between response time and our model's probability of reporting "true," this empirical result would seem to be at odds with our model's predictions. Interestingly, our model replicates Cree et al.'s results when applied to Cree et al.'s data (as we show below). These differences are likely due to fact that Cree et al. carefully selected the concepts and features in their stimuli, whereas our tests in this section were based on all the out-of-sample true sentences in our data set. Thus, the negative correlation between feature distinctiveness and BERT activation differences observed here could be caused by other variables that we do not control for in this analysis.

In contrast, to feature distinctiveness, production frequency had a positive correlation with BERT's probability of true. This is also sensible: BERT is likely to judge a concept–feature pair as true if that feature is given for that concept often. Intercorrelational strength had a very small positive correlation, consistent with the findings of McRae et al. (1997) discussed below. Finally, cue validity had a near-zero correlation. We are unsure why this may be the case and leave a detailed examination of this issue for future work.

The last potential correlate of prediction that we examined was the GloVe similarity between a concept and the feature (averaging over words in a feature, as we did for the GloVe similarity model). As seen in Figure 3B, we found that the model's predictions had a positive Spearman correlation with the GloVe-based similarity between concept and feature. This indicates that BERT may to some extent be relying on similarity as a heuristic for truth in semantic verification. This is consistent with prior work on semantic judgment (Glass et al., 1974; Smith et al., 1974). It is also consistent with work on high-level judgment (Bhatia, 2017; also see Evans, 2008; Kahneman, 2011; Sloman, 1996), which finds that such a heuristic, although quite useful in many settings, as features that are true of concepts are often semantically related to the concepts, can lead to occasional judgment errors.

Although not shown here, the above patterns are largely consistent across different cross-validation splits, except for the positive correlation with intercorrelation strength (which does not emerge for out-of-sample domains).

Model Errors

We also examined the five true (i.e., subject-generated) sentences with the highest false activation—that is, sentences that BERT thought were false but were actually generated by our subjects (and thus have the label true). These sentences are *belt drives an engine*, *beta is a fish*, *watch has needles*, *corn is a snake*, and *premium is a fuel*. It seems that BERT makes mistakes when sentences involve atypical uses of words and word disambiguation issues (which may also lead human subjects to judge these as false, on occasion). Likewise, the false (i.e., nonsubject generated) sentences with the highest true activation are *axe is made of metal*, *sofa is a chair*, *rabbit is a mammal*, *house is a house*, and *woodpecker has a beak*. These are sentences that BERT thought were true, but were generated by our "concept-replacement" algorithm (and thus have the label false). As can be seen here, these are all actually true sentences about the world, indicating that the accuracy rates shown above may underestimate BERT's ability to predict the features that people associate with concepts.

To further explore the possibility that in some sense we underestimate BERT's accuracy, we asked 19 subjects (M_{age} = 28 years, SD_{age} = 12 years, 52% Female) on Prolific Academic to judge the truth of 100 sentences labeled as false. The sentences labeled as false but with the highest model-based probability of truth (e.g., *sofa is a chair*) formed half of this set, and the sentences labeled as false but with the lowest model-based probability of truth (e.g., *jacket is a stair*) formed the other half. These stimuli were generated from just the McRae et al. sentences, we excluded all subject–object exchange sentences, and we slightly adjusted the grammar of each sentence to ensure that it was grammatical to a human reader.

Overall, subjects were generally likely to judge the former set as true, and the latter set as false. A mixed-effects logistic regression of subject judgment with fixed effects for the model-based probability of the sentence being true, and random intercepts for subject and sentence, confirmed this ($p < .001$). More simply, on average the high probability sentences were judged as true by 78% of subjects and the low-probability sentences as true by 10% of subjects. The fact that so many of our high probability "false" sentences may actually be true again suggests that our reported accuracy rates may be underestimates in certain respects.

Model Confidence

In a related test of model errors, we wished to examine how our full BERT model's accuracy rates varied based on the strength of its predictions, that is, its confidence. For this purpose, we calculated accuracy as a function of the model's predictions for the test data in our sentences cross-validation split. Here we took the absolute value of the model's activation difference, $|A_T - A_F|$, in order to quantify confidence, and further pooled observations based on a decile split of these absolute activation difference values. Accuracy rates for each decile split for out-of-sample sentences are shown in Figure 3C. Here, we can see that BERT was more likely to respond accurately if assigned a high activation to the sentence being true relative to false or false relative to true. In other words, BERT was less likely to have strong responses for concept–feature pairs it was likely to get wrong.

Relation-Sensitive Representations

In our final diagnostic test, we examined how different relations influenced the full BERT model's representation of concepts in the subject position of a sentence. Whereas distributional semantics

models like GloVe only have a single vector for a given word regardless of sentence context, BERT is able to provide contextualized representations for a given word, such that the representation for *dog* in *dog has* is different from that in *dog can*. Of course, this is reminiscent of classic connectionist models of semantic cognition. For example, Rogers and McClelland (2004, 2008) found that, in their trained neural network for predicting features, the hidden layer representations of living things shifted depending on whether they were joined with *can* or *is*. In particular, the context of *can* collapsed differences among plants—relative to their contextless representation in an earlier layer—since as far as their network knew, the only thing plants *can* do, is *grow* (see Figure 11 of Rogers & McClelland, 2008 for more). We thus sought to see if similar behavior arose in our full BERT model.

First, we computed similarities among all pairs of the 541 concepts in the McRae et al. feature norms twice, once for features containing the *is* relation (*is an animal*, *is furry*, etc.) and once for features containing the *has* relation (*has claws, has wooden legs*; we chose *is* and *has* simply because they were the most common relations). These similarities were simply calculated as the cosine similarity between two concepts' feature vectors, where the entries of each vector were the production frequency for that (concept, feature) pair. Second, we extracted our full BERT model's (trained on the entire training data) final hidden layer representation for each concept twice, once in the context of *is* (e.g., *dog is*) and once in the context of *has* (e.g., *dog has*). We then calculated cosine similarity among concept representations in the same context. As predicted, norm-based similarities among *is* features correlated better with BERT-based similarities in the context of *is* than with *has* (correlation of 0.15 relative to 0.09), whereas norm-based similarities among *has* features correlated better with BERT-based similarities in the context of *has* than with *is* (correlation of 0.11 relative to 0.09). Although the magnitude of these differences is small, they are reliable (95% confidence intervals only about .012 points in size). This suggests that, like its predecessor connectionist models, BERT is somewhat able to appropriately shift a concept's internal representation given the relation with which it is paired. This, in turn, allows BERT to predict that "similar" concepts have similar features slightly more when "similarity" is with respect to particular relations linking concept and feature.

Comparison With Feature2Vec

We are not the first to fine-tune a distributional semantic model using subject-generated feature norms. In recent work, Derby et al. (2019) have used feature norms data from McRae et al. and Devereux et al. (2014) to learn feature representations as vectors in a GloVe semantic space. This approach, known as Feature2Vec, is closely related to our GloVe model, in that the degree to which a feature is attributed to a concept is assumed to be proportional to the semantic similarity between the concept and the feature vector. Of course, the Feature2Vec feature representations are obtained not by a simple bag-of-words averaging of the GloVe vectors (as we did above) but by finding the feature vectors in the GloVe space that best predict participant responses. This approach allows Feature2Vec to potentially uncover feature representations that diverge from the averaged GloVe representations of their component words. Due to its reliance on participant data, Feature2Vec is able to accurately model feature norms and successfully predict features for out-of-sample concepts (it also outperforms previous approaches based on similar data, such as PSLR—Făgărăşan et al., 2015).

Feature2Vec does, however, have limitations: Its need for extensive training data makes it difficult for it to learn representations when feature data is sparse. This is why Derby et al. (2019) trained and tested Feature2Vec on experimenter-processed variants of the McRae et al. and Devereux et al. data sets, which pooled semantically related features. In this application, Feature2Vec learnt a single feature representation for distinct features like *has fur*, *is fluffy*, and *has hair*, and had representations for only 2,725 features in the Devereux et al. data set. BERT, in contrast, requires minimal experimenter preprocessing of features, which is why we were able to train and test it on all 26,075 raw subject-generated features in the Devereux et al. data set.

Feature2Vec's reliance on subject-generated features also explains why it is unable to extrapolate its learnt knowledge to out-of-sample features: Features that are absent in the training data just cannot be given vector representations. Again, BERT does not have this limitation, and successfully predicts out-of-sample features and features composed of novel word combinations (just as it does out-of-sample concepts).

Despite the unique strengths of BERT relative to Feature2Vec, it is useful to examine BERT's performance in the types of settings in which Feature2Vec excels. To do this, we trained and tested a separate set of models on variants of the McRae et al. (2005) and Devereux et al. (2014) data sets outlined above. Specifically, we (a) obtained the cross-validation splits used in Derby et al. (2019); (b) trained the full BERT model on the features for the training concepts in Derby et al. (400 concepts from McRae et al., and 500 from Devereux et al.); and (c) tested the full BERT model on the features of the test concepts in Derby et al. (141 concepts from McRae et al., and 138 from Devereux et al.). As in Derby et al., our tests measured the overlap between observed and predicted features. In the case of a concept with K features, we obtained the K features that yielded the strongest activation in favor of true (vs. false) for the concept in the trained BERT model. Overlap was calculated by measuring the proportion of predicted features in the subject-generated features data set. To further mimic the tests in Derby et al., we used experimenter-processed features (rather than raw subject-generated features) in our training and test data, and generated false sentences using only the "concept-replacement" negative sampling method (and not the "subject–object method"). In this way, we replicated Derby et al.'s tests with BERT. However, unlike Derby et al., who negatively sampled at a rate of 20x, we negatively sampled a rate of only 5x (due to computational constraints).

Table 3 shows the results of this analysis for the McRae et al. and Devereux et al. data sets. It also displays the results of Derby et al. for the Feature2Vec approach and the related PLSR approach of Făgărăşan et al. (2015), taken from Table 2 of Derby et al. (with 50 and 120 corresponding to the dimensionality of the trained PLSR model). Here, we can see that BERT obtained an overlap of 37% on McRae et al. and 43% on Devereux et al. test concepts, and thus matched the predictive power of Feature2Vec (which obtained an overlap of 35% and 44% on the two data sets respectively). Both BERT and Feature2Vec outperformed the PLSR approach. These results indicate that BERT has similar performance to Feature2Vec in the settings in which Feature2Vec performs well.

Table 3
Accuracy Rates For Derby et al. (2019) Tests

	McRae et al.		Devereux et al.	
Model	Train	Test	Train	Test
PLSR50	49.52	31.67	50.58	40.25
PLSR120	68.66	32.97	65.42	40.71
Feature2Vec	90.70	35.33	90.31	44.30
BERT	73.61	36.88	80.97	43.32

Note. These statistics measure the proportion of out-of-sample features in the McRae et al. (2005) and Devereux et al. (2014) data sets that are correctly predicted. PSLR corresponds to the Partial Least Square Regression (PLSR) approach of Făgărăşan et al. (2015), with 50 and 120 indicating the dimensionality of the model. The statistics for PSLR50, PSLR120, and Feature2Vec are taken from Table 2 of Derby et al. (2019). BERT = Bidirectional encoder representations from transformers.

Discussion

We have trained and tested BERT on feature and category norms data. Each prediction involved an assessment of whether a given concept and feature combination was generated by human participants (e.g., *cat has fur*) or whether it was generated by our "concept-replacement" (e.g., *tractor has fur*) or "subject–object exchange" (e.g., *fur has cat*) algorithms. As the participants typically generated features that were true of the concept and our "concept-replacement" and "subject–object exchange" algorithms typically generated features that were false for the concept, the task can equivalently be seen as involving semantic verification—that is truth or falsehood judgments for simple natural language propositions composed of common concepts and features.

We have found that our full BERT model was able to make accurate predictions, even with the most demanding train-test splits involving out-of-sample domains. Our full BERT models' reported accuracy rates may to some degree actually be underestimates, since many of our generated "false" sentences that the model predicted to be true, were also judged as true by human subjects, and conversely, some sentences our model "incorrectly" but confidently labeled as false seem likely to mislead human subjects as well (since they involve rare word senses). Despite being trained on equal numbers of false and true sentences (while most possible sentences are actually false), two separate tests revealed that this did not give our fine-tuned BERT model a strong false-positive bias. Moreover, this model's precision was higher than its recall, suggesting that its errors were more driven by being overly conservative in judging true sentences as true.

We conducted various tests exploring the settings and factors influencing our full model's performance. We found that the model's predictive power persisted for all three data sets, for the different types of features and relations in these data sets, for true sentences, and for the two types of false sentences. However, BERT did better on category norms, nonperceptual features, and false sentences involving "subject–object exchange," although we did find that these latter sentences were harder to predict correctly when the model was not trained on them (as was the case with the nSOE model). Predictive power also was higher when BERT had higher absolute activation values and was thus more confident in its beliefs. Our full model thought that a sentence was likelier to be true when the feature was less distinctive (more common) according to feature norms (but see our analysis of Cree et al., 2006 in the next section), when the feature was listed more often for that concept in feature norms, and when the concept and feature had high GloVe similarity. This last finding suggests that BERT may be using similarity as a heuristic, consistent with prior work on both semantic verification (Glass et al., 1974; Smith et al., 1974) and high-level human judgment (Bhatia, 2017; also see Evans, 2008; Kahneman, 2011; Sloman, 1996). Finally, we found that BERT appropriately adjusted a concept's internal representation given the relation with which it was paired, much like earlier connectionist models (e.g., Rogers & McClelland, 2004, 2008).

We compared our full fine-tuned BERT model against several alternative models, ranging from concept–feature GloVe similarity to logistic regression trained on final layer pretrained BERT representations. Our full model was better than all of these alternatives. The full BERT model also equaled the accuracy rates of Feature2Vec on out-of-sample concepts, implying that it retains the benefits of previous models that use feature norms to fine-tune distributional semantic representations.

The superiority of the full model over all of these alternatives leads to two conclusions regarding the source of its successes. First, the multilayer, interactive, attention-based processing of BERT adds considerable power over both purely similarity-based processing of GloVe or Feature2Vec, and simply finding a decision boundary in the input vector space which retains only the most rudimentary information about word order. Second, the full BERT model's success comes from both pretraining on a generic objective like masked word prediction or next sentence prediction on large language corpora (giving the model linguistic capabilities, specifically the ability to process basic syntactic structure), as well from fine-tuning all its parameters to the specific task of judging sentences to be true or false using psychological data (giving the model the ability to generalize its semantic and linguistic knowledge for semantic verification). We examine the theoretical implications of these results in greater detail in the general discussion of this article.

Patterns of Semantic Cognition

Overview of Tests

We also tested whether our full BERT model was able to generate well-known behavioral patterns in semantic cognition. Its ability to make predictions for truly out-of-sample concepts, features, and domains, allowed us to perform such tests using original stimuli from the articles that first documented these patterns. Such tests are crucial for evaluating the degree to which BERT's accuracy rates go beyond just predicting whether a feature is true of a concept, and extend to predicting response times, typicality ratings, and other patterns associated with psychological data. Transformer networks are highly flexible and perform well at most natural language processing tasks if trained on appropriate data. Thus, it may not be particularly surprising that a BERT model trained on category and feature norms data can learn to predict these data with high accuracy. But if this model is also able to account for additional behavioral patterns (that are not directly present in the data on which it was trained), then we may be able to conclude that its learnt representations, and processes for mapping these representations onto semantic judgments, closely resemble those used by humans.

We tested BERT on stimuli sets from 25 different experiments, corresponding to the 17 different behavioral patterns shown in Table 4. We obtained these stimuli sets from published articles

Table 4
Summary of Semantic Cognition Patterns and Stimuli Sources Examined in This Article

Response times in semantic verification

1. *Collins and Quillian (1969) Exp. 1 and 2:* Higher response time (RT) when the distance, in a taxonomic tree, between a concept and its superordinate category is larger. For example, RT(*canary is a bird*) < RT(*canary is an animal*).
2. *Rips et al. (1973) Exp. 1 and Smith et al. (1974) Exp. 2:* Reversals of Collins and Quillian (1969) patterns due to semantic relatedness. For example, RT(*bear is an animal*) < RT(*bear is a mammal*).
3. *Glass et al. (1974) Exp. 2 and 3:* Higher RT for true semantically unrelated statements but lower RT for false semantically unrelated statements. For example, RT(*all mothers are women*) < RT(*all mothers are parents*) and RT(*all buildings are roses*) < RT(*all buildings are houses*).
4. *Anderson and Reder (1974) Exp. 1:* RT for false sentences depends on time to generate superordinate category, and time to falsify superordinate. For example, RT(*tent is a metal*) ∼ RT(*tent is a dwelling*) + RT(*dwelling is a metal*).
5. *Glass et al. (1979) Exp. 1:* Higher RT when judging false sentences with indirect antonyms than direct antonyms. For example, RT(*all boys are girls*) < RT(*all boys are sisters*).
6. *Hampton (1984) Exp. 1 and 2; and Collins and Quillian (1969) Exp. 1 and 2:* Higher RT for true and false statements involving feature judgments compared to category judgments. For example, RT(*oak is a tree*) < RT(*oak is green*) and RT(*oak is a creature*) < RT(*oak eats mice*).
7. *McRae et al. (1997) Exp. 3:* Higher RT for true sentences with uncorrelated features than true sentences with correlated features. For example, RT(*deer is hunted by people*) < RT(*duck is hunted by people*).
8. *Cree et al. (2006) Exp. 1 and 2:* Higher RT for true sentences with nondistinctive features than true sentences with distinctive features. For example, RT(*ant lives in a colony*) < RT(*ant lives in the ground*).

Typicality ratings for concepts and categories

9. *Rosch (1975) Exp. 1:* Category exemplars vary based on their typicality. For example, *robins* are more typical *birds* than *penguins.*
10. *McCloskey and Glucksberg (1978) Exp. 1:* Higher inconsistency across subjects for concepts with intermediate typicality. For example, subjects agree with each other that *chair is furniture* is true, and that *cucumber is furniture* is false, but often disagree about whether *bookends are furniture* is true.
11. *Hampton (1982) Exp. 1 and 2:* Category membership judgments can violate transitivity due to typicality. For example, many subjects agree with *husky is a dog* and *dog is a pet*, but disagree with *husky is a pet.*
12. *Roth and Mervis (1983), Exp. 2 and 3:* Typicality can violate set-membership relations. For example, *strudel* is a more typical *pastry* than *pie*, whereas *pie* is a more typical *dessert* than *strudel.*

Distribution of features across concepts

13. *Rosch and Mervis (1975) Exp. 1:* Concepts typical of a category have many overlapping features with other concepts in the same category and few overlapping features with concepts in other categories. For example, *chair* has more features that overlap with members of the *furniture* category than does *telephone.*
14. *Malt and Smith (1984) Exp. 1 and 2:* Some features covary positively or negatively with each other. For example, *birds* that *fly* also usually *sit in trees.*

Similarity ratings

15. *Whitten et al. (1979) Exp. 1:* Similarity can be asymmetric. For example, *era* is judged to be highly similar to *age*, but not vice versa.
16. *Hill et al. (2015) Exp. 1:* Similarity can diverge from association. For example, *refrigerator* and *food* are highly associated, but not similar.
17. *Richie and Bhatia (in press) Exp. 1:* Similarity judgment rules vary across categories. For example, the features that make *robins* similar to *sparrows* are different to those that make *gloves* similar to *scarfs*.

that gave examples of the experimental stimuli in the article. We used every stimuli set involving real-world conceptual knowledge that we were able to find. We excluded semantic cognition tasks with artificial concepts or features. We also excluded tasks in which participants were asked to evaluate naturalistic propositions based on previously presented information (e.g., tasks in which participants first read a story and then evaluated whether statements about the story were true or false). Additionally, some seminal publications and findings were not accompanied by examples of actual experimental stimuli, making it impossible for us to replicate every classic result in the literature. Finally, due to the vast scope of the literature on concept combination (e.g., semantic verification of sentences involving adjective–noun combinations such as *red apple* and *brown apple*; e.g., Murphy, 1988; Smith & Osherson, 1984), we have not tried to replicate its findings in the current article. Thus, our tests are limited to the verification of sentences composed of simple concepts or nouns.

The BERT model used for our tests was trained on the full category and feature norms data set described in the Implementational Details section. This had 491,284 sentences with true and false labels, pertaining to 2,066 unique concepts and 29,048 unique features. This model was tested on the stimuli sets by manually transforming concept–feature pairs into natural language sentences, and calculating BERT's predicted classification probability (of true vs. false), P_T, for these sentences. We compared the full BERT model to the nSOE model, the nPT model, the first-layer model, the last-layer model, the language model, and the GloVe similarity model introduced above. These were also trained on the full category and feature norms data sets (in the case of all models except for nSOE, this only involved fitting the weights of a logistic regression model), and tested based on their classification probabilities for sentences in existing stimuli sets. We also evaluated the three "activation-based" models (the full model, the nSOE model, and the nPT model) based on their predicted activation differences for true versus false responses, $A_T - A_F$ (visualized through the normalized measure D_T). Thus, in Figures 2 through 7, P_T is visualized in bars labeled Full Prob., nSOE Prob., and nPT Prob. Likewise, D_T is visualized in bars labeled Full Act., nSOE Act., and nPT Act. Although P_T predicts the probability that an individual judges a sentence to be true, $A_T - A_F$ captures the underlying response tendencies, which are more predictive of response times, typicality judgment, and other psychological variables examined in existing work. Note again that P_T is simply a monotonic transformation of $A_T - A_F$ and D_T, and that both P_T and D_T are constrained to the range [0, 1], with a threshold of 0.5 for judging a sentence to be true or false.

Note that using the full training data set for our BERT model did imply that the model had been previously exposed to some of the experimental stimuli, and in this sense was occasionally making in-sample predictions. That said, the frequency of these in-sample predictions was quite low. Overall, only 7.67% of the sentences used in the experimental stimuli were in the training data. Likewise, only

28.20% of the target concepts, only 8.56% of the features, and only 63.98% of the unique words in the experimental stimuli were in the training data. The main reason why there is a large overlap in the number of words, but not in the number of sentences, is because experimental stimuli sets combine features with concepts in ways that are different to how human participants generate feature norms. For example, participants seldom generate semantically unrelated features for concepts in feature norm studies, but such features are often used in the experimental manipulations summarized below.

Response Times in Semantic Verification

Effect 1: Collins and Quillian (1969) Experiments 1 and 2

We began by testing our models' predictions for concept and feature pairs used in Collins and Quillian (1969) article. In this article, Collins and Quillian argued that semantic memory was organized hierarchically and that features were only stored at the highest level possible. For example, the feature *is an animal* would only be associated with *bird*, and not *canary*. Thus, to evaluate *canary is a bird*, one would first need to check whether canaries are animals, before moving up the hierarchy to check whether birds are animals. This generated the prediction that it would take longer for participants to verify *canary is an animal* than *canary is a bird.* Likewise, it would take longer for participants to verify *canary has skin* than *canary has wings* (because *has skin* is associated with animals, while *has wings* is associated with birds). Our BERT model does not involve explicit hierarchical organization, but may nonetheless be able to generate these patterns. This can be tested by assuming that the response time when responding true to a true sentence is proportional to BERT's prediction for true.

The stimuli we used to test this pattern were taken directly from Figure 1 of Collins and Quillian. This highly reproduced figure presents a hypothetical three-level semantic network with seven concepts (*animal*, *bird*, *fish*, *canary*, *ostrich*, *shark*, and *salmon*) and two to four features per concept. The concepts and features in this figure yield 73 distinct true sentences, spanning the three levels of the hierarchy. Figure 1 can also be used to generate false sentences. For this we swapped out features of a concept with other concepts at the same level of the hierarchy. Thus, for example, features of *bird* were attached to *fish* (e.g., *fish has wings*) and vice versa (e.g., *bird has fins*). In addition to Figure 1, we also used the stimuli presented in Table 1 of the article. These stimuli consisted of true and false sentences offered to participants in Experiments 1 and 2. Experiment 1 sentences spanned three levels of the hierarchy and involved different types of sports (e.g., *baseball*, *badminton*) and trees (e.g., *oak*, *spruce*), whereas Experiment 2 sentences spanned two levels and involved different types of drinks (*seven-up*, *ginger ale*).

In Figure 4A, we show the full BERT model's average predictions for sentences at each level of the hierarchy in the above stimuli. Here we can see that BERT was able to capture the Collins and Quillian effect, with lower predictions for sentences involving higher levels of the hierarchies. BERT also correctly judged most false sentences to be false and true sentences as true (though note that some Level-2 sentences were incorrectly judged to be false). These results are largely insensitive to whether we measure prediction using classification probability or difference in activation (note that this figure, and all subsequent figures plot $D_T = 0.5 + 0.05 \cdot (A_T - A_F)$ rather than $A_T - A_F$, to keep the *y*-axis on the 0–1 scale of probabilities). Figure 4A also presents analogous results for the five model variants. Here, we see that the nSOE and nPT models mimicked the predictions of the full BERT model and were able to successfully describe the data. The last-layer and GloVe similarity models also captured some of the level-of-hierarchy effects but did so imperfectly. Specifically, the last-layer model assigned equivalent predictions to Level-0 and Level-1 sentences, and fairly high predictions (probabilities around .5) to false sentences. The GloVe model likewise assigned high predictions to false sentences, and in fact gave these sentences higher predictions than Level-2 sentences. In this way, these alternate models were unable to distinguish true from false sentences. Finally, the first-layer model and language model were unable to capture these patterns and assigned all levels roughly similar predictions.

We examined the full BERT model's predictions more rigorously by regressing activation differences, $A_T - A_F$, on level-of-hierarchy for true sentences. Our regressions also included fixed effects for sentence type (i.e., whether it involved category membership, e.g., *is an animal*, or not, e.g., *has wings*). These regressions revealed a statistically significant negative relationship between sentence level and BERT activation difference, $\beta = -2.92$, $t(87) = -4.69$, $p < 0.001$, 95% CI = [−4.16, −1.68].

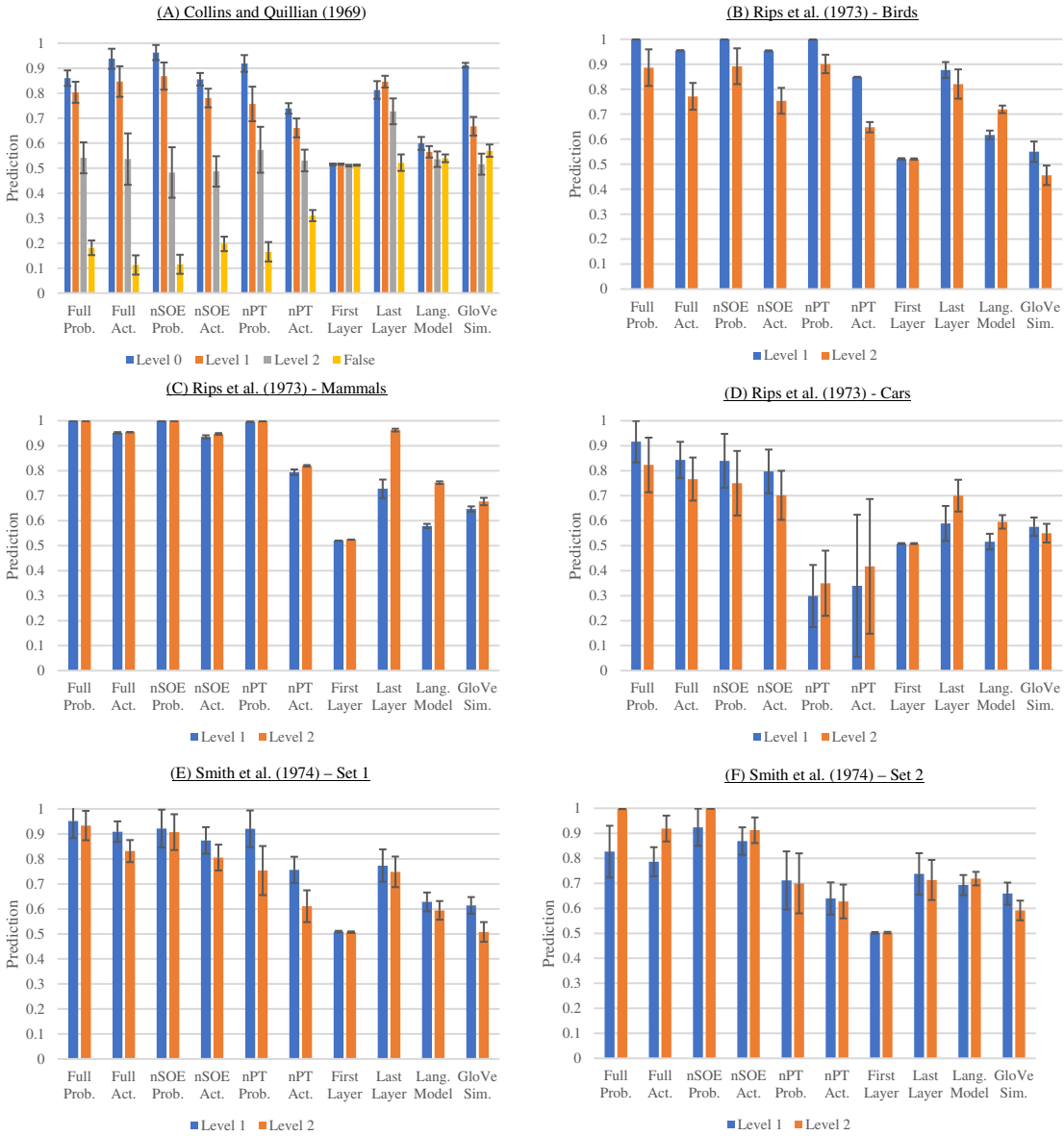
Effect 2: Rips et al. (1973) Experiment 1 and Smith et al. (1974) Experiment 1

Even though the simple GloVe model was unable to distinguish true from false sentences, it was able to capture Collins and Quillian's level-of-hierarchy effects for true sentences, indicating that semantic relatedness is a confound in Collins and Quillian's analysis. This is a point that was made by several researchers shortly after the publication of Collins and Quillian (1969). Out of these Rips et al. (1973) and Smith et al. (1974) showed that the effects predicted by Collins and Quillian's theory could be reversed for certain categories based on the semantic relatedness of the categories and their exemplars. Thus, even though *canary is a bird* was verified more quickly than *canary is an animal*, Rips et al. (1973) and Smith et al. (1974) found that sentences like *bear is a mammal* were verified more slowly than *bear is an animal*.

To see if our models were able to capture these reversals, we applied them to Table 2 of Rips et al., which summarizes the stimuli from Experiment 1 of their article. These stimuli involve 12 birds (e.g., *blue jay*, *cardinal*), 12 mammals (e.g., *bear*, *cat*), and 12 car brands (e.g., *Cadillac*, *Continental*). Participants were asked whether each of these concepts belonged to one of two superordinate categories: *bird* and *animal*, *mammal* and *animal*, and *car* and *vehicle* respectively. The first of these is a Level-1 category, whereas the second of is a Level-2 category. Rips et al. found that response times were faster for Level-1 categories in the case of *bird* and *car*, but slower in the case of *mammal*. In total, the experiment involved 72 different sentences generated by pairing the 36 concepts with each of the two superordinate categories.

Figure 4B–4D show that our full BERT model captured these patterns. Specifically, the model gave higher activations and probabilities for Level-1 categories relative to Level-2 categories for *bird* and *car*, but not for *mammal* (though note that the differences for *mammal* are fairly minor, as the model's predictions are saturated

Figure 4
Model Predictions for Behavioral Effects



Note. Average model predictions for level-of-hierarchy effects documented by Collins and Quillian (1969) (A) and moderators of the effects documented by Rips et al. (1973) for birds (B), mammals (C), and cars (D), and by Smith et al. (1974) for Set 1 stimuli (E) and Set 2 stimuli (F). Rips et al. and Smith et al. find that the level-of-hierarchy effects reverse for mammals and Set 2. Error bars denote +/−1 *SE* around the mean. See the online article for the color version of this figure.

very close to 1). We obtained similar results for the nSOE model and the GloVe similarity model, whereas the remaining models failed to capture all of the observed effects. Note that the nPT model did capture the effects for the *bird* and *mammal* categories, but failed at *car*, likely due to its poor performance for out-of-sample concepts and domains (there were many birds and mammals, but no car brands, in feature norms data sets used to train the models).

To rigorously test for a statistically significant shift between *bird* and *mammal* we regressed BERT activation difference, $A_T - A_F$, on category level (Level-1 or Level-2), Level-1 category name (*bird* or *mammal*), and their interaction. We found a statistically significant negative interaction, $\beta = -3.73$, $t(48) = -3.32$, $p < 0.01$, 95% CI = [−5.99, −1.47], indicating that BERT does capture the reduction in the level-of-hierarchy effect for *mammal* relative to *bird*.

We also performed a similar set of tests using stimuli from Table 3 of Smith et al. (1974). This table consists of two general sets of concepts, spanning a variety of domains including *animals*, *minerals*, *liquids*, *foods*, *body parts*, *universities*, *stones*, *musical instruments*, and *buildings*. In the first set, the concepts (e.g., *butterfly*) are semantically more similar or related (as measured by

ONLINE COLOR

production frequencies from Loftus & Scheff, 1971) to a Level-1 superordinate category (e.g., *insect*) than to a Level-2 superordinate category (e.g., *animal*). For this set, Smith et al. replicated the level-of-hierarchy effect, with quicker response times for Level-1 category verification judgments than Level-2 judgments. In the second set, however, the concepts (e.g., *aluminum*) are semantically more similar to a Level-2 superordinate category (e.g., *metal*) than to a Level-1 superordinate category (e.g., *alloy*). Here, Smith et al. found a reversal of the level-of-hierarchy effect, with faster responses for the Level-2 judgments than Level-1 judgments. There are a total of 13 concepts in each set, yielding 52 sentences total.

Figure 4E and 4F show that the full BERT model captured these patterns by giving higher activations for Level-1 relative to Level-2 categories in the first set, but lower activations for Level-1 relative to Level-2 categories in the second set. The nSOE model also captured this effect. The remaining models were unable to capture the effects. A regression of the full BERT model's activation difference ($A_T - A_F$) on category level (Level-1 or Level-2), category set (Set 1 or Set 2), and their interaction, revealed a marginally significant negative interaction, $\beta = -4.20$, $t(52) = -1.95$, $p = 0.06$, 95% CI = [−8.53, −0.12].

Effect 3: Glass et al. (1974) Experiments 2 and 3

The above tests were largely concerned with response times in semantic verification tasks in which the statements are true (although Smith et al., 1974 did test for semantic relatedness effects in false sentences, they did not provide their stimuli for these tests in their article). Glass et al. (1974), in contrast, documented semantic relatedness effects for both true and false sentences, and also provided examples of the stimuli used in their tests. Glass et al. found that participants were quicker to accept true sentences in which the concepts and features were semantically related, but slower to reject false sentences in which the concepts and features were related. Thus, semantic relatedness facilitates the verification of true sentences but inhibits the falsification of false sentences.

To test if our models were able to capture this pattern, we used the stimuli presented in Tables 2 and 5 of Glass et al. (1974). Table 2 has category membership sentences used in Experiment 2 of their article. These are eight sentences that are true and have semantically related categories (e.g., *all mothers are women*), eight that are true and have unrelated (or less related) categories (e.g., *all mothers are parents*), eight that are false and have related categories (e.g., *all buildings are houses*), and eight that are false and have unrelated categories (e.g., *all buildings are roses*). Likewise, Table 5 has feature verification sentences used in Experiment 3. These are 10 sentences that are true and have semantically related features (e.g., *all arrows are pointed*), 10 that are true and have unrelated features (e.g., *all arrows are narrow*), 10 that are false and have related features (*e.g., all fires are yellow*), and 10 that are false and have unrelated features (e.g., *all fires are rusty*). There are a total of 72 sentences in these two tables, involving a diverse set of domains, including *weapons*, *plants*, *animals*, *furniture*, *occupations*, and *social roles*. As with Smith et al., Glass et al. measured semantic relatedness using production frequencies.

Note that these sentences do have quantifiers. However, Glass et al. found that the *all*, *many*, *some*, and *few* quantifiers all displayed the same effect and that the *no* quantifier displayed the reversal of the effect. For this reason, our tests modified these sentences to exclude quantifiers and to reverse the truth value of the sentences prefixed by *no*. Thus, for example, we replaced *all arrows are pointed* in the original stimuli set with *arrows are pointed* for our tests. Likewise, we replaced *no forests are treeless* in the original stimuli set with *forests are treeless,* and additionally recoded this sentence as false.

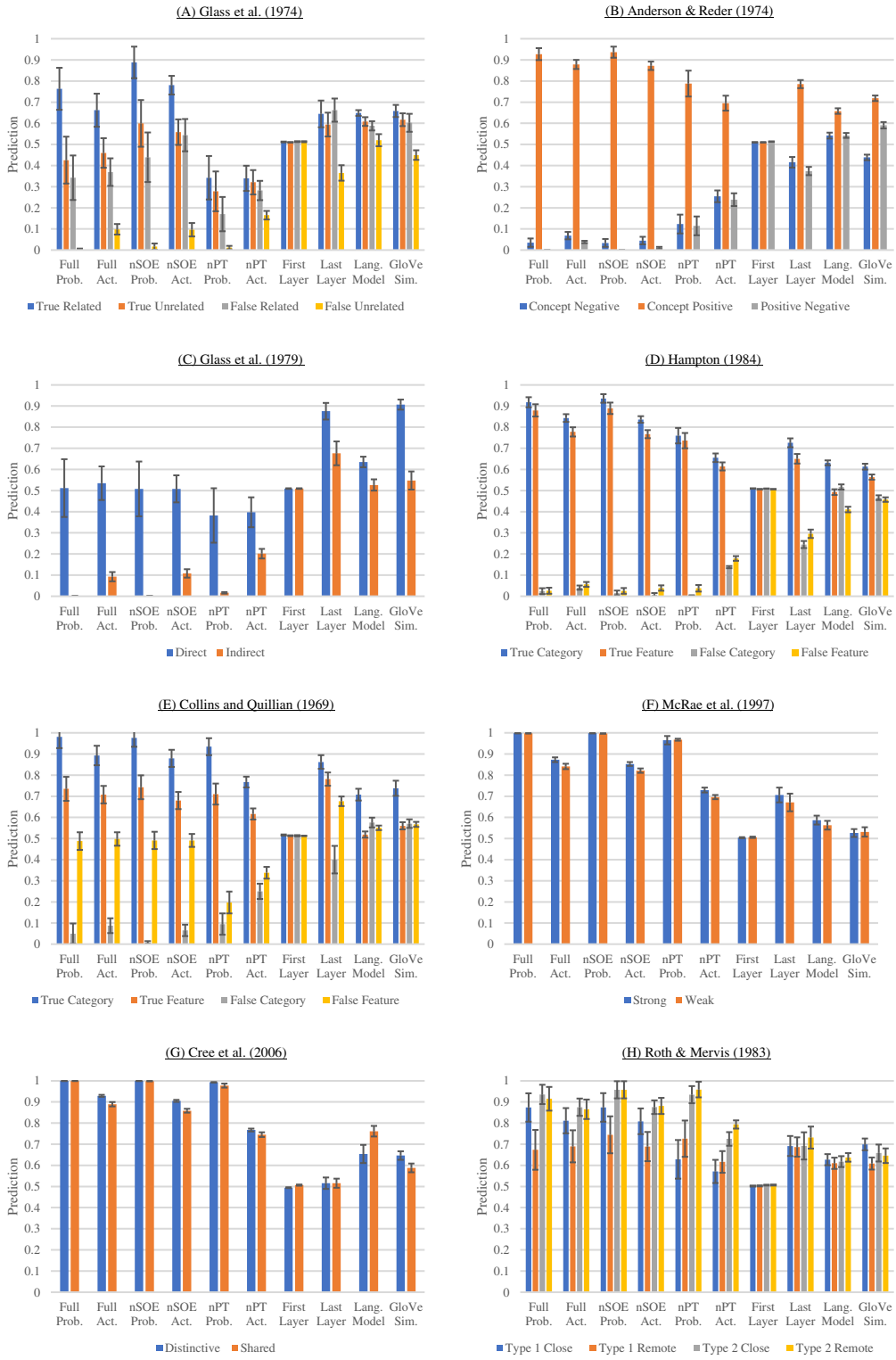
Figure 5A shows model predictions for true and false sentences involving related and unrelated categories or features. Here we see that the full BERT model generated higher probabilities and activation differences for true-related sentences than true-unrelated sentences, capturing the finding that true-related sentences are verified more quickly than true-unrelated sentences. Likewise, this model generated lower probabilities and activation differences for false-related sentences than false-unrelated sentences, capturing the finding that false-related sentences are verified slower than false-unrelated sentences. Additionally, although true-unrelated sentences had predictions close to 0.5 (indicating that BERT sometimes judged them to be false), true activation was always higher than false activation.

We obtained the same results for the nSOE model, but not for the remaining models. The nPT model, last-layer model, language model, and GloVe similarity models failed to distinguish between true and false sentences, whereas the first-layer model was insensitive to the semantic relatedness of the concepts and features. We further tested the predictions of the full BERT model by regressing the absolute activation difference ($|A_T - A_F|$) for each sentence on main effects for truth value (true or false), semantic relatedness (related or unrelated), sentence type (category membership or features), as well as the interaction between truth value and relatedness. Absolute activation corresponds to the strength of BERT's beliefs and should be proportional to response time. As expected we observed a significantly positive interaction effect, $\beta = 3.87$, $t(72) = 3.38$, $p < 0.001$, 95% CI = [1.58, 6.16], indicating related true sentences have higher absolute activations and related false sentences have lower absolute activation, and that BERT is able to predict the effect of semantic relatedness on response times.

Effect 4: Anderson and Reder (1974) Experiment 1

Anderson and Reder (1974) also examined false sentences. They proposed that falsifying a sentence involving a concept (e.g., *tent*) and a superordinate category that it does not belong to (e.g., *a metal*) involved first generating a true superordinate category for the concept (e.g., *a dwelling*) and then falsifying the true superordinate category and the original false superordinate category. Thus, a sentence like *tent is a metal* would be falsified by first verifying the sentence *a tent is a dwelling* and then falsifying the sentence *a dwelling is a metal.* In their experiments, Anderson and Reder found that response times for the falsification of a concept on a false superordinate category were predicted both by the response times for the verification of the concept on the true superordinate category and the response times for the superordinate category falsification, providing support for their hypothesis. Anderson and Reder (1974) provided the full set of concepts, as well as false superordinate categories (which they referred to as negative categories) and true superordinate categories (which they referred to as positive categories) in their appendix. There were a total of 82 concepts, with 41 positive and 41 negative categories, spanning a diverse set of domains (including *buildings*, *materials*, *vehicles*, *weapons*, *plants*, *furniture*, and *beverages*). This generated a total of 246 sentences.

Figure 5
Model Predictions for Behavioral Effects



Note. Aggregate model predictions for semantic relatedness effects documented by Glass et al. (1974) (A), falsification effects documented by Anderson and Reder (1974) (B), antonym effects documented by Glass et al. (1979) (C), category versus feature effects documented by Hampton (1984) (D) and by Collins and Quillian (1969) (E), feature correlation effects documented by McRae et al. (1997) (F), feature distinctiveness effects documented by Cree et al. (2006) (G), and set-membership reversals in Roth and Mervis (1983) (H). See the online article for the color version of this figure.

We extracted these sentences from the appendix and attempted to replicate Anderson and Reder's findings using our six models. Specifically, for each target concept, positive category, and negative category, we calculated model predictions for the sentences generated for the concept/negative category pair (i.e., the original sentence, e.g., *tent is a metal*), the concept/positive category pair (i.e., the superordinate category verification, e.g., *tent is a dwelling*), and the positive category/negative category pair (i.e., the superordinate category falsification, e.g., *dwelling is a metal*). Figure 5B shows average model predictions for these three sets. Here we see that the full BERT model was once again able to successfully falsify the concept/negative category and positive category/negative category sentences, and verify the concept/positive category sentences. This was also the case with the nSOE model, the nPT model, and the last-layer model. The language model and GloVe similarity did typically give lower predictions to the concept/negative category sentences and positive category/negative category sentences than the concept/positive category sentences but did not correctly assign the false sentences predictions less than 0.5. The first-layer model again completely failed at capturing the results.

To predict the response time relationships observed in Anderson and Reder (1974) we calculated the absolute value of the full BERT model's activation difference ($|A_T - A_F|$). This gave us a measure of the strength of belief for the model, which should be proportional to response time. We regressed absolute BERT activation difference for the concept/negative category pair on the absolute activation for the concept/positive category pair and the positive category/negative category pair and found a significant positive effect of the two predictor variables, $\beta = 0.34$, $t(82) = 3.57$, $p < 0.001$, 95% CI = [0.15, 0.53] and $\beta = 0.51$, $t(82) = 3.75$, $p < 0.001$, 95% CI = [0.24, 0.78] respectively, replicating Anderson and Reder's results.

Effect 5: Glass et al. (1979) Experiment 1

Glass et al. (1979) examined response times for falsifying false sentences composed of direct and indirect antonyms. In contrast to their prior work (Glass et al., 1974), which showed that semantic relatedness increased response times for false sentences, Glass et al. (1979) found that participants were typically quicker to falsify direct antonyms (e.g., *boy-girl*) than indirect antonyms (e.g., *boy-sister*). We obtained the stimuli used by Glass et al. in Experiment 1 of their article from their Table 1. These stimuli involved 12 direct antonym pairs and 12 indirect antonym pairs, totaling 24 pairs. In all cases stimuli involved gender-based social categories, and antonyms were constructed by pairing male and female categories with each other. Sentences generated using these word pairs were prefixed with *all* and *some* quantifiers. However, as these quantifiers did not influence Glass et al.'s results, we excluded them from our tests. Thus, we tested sentences like *boys are girls* or *boys are sisters* rather than *all boys are girls* or *all boys are sisters*.

We were unable to capture the findings of Glass et al. (1979). Specifically, both the full BERT model and its competitors considered direct antonyms to be more likely to be true than indirect antonyms. This is shown in Figure 5C. Here we can see that BERT generated very strong negative activation for indirect antonyms, but neutral activation for direct antonyms. Thus, although it was able to correctly falsify indirect antonyms, it sometimes considered direct antonyms to be true. This was also the case for the nSOE model, and to a lesser extent, the nPT model. The remaining models additionally failed at classifying the sentences as false. We also regressed absolute BERT activation difference ($|A_T - A_F|$) on a binary variable indicating whether the sentence was a direct or indirect antonym and found a significant positive effect of direct antonym, $\beta = 8.84$, $t(24) = 5.13$, $p < 0.001$, 95% CI = [5.26, 12.41], showing that this model behaved according to the semantic relatedness effect documented by Glass et al. (1974), and was unable to predict the reversal of this effect for antonyms documented by Glass et al. (1979).

Effect 6: Hampton (1984) Experiments 1 and 2, and Collins and Quillian (1969) Experiments 1 and 2

Our sixth test of response times in semantic verification involved Hampton (1984) finding that participants are typically quicker to verify true statements about category membership (e.g., *oak is a tree*) than true statements about features (e.g., *oak is green*) and also typically quicker to falsify false statements about category membership (e.g., *oak is a creature*) than false statements about features (e.g., *oak eats mice*). Hampton presented the stimuli used for these tests in the appendix of his article. These stimuli involved 115 concepts, each with a true superordinate category, a true feature, a false superordinate category, and a false feature. This generated 460 distinct sentences that were used in Experiments 1 and 2 of the article. The properties and features were obtained from an existing feature norms data set and were matched in terms of production frequency. They spanned a large set of domains, including *trees*, *vehicles*, *foods*, *sports*, *weapons*, and *animals*.

Figure 5D shows model predictions for the four types of sentences used in Hampton (1984). Here we see that the full BERT model gave higher probabilities and activations to true category membership sentences than true feature sentences, as well as lower probabilities and activations to false category membership sentences than false feature sentences (though effects appear small for the false sentences as predictions are saturated close to 0). The nSOE model, nPT model, and the last-layer model also captured this effect. The remaining models failed at capturing the full set of effects, and sometimes had difficulties distinguishing true from false sentences.

Interestingly, Collins and Quillian (1969) also found that participants were quicker to verify true statements about category membership than true statements about features but and quicker to falsify false statements about category membership than false statements about features. Even though their stimuli were not designed to rigorously test for this effect, and this effect was not the central focus of their article, we nonetheless tried to replicate the above results using their stimuli. For this purpose, we once again obtained the true and false sentences from Experiments 1 and 2 of Collins and Quillian (1969), presented in their Figure 1 and Table 1, and used in our analysis of Effect 1 above. Our model predictions for these sentences are shown in Figure 5E. Here we once again find that the full BERT model captured the effect by giving true statements about category membership higher probabilities and activations than true statements about features, and giving false statements about category membership lower activations than false statements about features. The nSOE and nPT models closely mimicked these predictions. Although the last-layer model also captured these effects and was able to explain the Hampton (1984) results in Figure 5D, it did not correctly falsify false feature sentences. The remaining models failed to capture these effects and often had difficulties distinguishing true from false.

To test the above effects more rigorously we regressed absolute BERT activation difference ($|A_T - A_F|$) on binary variables indicating whether the sentences involved category membership or feature judgments, with a control for the truth value of the sentence. We found a positive significant effect of category membership for both Hampton (1984) stimuli, $\beta = 0.41$, $t(460) = 2.45$, $p < 0.05$, 95% CI = [0.08, 0.74], and for Collins and Quillian (1969) stimuli, $\beta = 1.39$, $t(151) = 3.47$, $p < 0.01$, 95% CI = [0.60, 2.18].

Effect 7: McRae et al. (1997) Experiment 3

We also examined McRae et al. (1997) finding that feature correlations predict response times in semantic verification. This effect is the basis of a recurrent neural network theory in which feature representations guide semantic cognition. Although McRae et al.'s theory is somewhat different to the transformer networks tested here, it is still possible that our networks reproduce their main results. Indeed, as we show in the Distribution of Features Across Concepts section below, our full BERT model successfully predicts feature correlations documented in prior research, implying that it should be able to capture response time effects attributed to these correlations.

The stimuli we used in our analysis were taken from Appendix C of McRae et al. This appendix contains a set of 37 target features (e.g., *hunted by people*), each associated with a "strong" (e.g., *deer*) and a "weak" (e.g., *duck*) concept. Target features were highly correlated with other features belonging to the strong concept, but weakly or negatively correlated with other features belonging to the weak concept. McRae et al. hypothesized that semantic verification should be quicker for strong concepts relative to weak concepts for a target feature, and in Experiment 3A they successfully documented this result. Additionally, in Experiment 3B, they found that target features were considered to be more typical of strong concepts than weak concepts.

As shown in Figure 5F, we found that the full BERT model was able to successfully reproduce the results of Experiment 3 in McRae et al., by assigning slightly higher activation values to sentences generated by pairing target features with strong concepts compared to sentences generated by pairing target features with weak concepts (the effect also persists for BERT's probability predictions, though predictions are saturated close to 1, and differences are not visible in the figure). The nSOE model, nPT model, last-layer model, and language model also captured these effects, whereas the remaining two models did not. We tested the full model's predictions more formally by regressing BERT activation difference ($A_T - A_F$) on a binary variable for strong versus weak concepts. This regression revealed a significant effect, $\beta = 0.64$, $t(74) = 1.99$, $p < 0.05$, 95% CI = [0.00, 1.27].

Effect 8: Cree et al. (2006) Experiments 1 and 2

Finally, we tested Cree et al. (2006) result, that sentences composed of distinctive features are verified more quickly than sentences composed of shared features. As with the feature correlation effect discussed above, the feature distinctiveness effect is also derived from McRae et al.'s (1997) recurrent network theory.

The stimuli we used in our analysis were taken from Appendix A of Cree et al. This appendix contains a set of 36 concepts (e.g., *ant*), each associated with a "distinctive" feature that is unique to the concept (e.g., *lives in a colony*) and a nondistinctive feature that is "shared" with other concepts (e.g., *lives in the ground*). Cree et al. hypothesized that semantic verification should be faster for distinctive features relative to shared features, and in Experiments 1 and 2 they verified this hypothesis.

As shown in Figure 5G, we found that the full BERT model was able to successfully reproduce the results of Cree et al., by assigning higher activation values to sentences generated by pairing target concepts with distinctive features compared to sentences generated by pairing target concepts with shared features (the effect also persists for BERT's probability predictions, though predictions are saturated close to 1, and differences are not visible in the figure). The nSOE model, nPT model, and GloVe similarity model also captured these effects, whereas the remaining models did not. We tested the full model's predictions more formally by regressing BERT activation difference ($A_T - A_F$) on a binary variable for distinctive versus shared features. This regression revealed a significant effect, $\beta = 0.79$, $t(72) = 3.49$, $p < 0.01$, 95% CI = [0.34, 1.25].

Typicality Ratings for Concepts and Categories

Effect 9: Rosch (1975) Experiment 1

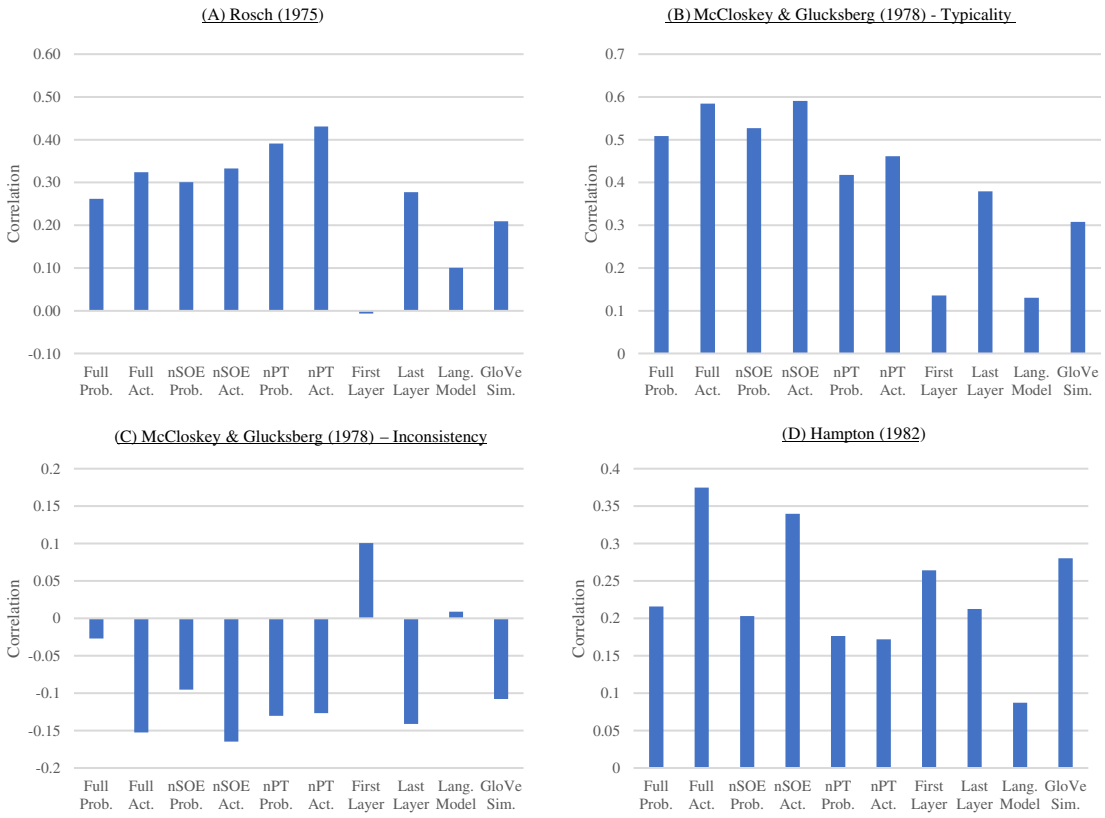
We also attempted to use our models to study typicality effects in conceptual knowledge. It is well known that exemplars vary based on how representative they are of the category, a property of human cognition that can account for some of the semantic relatedness effects discussed above (Glass et al., 1974; Rips et al., 1973; Smith et al., 1974). Although BERT does not explicitly represent concepts in terms of typicality on a superordinate category, we expected that its representations would implicitly reflect typicality relationships, which could, in turn, explain behavioral patterns attributed to typicality. We began by testing our models on the typicality ratings data set collected by Rosch (1975) in Experiment 1 and presented in the appendix of the article. This data set involves 560 goodness-of-example ratings for exemplars taken from 10 semantic categories: *birds*, *carpenter's tools*, *clothing*, *fruit*, *furniture*, *sports*, *toys*, *vegetables*, *vehicles*, and *weapons*. We used these exemplars and categories to generate 560 category membership sentences (e.g., *robin is a bird*, *penguin is a bird*, etc.), which we offered to our BERT model. We correlated its predictions for these sentences with typicality scores (higher score indicates more typical exemplars) for these sentences, separately for each of the 10 categories, and then averaged the correlations across the 10 categories to get a single measure of model performance.

In Figure 6A, we can see that the full BERT model's classification probabilities and activation differences both had positive average correlations with typicality scores, though the latter achieved somewhat better predictions (with an average correlation of 0.32 across the 10 categories). Figure 6A also shows average correlations for other models. Here we see that nSOE closely mimicked our full model's predictions and the nPT model surprisingly exceeded our full model's predictions. We can also see that the last-layer model achieved good correlations and that both the GloVe similarity and language models had moderate correlations, though these were lower than the full model's correlations. Once again, the first-layer model completely failed to predict the results.

Why are the observed correlation rates not higher, given the fact that typicality is a key variable in human semantic cognition that should be easily captured both by simple semantic relatedness models like GloVe and modern extensions like BERT? Here it is useful to note that the above statistics reflect averaged correlations across categories. There is some variability in model accuracy across

Figure 6

Model Predictions for Behavioral Effects



Note. Average model correlations with typicality ratings for categories and exemplars in Rosch (1975) (A) and McCloskey and Glucksberg (1978) (B), average model correlations for inconsistency rates for categories and exemplars in McCloskey and Glucksberg (1978) (C), model correlations for intransitivity rates in Hampton (1982) (D). See the online article for the color version of this figure.

categories. Categories for which predicted correlations were low (such as *sports*, for which we achieved $r = 0.10$) usually have many exemplars that are fairly typical and few that are highly atypical (e.g., *dancing*, *checkers*). Computing correlations on continuous scales (e.g., BERT's output predictions and subject's typicality rating scores) thus lead to weaker correlations. By contrast, categories for which predicted correlations were high (such as *clothing*, for which we achieved $r = 0.70$) usually have a more graded distribution of typicality (e.g., some clothes are highly typical, some are moderately typical, and some are atypical), allowing for higher predictive accuracy rates. Indeed, we found that the category-level correlations obtained from the full model correlated positively ($r = 0.37$) with the standard deviation of the typicality ratings of the items in these categories in Rosch's data. Thus, our model is more accurate for categories that have highly variable typicality ratings.

Effect 10: McCloskey and Glucksberg (1978) Experiment 1

McCloskey and Glucksberg (1978) used typicality ratings to explain participant inconsistencies in category judgments. They found that participants were fairly consistent for sentences involving highly typical exemplars (e.g., *chair is furniture*), as well as for sentences involving unrelated exemplars (e.g., *cucumber is furniture*). However, participants were inconsistent for sentences involving atypical exemplars (e.g., *bookends are furniture*). McCloskey and Glucksberg took their results as evidence that natural categories are fuzzy sets with graded set-membership relations described by typicality, rather than clear boundaries or formal definitions.

The stimuli used in McCloskey and Glucksberg (1978) involved 492 concepts in 18 different superordinate categories (including *birds*, *diseases*, *tools*, *furniture*, *vehicles*, and *weather phenomena*), and the appendix of their article has data for both typicality ratings as well as inconsistency statistics for each exemplar-category pair. We began by first predicting typicality ratings. Here, as with the tests for Rosch (1975), we measured our full BERT model's predictions for sentences composed of the exemplar-category pairs, and computed average correlations with subject ratings for exemplars in the 18 categories. These average correlations are shown in Figure 6B. Again, we see that the model's probability and activations correlated highly with subject ratings, though the latter were slightly higher (average correlation of 0.58 across the 18 categories). The average correlation achieved by the nSOE model was nearly identical, and the correlations for the remaining models were lower. Note that the correlations presented here are higher than those for the Rosch (1975) data, as McCloskey and Glucksberg (1978) included many exemplars that are highly atypical of the category and were judged

by subjects as not belonging to the category. For example, in the *birds* category, McCloskey & Glucksberg included exemplars like *bees* and *chicken eggs*. By contrast, Rosch (1975) did not have such highly atypical exemplars.

We also found that the full BERT model was able to predict participant inconsistencies. To test this, we first calculated absolute BERT probabilities ($|P_T - 0.5|$) and activation differences ($|A_T - A_F|$), and found that both metrics correlated negatively with the proportion of inconsistent participant responses for sentences for most categories. These correlations (averaged across categories) are shown in Figure 6C. Again, activation differences were more predictive, achieving an average correlation of −0.15 for the 18 categories. The nSOE, nPT, last-layer, and GloVe similarity models were able to capture this pattern, whereas the first-layer and language models were not. We suspect that the correlations here were not as high as those for typicality ratings, because McCloskey & Glucksberg's inconsistency data set involved only 30 subjects per ratings. Thus, observed measures of inconsistency on a trial level were likely to be very noisy. This implies that there is lots of unsystematic variance in the data that we are trying to explain, reducing our accuracy rates.

Effect 11: Hampton (1982) Experiments 1 and 2

In related work, Hampton (1982) examined intransitives in category membership judgments. Most theories that propose a hierarchical structure to the organization of semantic memory assume that the relations between sets of concepts involve class inclusion, which necessarily satisfies transitivity. Thus, if *X* is a member of category *Y* and *Y* is a member of category *Z*, *X* must also be a member of category *Z*. However, Hampton found that many concept and superordinate category combinations violated transitivity. For example, participants agreed that *husky is a dog* and *dog is a pet*, but disagreed that *husky is a pet*. He argued that this finding suggests that semantic memory is not hierarchically structured, but rather influenced by typicality and fuzzy set-membership relations, according to which concepts can be typical of their close superordinate categories (direct hypernyms) but not the remote superordinate categories (indirect hypernyms) above them.

Hampton's experiments involved 74 triplets of concept combinations, consisting of a subordinate category, for example, *husky*, a basic category, for example, *dog*, and a superordinate category, for example, *pet*. These categories spanned a large variety of domains, including *vehicles*, *household appliances*, *stones*, *tools*, *weapons*, and *animals*, and were associated with varying levels of transitivity violations in the two experiments. Hampton listed the triplets in the appendix of his article, and we used these triplets to generate pairs of natural language sentences to offer to our models. The first sentence in each pair had the form [SUBORDINATE] *is a* [BASIC] (e.g., *husky is a dog*) and the second had the form [SUBORDINATE] *is a* [SUPERORDINATE] (e.g., *husky is a pet*). Each triplet in Hampton's data was associated with an "intransitivity percentage" statistic describing the proportion of participants violating transitivity for the triplet. Our goal was to predict this statistic. We did this by taking the difference in our models' predictions for the first sentence and the second sentence. Higher differences would indicate that the model is more likely to think [SUBORDINATE] *is a* [BASIC] is true but [SUBORDINATE] *is a* [SUPERORDINATE] is false.

Our tests showed that the full BERT model's probabilities and activations both correlated with intransitivity rates, but that activations once again outperformed probabilities in terms of predictive accuracy (with a correlation of 0.37). This is shown in Figure 6D. We again obtained identical results for the nSOE model. Unlike our previous tests, most of our remaining models also achieved moderate positive correlations.

Effect 12: Roth and Mervis (1983) Experiments 2 and 3

In response to the finding that typicality plays a role in semantic organization, McCloskey and Glucksberg (1978) and many others (Lakoff, 1982; Osherson & Smith, 1981; Zadeh, 1982) advanced "fuzzy set theory," which is a formal approach to describing relations of typicality across categories. One prediction of fuzzy set theory is that the typicality of an exemplar for a remote superordinate category (indirect hypernym) cannot be higher than the typicality of the exemplar for a close superordinate category (direct hypernym). Roth and Mervis showed that this prediction is violated and that there are concepts (e.g., *pie)* that are more typical of remote superordinate categories (e.g., *dessert*) than close superordinates (e.g., *pastry*). In many ways, this is a corollary to Hampton's findings discussed above, in which some concepts are more likely to be seen as members of close superordinate categories than remote superordinates. Overall, Roth and Mervis concluded that fuzzy set theory cannot provide a good account of concept typicality.

Although Roth and Mervis replicated their finding with many different experiments, their article only presented the stimuli used in Experiments 2 and 3 (in Tables 2 and 3 respectively). In these experiments, Roth and Mervis asked participants to give typicality ratings for two types of items on both close and remote superordinate categories. Type 1 items (e.g., *strudel*) were those that were predicted to have higher typicality for the close (e.g., *pastry*) rather than the remote (e.g., *dessert*) superordinate, whereas Type 2 items (e.g., *pie*) were predicted to generate the opposite pattern. The results of Experiments 2 and 3 verified these predictions. There were a total of 12 Type 1 and 2 pairs (e.g., *strudel-pastry*) in each experiment. As each item was rated on two superordinate categories, this generated a total of 96 sentences across the two experiments. The categories included *musical instruments*, *drinks*, *car brands*, *flowers*, *precious stones*, and *fruits.*

We offered the sentences to the full BERT model, whose average probabilities and activations on the Type 1 close, Type 1 remote, Type 2 close and Type 2 remote superordinates are shown in Figure 5H. Here, we can see that BERT generated higher predictions for Type 1 close sentences than Type 1 remote sentences. In contrast, it generated a nearly equivalent predictions for Type 2 close and Type 2 remote sentences. The nSOE and GloVe similarity models also generated these patterns, however, the remaining models were unable to capture observed differences between either Type 1 or Type 2 remote and close sentences.

We tested these differences by regressing BERT activation differences ($A_T - A_F$) for the sentences on binary variables for sentence type (Type 1 vs. 2), superordinate type (close vs. remote) and their interaction. We documented a positive interaction, though we did not achieve significance ($p = 0.34$). These results indicate that our model is able to capture the qualitative trends underlying Roth and Mervis's results but is not able to do so in a statistically reliable manner (we examine the causes of model failures in detail in the discussion section below).

Distribution of Features Across Concepts

Effect 13: Rosch and Mervis (1975) Experiment 1

The distribution of features across concepts has important implications for the organization of semantic memory. Perhaps the most notable empirical result regarding feature distribution is Rosch and Mervis (1975) finding that exemplars that are highly typical of their categories also have a high degree of overlap with the features of other exemplars in the category and a low degree of overlap with the features of exemplars of other categories. This finding explains why some exemplars are more typical of a category than others. Our BERT model, which can make judgments regarding nearly any natural concept and feature pair, is uniquely suited to testing such predictions. We attempted such tests using the approach outlined in Experiment 1 of Rosch and Mervis (1975). In this experiment, Rosch and Mervis asked participants to generate features for a subset of the concepts and superordinate categories used in Rosch (1975). Rosch and Mervis then measured the frequency with which the different features were listed for each category, and subsequently averaged the feature frequencies for a given exemplar to calculate the degree to which that exemplar's features were present in other exemplars for the category and in exemplars in other categories. These were the measures of feature overlap used by Rosch and Mervis.

Due to the generality of our model, we decided not to limit our tests to the subset of concepts used in Rosch and Mervis, but rather implement the tests for all 560 concepts and 10 superordinate categories for which Rosch (1975) collected typicality ratings. Specifically, we first obtained a large list of 4,654 features. These were the features in the Devereux et al. (2014) and McRae et al. (2005) data sets that were listed more than three times by participants. We then paired these features with each of the 560 Rosch (1975) concepts to generate 2,606,240 distinct sentences. These sentences were evaluated by all our models, and their binarized predictions (true or false) were used to calculate measures of feature overlap that were identical to the two used by Rosch and Mervis (1975). Specifically, our first measure took the dot product of the vector of binary features for each concept with the vector of total feature frequencies for other exemplars in its category. This calculated the degree of feature overlap the concept had with members of its superordinate category. Our second measure took the dot product of the vector of binary features for each concept with the vector of total feature frequencies for all concepts in other superordinate categories. This calculated the degree of feature overlap the concept had with nonmembers of its superordinate category. Finally, we correlated these two measures of feature overlap with typicality scores (higher scores for more typical items) for each of the 10 categories in Rosch (1975) and averaged the correlations for each model. We expected to observe positive correlations for feature overlap with category members and negative correlations for feature overlap with nonmembers.

The results of this analysis are shown in Figure 7A. Here, we can see that the full BERT model captured the predicted patterns, with average correlations of 0.45 with the features of members and −0.16 with the features of nonmembers. The nSOE and nPT models also obtained similar correlations (though nPT's correlations were slightly weaker). However, none of the other models were able to capture these patterns.

An alternate approach to evaluating the GloVe model is to use semantic similarity between exemplars, rather than feature overlap, to predict typicality. It could be the case that concepts that are centrally located in the GloVe space relative to other members of their superordinate category are more typical of their superordinate category. This would be consistent with Rosch and Mervis (1975) if we interpret the dimensions of the GloVe space as representing continuous features. To test if GloVe category centrality predicted typicality, we measured the average pairwise cosine similarities of concepts with all other concepts in their superordinate category, as well as with concepts not in their superordinate category. This gave us, for each concept, measures of aggregate GloVe space in-category and out-category centrality. We then correlated these measures with typicality ratings. These correlations are also presented in Figure 7A, which shows that the GloVe centrality approach did yield higher correlations than the GloVe feature overlap approach for feature overlap with category members, but that it could not generate negative correlations for nonmembers.

Effect 14: Malt and Smith (1984) Experiments 1 and 2

Another important result regarding feature distributions was documented by Malt and Smith (1984), who found that certain features were systematically correlated with each other (e.g., *sit on trees* occurred in the same concepts as *flies*, but in different concepts as *is near ocean*, for the category *birds*) and that people were able to perceive and predict these correlations. This result was important as most models of semantic cognition prior to Malt and Smith had assumed that the features of category members were independent of each other.

We attempted to replicate these results using the stimuli from Experiment 1, which were presented in Figure 1F and 1G of Malt and Smith (1984). These figures displayed significantly correlated features from six categories: *birds*, *furniture*, *clothing*, *flowers*, *trees*, and *fruit.* There were a total of 103 positively correlated feature pairs and 41 negatively correlated feature pairs, with an average of 24 pairs per category. We paired each of these features with concepts in the six categories presented in Table 4 of Malt and Smith. This table had a total of 96 concepts with an average of 16 concepts per superordinate category, generating 768 distinct concept–feature pairs. We gave these concept–feature pairs (in the form of sentences) to the full BERT model and then tested the degree to which its predictions for sentences generated using one feature in a pair were correlated with its predictions for sentences generated using the other feature in a pair. For example, as there were 15 different birds in Table 4 of Malt and Smith, this method yielded two sets of 15 sentences for the feature pair (*sits on trees*, *flies*; one sentence for each bird and feature combination). These sentences gave us two 15-dimensional vectors of continuous predictions (activation differences), which were correlated with each other to get a single measure of predicted correlation for (*sits on trees*, *flies*). Intuitively, this approach measures the degree to which *sits on trees* occurs in the same concepts as *flies* for the BERT model.

We applied a similar analysis to the remaining models. However, for the GloVe model, we merely measured the semantic similarity of the bag-of-words GloVe vectors for the feature pairs (e.g., the similarity of the bag-of-words vector for *sits on trees* and the bag-of-words vector for *flies*). This is a reasonable shortcut as the correlation between vectors of distances to two points in a space is proportional to the direct distance between the two points. In other words, we can proxy the degree to which *sits on trees* occurs in the same concepts as *flies* for the GloVe model by simply measuring the distance between *sits on trees* and *flies.*

Figure 7
Model Predictions for Behavioral Effects

Note. Model correlations for feature overlap effects in Rosch and Mervis (1975) (A), feature correlation effects in Malt and Smith (1984) (B), asymmetric and symmetric similarity effects in Whitten et al. (1979) (C), symmetry versus association effects in Hill et al. (2015) (D), and within-category similarity effects in Richie and Bhatia (in press) (E). See the online article for the color version of this figure.

The results of this analysis are shown in Figure 7B. Here, we present average correlations for the full BERT model for feature pairs that were positively and negatively correlated in Malt and Smith. We can see that BERT generated highly positive correlations for positively correlated Malt and Smith features, but neutral correlations for negatively correlated Malt and Smith features. This is also the case for the nSOE model, but not for the remaining models which were able to capture positive correlations for positively correlated features, but also generated positive correlations for negatively correlated features. Interestingly, the first-layer model achieved a correlation close to 1 in all tests. This was because its predictions depended only on the underlying semantic content of the sentence and were thus largely constant regardless of which concepts the features were paired with. Note that the negatively correlated feature for which the full model had the biggest problem was *is in a warm climate* for the *tree* category. The model thought that this feature was positively correlated with other *tree* features (e.g., *has branches*, *is green*, *is shady*), whereas Malt and Smith

documented a negative correlation. When removing the *is in a warm climate* from our analysis, the average predicted correlations are (mildly) negative.

We also tried to replicate the results of Experiment 2 of Malt and Smith (1984). In this experiment, Malt and Smith offered a subset of the feature pairs from Experiment 1 to participants. Participants were asked to rate the extent to which they thought that these pairs went together or were related. Malt and Smith found that 44 of these pairs were rated as positively related and 13 were rated as negatively related by participants. These relationships are displayed in Table 5 of their article. As in Experiment 1, we used the concepts in Table 4 of the article to generate sentences for each positively or negatively related feature pair listed in Table 5 of the article. The full BERT model's predictions for these sentences were then correlated to get a single measure of predicted correlation for each Malt and Smith feature pair. We also used this method to obtain predictions for the remaining models, except for GloVe for which we used the feature similarity analysis outlined above. The results of this analysis are also shown in Figure 7B. As with Experiment 1, the full BERT model was able to give positive correlations to positively related Malt and Smith feature pairs and negative correlations to negatively related Malt and Smith feature pairs. Again, the nSOE model replicated the results, and the remaining models failed to generate negative correlations for negatively correlated features.

The differences shown above are significant according to a regression model in which BERT's correlations (obtained using BERT's output activations) are regressed on a binary variable indicating whether or not the feature pair was positively or negatively correlated in Malt and Smith. This is the case both for Experiment 1, $\beta = 0.53$, $t(144) = 9.03, p < 0.001$, 95% CI = [0.41, 0.64], and Experiment 2, $\beta = 0.99$, $t(57) = 9.61$, $p < 0.001$, 95% CI = [0.91, 1.08].

More recently, McRae et al. (1997) have replicated the findings of Malt and Smith (1984) for a larger set of feature pairs. These findings were the basis of Experiment 3 of McRae et al., analyzed in the Response Times in Semantic Verification section of this article.

Similarity Ratings

Effect 15: Whitten et al. (1979) Experiment 1

It is well known that distributional semantic models predict similarity judgments between pairs of words. BERT, which comes equipped with word vectors in its first set of layers is also likely to achieve similar accuracy rates (though to our knowledge, the specific set of vectors used in the BERT "base" model Wu et al., 2016 have not been tested on semantic similarity judgements). In any case, our goal here is not to evaluate BERT on standard measures of semantic similarity but rather explore its application to similarity ratings phenomena that are particularly problematic for distributional semantic methods like GloVe (at least with the typical application of cosine similarity).

We begin by examining a classic finding in semantic cognition: Asymmetric similarity. Tversky (1977) noted that spatial models of similarity (which include word vector models like GloVe and LSA), when using simple distance measures like cosine or Euclidean distance, predict that similarity ratings are symmetric, that is, that participants believe X to be as similar to *Y* as *Y* is to *X*. Tversky found that these predictions were violated in specific settings, mostly involving countries (e.g., *Red China* and *North Korea*), as well as perceptual stimuli like figures and artificial letters. Whitten et al. (1979) provided a more extensive data set of similarity ratings for natural language concepts. In their experiment, Whitten et al. asked participants to rate the synonymy of 464 word pairs. The words were all nouns, and were taken from Roget's Thesaurus—they were not limited to specific semantic categories, as in prior work. Although each word pair was considered to be a synonym according to the Roget's Thesaurus, participants gave some pairs higher ratings than others. Additionally, participant ratings varied based on the word in the pair that was presented first. Thus, for example, *age* was judged to be a good synonym for *era* but not vice versa.

Whitten et al. presented word pairs and associated ratings in Table 1 of their article. Our goal was to use the full BERT model to predict the asymmetries in this table. We obtained its predictions by asking the model to judge natural language sentences of the form [WORD 1] *is a* [WORD 2] (e.g., *age is an era*) and [WORD 2] *is a* [WORD 1] (e.g., *era is an age*). The activation states outputted by BERT for these sentences were subtracted from each other to obtain a predicted asymmetry rating for the word pairs. This asymmetry rating was compared against Whitten et al.'s *t*-statistic measurement of asymmetry for the word pairs. Note that we used the "is a" relation rather than the "is similar to" relation to construct sentences, as our BERT model was not trained on similarity. By contrast, it had a lot of experience with "is a" (including in the form of tautology sentences like *cat is a cat*).

The correlations between BERT's predictions and the *t*-statistic are shown in Figure 7C. Here, we see that the full BERT model achieved positive correlations (0.24 for probabilities and 0.26 for activations), showing that BERT can capture observed asymmetries in similarity judgment. Unlike our previous tests, the nSOE model performed relatively poorly. This is because predictions of asymmetric similarity likely rely on trained examples of asymmetry on relations (in the subject–object exchange sentences), which were excluded from the nSOE model's training set. We found similarly poor performance for the remaining models, with the GloVe model not predicting any asymmetry whatsoever (due to the symmetric measure of cosine similarity used for the word vectors).

We also examined our models in terms of their ability to predict the symmetrized similarity ratings for the word pairs in Whitten et al. We obtained predictions for the full BERT model by averaging its output activations for the two sentences associated with each word pair (e.g., averaging $A_T - A_F$ for *age is an era* and *era is an age*). GloVe's predictions were simply the cosine similarities of the vectors corresponding to the words in the pair (e.g., cosine similarity of *age* and *era*). This analysis showed that BERT achieved relatively high correlations with the symmetrized similarity ratings data (0.37 and 0.39 for probabilities and activations respectively). The remaining models also performed quite well (though slightly worse than the full model), except for the first-layer model which achieved near-zero correlations.

Effect 16: Hill et al. (2015) Experiment 1

Standard distributional semantics models like GloVe are trained on word co-occurrence statistics. For this reason, their word vectors represent the degree to which different words are semantically related or associated with each other. Although generic semantic relation or association does often correlate with semantic similarity, there are settings in which words can be positively associated but dissimilar (e.g., *cat* and *fur* or *refrigerator* and *food*). To disentangle association

from similarity, Hill et al. (2015) curated a set of 999 word pairs and asked participants to judge these word pairs based on their synonymy or similarity. Participants were specifically instructed to give low ratings to word pairs that were related but dissimilar. The word pairs used in Hill et al. were obtained from the USF free association norms data set (Nelson et al., 2004) and were not restricted to a specific set of semantic categories. Participant ratings for these word pairs were released as part of the SimLex-999 data set.

We used the full BERT model to predict similarity ratings on the SimLex-999 data set. For BERT, we again transformed word pairs into natural language sentences of the form [WORD 1] *is a* [WORD 2] and [WORD 2] *is a* [WORD 1]. We averaged the predictions of the model for the two sentences to get a single prediction for the word pair. We then correlated this prediction with subject similarity ratings. These results are shown in Figure 7D. Here, we see that the full BERT model achieved high correlations (0.47 and 0.51 for probabilities and activations). The nSOE model achieved similar correlations, and nPT model, last-layer model, language model, and GloVe similarity model achieved slightly worse (though still positive) correlations. The first-layer model again failed at capturing the data.

Effect 17: Richie and Bhatia (in press) Experiment 1

Most similarity data sets generate stimuli by selecting word pairs from a larger lexicon, and thus often pair up distinct concepts from different categories. For this reason, these data sets ignore differences in similarity judgment rules across categories. The features that make *robins* similar to *sparrows* are different to those that make *gloves* similar to *scarfs*, and existing similarity data sets do not test the ability of models to account for these differences.

To address the limitations of existing research, we recently examined participant similarity judgments for word pairs with a shared category (Richie & Bhatia, in press). Specifically, we offered participants 2,040 word pairs, each of which involved concepts from one of seven superordinate categories: *furniture*, *clothing*, *birds*, *vegetables*, *sports*, *vehicles*, and *fruit*. Thus, for example, participants were asked to judge the similarity of two types of furniture or two types of clothing or two types of birds. We found that the performance of leading distributional semantics models like GloVe was lower in our data set relative to previous similarity data sets. This was partially due to the fact that the similarity rules used by our participants varied across categories.

In the current article, we wanted to test whether the full BERT model was able to avoid these issues. For this purpose, we applied it to the word pairs in Richie and Bhatia (in press). Again, we obtained BERT predictions by averaging the activations of the [WORD 1] *is a* [WORD 2] and [WORD 2] *is a* [WORD 1] sentences, and correlated these predictions separately for each category. These correlations were subsequently averaged across categories, and are shown in Figure 7E. Here, we can see that BERT achieved high average aggregate correlations (of 0.53 for probabilities and 0.58 for activation). Again, the nSOE model performed equivalently; the nPT model, last-layer model, language model, and GloVe similarity model performed worse (but still yielded positive correlations); and the first-layer model performed poorly.

Discussion

We have tested the full BERT model on stimuli from 25 experiments, spanning 17 distinct behavioral effects in semantic cognition research. Our first set of tests used the probability predictions and differences in activation generated by BERT to describe response times in semantic verification. We found that BERT was able to successfully capture classic level-of-hierarchy effects (Collins & Quillian, 1969), as well as reversals of these effects due to semantic relatedness (Rips et al., 1973; Smith et al., 1974). BERT also predicted response time differences between judgments involving related and unrelated sentences (Glass et al., 1974), between category membership and feature judgments (Collins & Quillian, 1969; Hampton, 1984), between correlated and uncorrelated features (McRae et al., 1997), and between distinctive and nondistinctive features (Cree et al., 2006). It also captured response time patterns for judgments involving false sentences (Anderson & Reder, 1974; Glass et al., 1974). In our second application, we used BERT to predict findings related to typicality. We found that predictions for category membership judgments were proportional to the typicality of concepts in superordinate categories (Rosch, 1975). This allowed BERT to capture inconsistencies across participants (McCloskey & Glucksberg, 1978), transitivity violations (Hampton, 1982), and violations of set-membership relations (Roth & Mervis, 1983) in semantic judgment, previously attributed to typicality. Our third application exploited the fact that BERT can make predictions for arbitrary concept–feature pairs, and thus can be used to measure the distribution of thousands of features across thousands of concepts. Our tests showed that feature overlap in BERT predicted concept typicality (Rosch & Mervis, 1975), and that feature correlations uncovered by BERT matched those obtained in previous experimental data (Malt & Smith, 1984). Our final application showed that BERT was able to predict patterns in similarity judgment that are problematic for existing distributional semantics models. These patterns involve asymmetry in similarity judgment (Whitten et al., 1979), the distinction between association and similarity (Hill et al., 2015), and the measurement of similarity within (rather than across) categories (Richie & Bhatia, in press). Note that BERT's activation differences provided better predictions of semantic verification response times, as well as typicality and similarity ratings, relative to BERT's predicted probabilities. The latter were often saturated at 0 or 1, and were thus not able to capture nuanced patterns in subject data.

The full BERT model outperformed many alternative models in its ability to capture the above patterns. The nSOE model (which is identical to the full model, except that it is not trained on subject–object exchange sentences) mimicked the performance of the full model for all tests except for those involving asymmetries in similarity judgment (Whitten et al., 1979). This again shows that it is necessary to specially train BERT to handle asymmetries in relations and propositions. The nPT model (which is similar to the full BERT model, except that it avoids pretraining the weights using masked word and next sentence prediction) did fairly well on some tests, but performed poorly on others, likely due to its inability to make good predictions for out-of-sample concepts and domains. We also found that the last-layer model (which is similar to the full BERT model except that its sentence representations are not fine-tuned on subject norms) did well on many of the tests, however, there were also some findings that it was unable to capture. This was partially due to its difficulties distinguishing true sentences from false sentences, likely a consequence of it not being fully fine-tuned. This was also the case for the language model and the GloVe similarity model, which use either the PLL of the sentence in the

nonfine-tuned BERT model or the semantic similarity of the concept and feature in the sentence, to make predictions. These models typically performed worse than the last-layer model. Finally, the first-layer model, which relies purely on the semantic content of the sentence, performed very poorly, and often gave sentences in different experimental conditions the same predictions.

These results are similar to those documented in the predictive accuracy section above, and as in that section, show that BERT's successes come from (a) multilayer, interactive, attention-based processing, as well as pretraining using masked word and next sentence prediction, which allow it to appropriately respond to aspects of sentence structure, and (b) the fine-tuning procedure, which adjusts the representation of sentences for the task of semantic verification. This is why the nPT model (which avoided the pretraining step) and the last-layer model (which avoided the full fine-tuning step) performed worse than the full model.

Note that even our full BERT model was unable to capture all previously documented behavioral patterns. In particular, it failed to predict Glass et al. (1979) finding that response times for category membership judgments for direct antonyms are quicker than judgments for indirect antonyms (which itself is a reversal of Glass et al.'s earlier 1974 finding the semantic relatedness slows down judgments for false sentences). Relatedly, BERT gave nonsignificant predictions for Roth and Mervis (1983) set membership typicality violations (though its predictions were in the correct direction). Other limitations of the full BERT model include its inability to give reliable true responses for Level-2 sentence in Collins and Quillian (1969) and in the true-unrelated condition of Glass et al. (1974), its inability to generate reliability negative correlations for negatively associated features in Experiment 1 of Malt and Smith (1984), its weak predictions for the false feature versus category condition of Hampton (1984), and its weak predictions for the Rips et al. (1973) mammals condition, and its weak predictions for the McRae et al. (1997) experiment. These limitations were also present in our alternative models.

What may be the source of these errors? We suspect that the model's most serious failings stem from an overreliance on semantic relatedness (of the concept and feature) to judge the truth of the sentence. True Level-2 propositions in Collins and Quillian (1969) and true propositions in the true-unrelated condition of Glass et al. (1974) involve semantically unrelated concepts and features (e.g., *sharks* and *have skin*), which is why they are sometimes judged by the model to be false. By contrast, false sentences with direct antonyms in Glass et al. (1979) have semantically related concepts and features (e.g., *boys* and *are girls*), which is why they are sometimes judged by the model to be true. Unlike the GloVe similarity model, BERT can go beyond semantic relatedness, but perhaps not enough to give the correct response in all settings. In the general discussion, we examine the implications of this issue for our understanding of human semantic cognition.

Now, there are also other empirical findings for which the model's predictions are somewhat weak. These include, for example, certain conditions in Hampton (1984), McRae et al. (1997), Rips et al. (1973), and Roth and Mervis (1983). We believe that these do not have to do with fundamental limitations of the model but rather subtleties in the empirical data. Here, the model fails because its predictions are saturated close to 1 or 0 (strongly true or strongly false) and it is unable to capture highly nuanced differences across some experimental conditions. For Malt and Smith (1984) negatively correlated features, the model fails only because of a single mistake (i.e., the belief that *trees* that are *in a warm climate* also have other prototypical tree features such as *have branches*, *have bark*, *are green*, *are shady*, etc.—Malt and Smith documented a negative correlation between *in a warm climate* and these features). It is possible that contemporary human subjects generate similar results to BERT, which should be tested in future empirical work. The ability of our model to make quantitative predictions that can be used to motivate theory-driven empirical tests showcases its value for semantic cognition research.

On this note, we would also like to point out that it is quite likely that when testing a large number of independent effects (17 in the current article), one or more effects fail to emerge solely due to chance. This is a property of all statistical tests (both those done on human outputs, as well as those done on model outputs). We believe that our model's occasional mispredictions should not detract from what we believe to be its strong performance and unprecedented empirical scope.

General Discussion

We have trained and tested a transformer network of human conceptual knowledge. Our BERT model can predict what people know about thousands of concepts and features, including concepts and features not in the human data on which it is fine-tuned. For this reason, it can be applied to stimuli from previous empirical research on semantic cognition, and we show that our model reproduces several classical effects. The predictive power and generality of our model are unparalleled. There have been many cognitive theories of semantic representation and retrieval in the past 50 years, yet none are able to describe human responses for arbitrary (linguistically encodable) concepts and features, or quantitatively model 17 key patterns of semantic cognition by predicting responses on existing experimental stimuli sets.

Theoretical Implications

Connectionist Theories of Semantic Cognition

What are the implications of our results for human cognition? At the most basic level, the results of this article offer a type of existence proof—a connectionist model in which "semantic abilities arise from the flow of activation among simple, neuron-like processing units, as governed by the strengths of interconnecting weights" (Rogers & McClelland, 2004, p. 689) is sufficient to predict simple semantic verification and capture many observed patterns of human semantic cognition (even though the training data itself does not "build in" the patterns). Connectionist models have considerable precedent in psychological research (see Rogers & McClelland, 2004, for a review), and the transformers analyzed in the current article are closely related to Hinton (1981, 1986) propositional network. Just like Hinton's model, BERT relies on interactions between distributed word vectors to obtain a vector representation of the sentence. These interactions occur in neural network layers with a deep feed-forward structure (rather than a recurrent structure, as in Hinton's model), and involve a number of new technical innovations such as the self-attention mechanism (which is not present in Hinton's model), however they are nonetheless able to yield

sentence representations that offer the same benefits as those in Hinton's framework.

For example, sentence representations in both transformer networks and in Hinton's network depend on word order; both networks permit different sets of representations for words in different positions. This implies that the resulting sentence vectors can respond to word order and thus capture aspects of linguistic structure, including hierarchies and syntactic dependencies (Jawahar et al., 2019; Linzen & Baroni, 2020; Manning et al., 2020; McClelland et al., 2020; Tenney et al., 2019). Sentence representations can also be generated for a vast set of different sentences (using a finite set of network weights), giving the two models a high degree of generality and a wide domain of applicability. Finally, both transformer networks and Hinton's network generalize learnt information based on the similarity of word vector and sentence vector representations. This is why they are able to accurately extrapolate from a limited training data set to predict semantic judgments for new sentences (as we discuss below, this also seems to be the primary cause of our model's failures).

Learning Processes

Hinton instantiated his model using a small set of words and experimenter-generated weights. However, the BERT model used this article can accurately represent tens of thousands of words, and has pretrained weights based on large-scale natural language data. We further fine-tuned these weights based on psychological data such as concept and feature norms. These pretraining and fine-tuning steps are necessary to obtain good predictions, as illustrated by the relatively poor performance of our six alternative models. For example, our "no pretraining" (or nPT) model was initiated with a random set of weights (except in the first layer) and subsequently fully fine-tuned on the same feature norms data as our full model. This model made reasonable predictions and mimicked observed empirical regularities when the tested concepts were in its fine-tuning data but was unable to extrapolate its learnt knowledge to new concepts and domains. For this reason, it failed in some of our more challenging cross-validation tests and was unable to capture all tested empirical effects. By contrast, our "last-layer" model came with pretrained weights, but these weights were not updated by fine-tuning. This model also performed reasonably in many settings, and unlike the nPT model could handle some out-of-sample concepts and domains. However, its performance fell short of the full model in all cases, mostly due to difficulties distinguishing true sentences from false sentences. Other models, like the first-layer model (which averaged the BERT input word vectors for sentence), the language model (which computed the sentence's pseudolog-likelihood), the GloVe model (which measured the semantic similarity between the concept and feature GloVe vectors) also performed poorly, showing that semantic cognition relies on more than just the thematic content of the sentence, the (language) probability of the sentence, or the similarity of the concept and feature in the sentence.

Unlike many previous connectionist models, our transformer model is not a model of semantic learning. After all, the pretraining corpora has 3.3 billion tokens and we fine-tuned the model on nearly 500,000 examples of true and false sentences. These steps are clearly not developmentally realistic: Children learning (spoken) language only hear on the order of millions, not billions, of word tokens (Golinkoff et al., 2019; Sperry et al., 2019), and are unlikely to be exposed to half a million statements and associated true or false labels. At the same time, we believe that the combination of pretraining and fine-tuning steps necessary to generate successful performance in our model may also be at play in human learning. That is, it does seem reasonable that children develop some conceptual knowledge independently through nonlinguistic information sources (whether innate or by observing the world), further learn linguistic structure through exposure to language, and then, when they can understand sentences, bolster their knowledge when parents or other sources explicitly describe the nature of the world through positive and (implicitly) negative examples (*cats are animals* vs. *whales are not fish*). Ultimately, rich concept representations and knowledge of language structure, combine with knowledge of the goals and properties of the semantic verification task, to allow children to judge new sentences with unknown truth values. Whatever the real algorithms and data sources by which children implement these steps, they arrive at knowledge and abilities that can be described, in many ways, by our full (pretrained and fine-tuned) BERT model.

Semantic Relatedness and Complex Reasoning

It is clear that BERT relies on statistical cues in language to make predictions, and to the extent that there is a correspondence between BERT and human judgment, we would expect humans to also rely on a similar set of cues. We have performed a limited analysis of these cues and have found (perhaps unsurprisingly) that sentences with highly frequently generated features and features that are semantically related to the concepts are more likely to be judged as true. The semantic relatedness cue also seems responsible for the settings in which the model performs poorly. For example, BERT sometimes considers Level-2 propositions in Collins and Quillian (1969) and propositions in the true-unrelated condition of Glass et al. (1974) to be false. Likewise, the model sometimes considers propositions with direct antonyms in Glass et al. (1979) to be true. Thus, it seems that BERT overweighs the semantic relatedness of concepts and features when judging a sentence.

Why do human participants not make these mistakes? There are two possible reasons for this. First, people may also be relying on semantic relatedness (i.e., similarity in concept representation) in a manner similar to BERT, but their concept representations may be more sensitive to perceptual features. This is why people can correctly answer sentences like *sharks have skin*, a Level-2 Collins and Quillian sentence that BERT thinks is false. It is possible that a BERT-like language model, equipped with concept representations from grounded semantics (e.g., images of sharks), could avoid these problems.

Second, people are capable of logical inference and reasoning. BERT is clearly not an appropriate model of complex reasoning, and thus fails in settings where people use reasoning in the service of semantic verification. That said, BERT's outputs could be used in such reasoning processes. For example, we can judge sentences like *sharks have skin* not only by directly querying BERT with the sentence but also by using a graph search in which we first ask BERT to evaluate *sharks are animals*, then ask it to evaluate *animals have skin*, and then, if both sentences are judged to be true, consider *sharks have skin* to be true (see Collins & Loftus, 1975; Collins & Quillian, 1969). Interestingly, BERT does consider both *sharks are animals* and *animals have skin* to be true, which indicates that a hybrid neural-symbolic model that implements graph search on

BERT's outputs could be used to correctly respond to more difficult (Level-2) propositions. We speculate that people may also reason using a similar set of steps. That is, they may use a cognitive process like BERT to judge the intuitive truth of simple sentences, and query such a system in a structured manner to reason about more complex sentences. Such an account is also consistent with dual-process theories of high-level judgment and reasoning, and can also predict human errors, many of which stem from an overreliance on semantic relatedness, or "representativeness" (Evans, 2008; Kahneman, 2011; Sloman, 1996; see also Bhatia, 2017).

Limitations and Future Directions

Other Tasks in Semantic Cognition

While our approach is powerful, we do acknowledge our article has certain limitations. First, we certainly have not tested our model on the full range of semantic cognition and truth-value judgment tasks that we could have, or might be able to. One important task missing from our analysis involves complex concepts composed of adjective–noun combinations such as *red apple* and *brown apple*. There is an extensive literature on this topic, mostly concerned with how features are applied to such complex concepts, and how semantic verification depends (in fairly subtle ways) on the adjective–noun combination used (see, e.g., Murphy, 1988 and Smith & Osherson, 1984, for early results). Our model is trained to predict the features of simple concepts, and we suspect that it can be extended to handle complex concepts as well. In fact, in preliminary tests, we have found that our full BERT model accurately predicts the findings of Smith and Osherson (1984) using their original stimuli. It also manages to generate some of the patterns observed in Murphy (1988).

We have also not attempted to use our model to study tasks involving the use of quantifiers (e.g., *some cats have fur* vs. *all cats have fur*), and in our analysis of Glass et al. (1974) and Glass et al. (1979), have simplified our model inputs to avoid quantifiers. Some preliminary tests (not reported here) have shown that BERT is indeed sensitive to quantifiers, though it typically has difficulty with negating sentences that are not true (e.g., *no cats have fins*). Likewise, this article has not examined the BERT model on tasks involving logical connectives (e.g., *cats have fur and like to hunt mice*), though again preliminary tests have found that BERT has some ability to respond to connectives. We suspect that teaching the model about quantifiers and connectives through appropriate training data could somewhat improve its predictions, though we doubt that this would be a satisfactory solution. Hybrid neural-symbolic systems may provide a more promising approach; that is, it could be possible to use BERT predict the truth values of simple atomic sentences, and then to build a second, high-level module to use the knowledge base for more complex predicate logic.

Finally, we have not attempted to evaluate semantic verification data when other sentences are given as premises, when participants are asked to judge the truth or falsehood of sentences based on information that was presented to them previously (e.g., tasks in which participants read a story before evaluating facts about the story), or when participants are primed with words prior to semantic verification. This would require additional, contextual representations (e.g., of premises, previously presented information, or primes), and could potentially be tested in variants of the BERT model equipped with such representations in additional layers. However, Dasgupta et al. (2020) have recently shown that such an approach has many limitations. Thus, the hybrid neural-symbolic approach may again present a better strategy: BERT can provide a knowledge base for simple concept–feature pairs, which can be combined with a second module that reasons and responds based on additional information.

Interpreting Neural Networks

Second, perhaps even more than their predecessor neural networks and distributional semantic models, transformers like BERT might be criticized as "black boxes" that offer little insight into how they solve a particular problem like sentence verification. We are sympathetic to this concern. As such, we conducted analyses of the factors influencing our full, fine-tuned model's performance (section Correlates of Model Judgment). We also note that there is an emerging cottage industry—sometimes (pejoratively) termed BERTology (Rogers et al., 2020)—attempting to better understand how BERT and other transformers solve natural language processing problems. For example, Jawahar et al. (2019) showed that the intermediate layers of BERT follow a hierarchy of processing not unlike that proposed by many traditional cognitive models of linguistic processing (e.g., Frazier, 1987; for similar findings, see Tenney et al., 2019). Separately, Manning et al. (2020) found that BERT represents aspects of hierarchical linguistic structure such as syntactic dependencies. It is likely that further insights like these could be gleaned from our fine-tuned model, thereby shedding light on why the model is able to reproduce the large set of semantic cognition effects studied in this article. Besides the techniques utilized in the BERTology work referenced above, it may be useful to rely on the methods used by previous neural network modelers of semantic cognition, for example, Rogers and McClelland (2004) and Saxe et al. (2019), to understand the hidden layer representations of our networks. Our analysis in the section Relation-Sensitive Representations is one example of this, and we expect much more is possible in this vein.

Limits of Neural Network Approaches to Semantics

Finally, and perhaps most importantly, there are some concerns that BERT and other deep neural network models are just not up to the task of modeling meaning and reasoning in human-like ways. Some work (e.g., Ettinger, 2020; Ettinger et al., 2018) shows BERT or related models struggling on benchmarks for various semantic tasks like semantic role interpretation (e.g., knowing that *professor* is the agent for *help* in the sentence *the student who is sleeping was helped by the professor*). More seriously, some cognitive scientists view neural network models like BERT as just too dependent on superficial statistics of text, and thus inherently limited in their ability to explain human knowledge and reasoning (Lake & Murphy, in press, Marcus, 2020, or Bender & Koller, 2020). For example, Marcus (2020) shows that GPT-3 (Brown et al., 2020), a very popular, state-of-the-art transformer model that can generate text following a prompt, often displays remarkable lack of conceptual knowledge and often simplistically responds with continuations that are likely given nearby textual context (Marcus, 2020) suggested remedy to this is hybrid neural-symbolic systems, which, as discussed above, may also address some of the limitations of our model). The knowledge capabilities touted by BERT similarly might not reflect actual reasoning, and instead may, to at least

some extent, reflect BERT merely learning superficial patterns in its training data (Kassner et al., 2020).

Given all the above, we certainly are not ready to claim that our fine-tuned BERT network is a complete model of the human computation of meaning and reasoning that underlies sentence verification tasks in their full complexity. But if the foregoing sorts of criticisms are correct, then why does our model work as well as it does in both raw out-of-sample accuracy, and accounting for classic patterns in the literature? We suspect that our fine-tuned BERT model is able to explain so many classic semantic cognition results because these results reflect shallower aspects of knowledge and processing than previously thought, aspects of knowledge that might be reflected in patterns of word co-occurrence and sentence frequency in text. This view is largely consistent with a common perspective in cognitive science (Evans, 2008; Kahneman, 2011; Sloman, 1996), which holds that human judgments often rely on a fast and shallow (and therefore sometimes inaccurate) heuristic system rather than on a slow and effortful deliberative system. It may be the case that much of semantic cognition, and the effects in classic studies that we examined here, have more in common with the former system than previously thought (see also Bhatia, 2017, for a discussion of this point in the context of judgment research). The debate about the validity of deep neural network models as theories of human cognition is ongoing (see Cichy & Kaiser, 2019; Lillicrap et al., 2020; Ma & Peters, 2020; McClelland et al., 2020; Richards et al., 2019; Saxe et al., 2021; and Yarkoni & Westfall, 2017, for recent perspectives). Fully engaging with this debate is out of the scope of the present article, though we believe that it is clear that a complete model of knowledge and reasoning needs both connectionist components, as BERT clearly provides, as well as the symbolic or deliberative components that it might not.

New Applications

Cognitive Process Modeling of Semantic Cognition

The predictive accuracy and generality of BERT offer many practical benefits for semantic cognition research, which do not depend on its plausibility as a theory of human conceptual knowledge. One such benefit involves the cognitive process modeling (Busemeyer & Diederich, 2010; Lewandowsky & Farrell, 2010) of semantic cognition: BERT's outputs can be used to equip existing cognitive process models of binary choice or memory search with the knowledge necessary to predict naturalistic semantic verification and knowledge retrieval. Process models equipped with this world knowledge can then be applied to a wide range of everyday semantic cognition tasks. With parametric model fitting, researchers can recover parameters that describe individual-level semantic mechanisms and study the variability in these mechanisms across individuals and the sensitivity of these mechanisms to experimental manipulations and task-relevant variables.

For semantic verification, for example, we can quite simply use the output activation states of our trained BERT model to quantify the strength of evidence favoring true or false in response to a given concept–feature pair. Strength of evidence can subsequently be fed as a drift rate into dynamic decision models of two-alternative-forced-choice (Brown & Heathcote, 2008; Ratcliff, 1978; Usher & McClelland, 2001), thereby predicting response times in semantic verification (but see Ratcliff & McKoon, 1982, who provide an early test of the feasibility of this approach and find mixed results). Other task-relevant variables (e.g., sentence length, word frequency, time pressure) can also be incorporated into this framework, and the semantic similarity between the concept and feature can be used to specify response biases, thereby instantiating dual-process theories of cognition (which propose that responses are a product of both a fast associative process and a slow deliberative process—see Bhatia, 2017; Evans, 2008; Kahneman, 2011; Sloman, 1996). Using such an approach, model predictions can be obtained for thousands of concepts and features, allowing for a quantitative cognitive process model of semantic verification with a very large domain of applicability.

We can also apply a variant of this approach to memory modeling. This would first require a prespecified set of features and concepts over which memory processes can be assumed to operate. We can obtain this set from existing norms data and expand the set by algorithmically creating new features based on existing features. Thus, features like *hunts mice* can first be parsed into a "feature type" like *hunts* [ANIMAL], and then be used to generate new features like *hunts rabbits* and *hunts lions*, based on sister terms for *mice* in lexical resources like WordNet (Miller, 1995). Our trained BERT model can be given sentences generated from millions of algorithmically generated features to derive the universe of features that are reasonably true for the concept. Finally, cognitive process models of memory (e.g., Polyn et al., 2009; Raaijmakers & Shiffrin, 1981; see also McRae et al., 1997) can be equipped with model-derived representations for these features, in order to quantitatively characterize retrieval processes in feature norm generation studies. Evaluating the feasibility of such models is a promising avenue for future research.

Models of Real-World Cognition

There are many successful theories of categorization, judgment, decision-making, and reasoning, but these are typically tested using abstracted stimuli involving small sets of experimenter-defined features and relations. Occasionally, researchers may use experimental techniques like similarity ratings to uncover concept representations for more naturalistic variants of these tasks (e.g., Nosofsky et al., 2018). However, models trained in these experiments can only be applied within a narrow domain, and cannot be used to make quantitative predictions for most common naturalistic concepts (for which additional experimental data have not been collected). There are likewise few computational models capable of making quantitative predictions for naturalistic cognitive tasks involving explanation, argumentation, moral judgment, and creativity, as responses in these tasks are nearly always based on rich real-world knowledge that is difficult to formally model.

The transformer networks analyzed in this article again offer a solution to this problem. As with the cognitive process models discussed in the previous section, we can use BERT's outputs to equip theories in other domains of cognitive science with the world knowledge necessary for successful deliberation and response in naturalistic tasks. Thus, for example, we can use our fine-tuned BERT model applied to both human-generated and algorithmically generated features to derive the universe of features that are reasonably true for a set of concepts. These features can then be used to model the learning of new categories for naturalistic concepts, probability judgment for events involving concepts, and preferential choice between various objects and concepts (see Bhatia et al., 2019, for a discussion).

One could also, in principle, apply such an approach to model naturalistic high-level cognition. Thus, for example, existing models of analogical reasoning such as BART (Lu et al., 2012, 2019) could be given our model's fine-tuned vector representations of sentences (made of relations and entity pairs) as inputs for further relational reasoning. Likewise, cognitive architectures like ACT-R (Anderson, 1990) could be equipped with facts about the world derived from our trained BERT model for use alongside production rules and other cognitive operations. Similarly, Bayesian models (Oaksford & Chater, 2007) could be given the world knowledge to aid probabilistic reasoning and inference for thousands of concepts and features. It may also be possible to directly embed the representations possessed by our BERT model into hybrid neural-connectionist models (e.g., Doumas et al., 2008; Hummel & Holyoak, 2003), thereby combining the expressiveness and productivity of symbolic models with BERT's ability to capture learning and gradations of similarity and meaning.

A complete theory of the full range of naturalistic high-level cognition will surely be extraordinarily complex—indeed, solving this problem is likely to be as hard as solving the problem of general human intelligence. We therefore believe that such a theory will likely have need for both connectionist principles as embodied in our fine-tuned BERT model, and symbolic and structured principles as embodied in other theories of high-level cognition. Importantly, such theories would benefit from the large scope and generality of the knowledge contained in our fine-tuned BERT model, and therefore be able to reason over a much broader set of concepts than previously possible, as well as uncover new reasoning rules in a data-driven manner. Synthesizing and reconciling diverse approaches to human semantic cognition, by equipping existing theories with rich world knowledge, is likely to be a fruitful area of research for years to come.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Lawrence Erlbaum.

Anderson, J. R., & Reder, L. M. (1974). Negative judgments in and about semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *13*(6), 664–681. https://doi.org/10.1016/S0022-5371(74)80054-X

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, *116*(3), 463–498. https://doi.org/10.1037/a0016261

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. In S. Padó & R. Huang (Eds.), *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. 3606–3611). Association for Computational Linguistics.

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5185–5198). Association for Computational Linguistics.

Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological Review*, *124*(1), 1–20. https://doi.org/10.1037/rev0000047

Bhatia, S. (2019). Predicting risk perception: New insights from data science. *Management Science*, *65*(8), 3800–3823. https://doi.org/10.1287/mnsc.2018.3121

Bhatia, S., Olivola, C., Bhatia, N., & Ameen, A. (in press). Predicting leadership perception with large-scale natural language data. *The Leadership Quarterly*.

Bhatia, S., Richie, R., & Zou, W. (2019). Distributed semantic representations for modelling human judgment. *Current Opinion in Behavioral Sciences*, *29*, 31–36. https://doi.org/10.1016/j.cobeha.2019.01.020

Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., & Choi, Y. (2019). COMET: Commonsense transformers for automatic knowledge graph construction. In P. Nakov & A. Palmer (Eds.), *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 4762–4779). Association for Computational Linguistics.

Bouraoui, Z., Camacho-Collados, J., & Schockaert, S. (2020). Inducing relational knowledge from BERT. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(5), 7456–7463. https://doi.org/10.1609/aaai.v34i05.6242

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178. https://doi.org/10.1016/j.cogpsych.2007.12.002

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., & Agarwal, S. (2020). *Language models are few-shot learners* [Conference session]. In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems. 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Busemeyer, J. R., & Diederich, A. (2010). *Cognitive modeling*. Sage Pubications.

Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Strope, B., & Kurzweil, R. (2018). Universal sentence encoder for English. In E. Blanco & W. Lu (Eds.) *Proceedings of the 2018 conference on empirical methods in natural language processing: System demonstrations* (pp. 169–174). Association for Computational Linguistics.

Cichy, R. M., & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in Cognitive Sciences*, *23*(4), 305–317. https://doi.org/10.1016/j.tics.2019.01.009

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428. https://doi.org/10.1037/0033-295X.82.6.407

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*(2), 240–247. https://doi.org/10.1016/S0022-5371(69)80069-1

Cree, G. S., McNorgan, C., & McRae, K. (2006). Distinctive features hold a privileged status in the computation of word meaning: Implications for theories of semantic memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(4), 643–658. https://doi.org/10.1037/0278-7393.32.4.643

Dai, Z., Yang, Z., Yang, Y., Carbonell, J. G., Le, Q., & Salakhutdinov, R. (2019). Transformer-XL: Attentive language models beyond a fixed-length context. In P. Nakov & A. Palmer (Eds.), *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 2978–2988). Association for Computational Linguistics.

Dasgupta, I., Guo, D., Gershman, S. J., & Goodman, N. D. (2020). Analyzing machine-learned representations: A natural language case study. *Cognitive Science*, *44*(12), Article e12925. https://doi.org/10.1111/cogs.12925

Derby, S., Miller, P., & Devereux, B. (2019). Feature2Vec: Distributional semantic modelling of human property knowledge. In S. Padó & R. Huang (Eds.), *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. 5856−5862). Association for Computational Linguistics.

Devereux, B. J., Tyler, L. K., Geertzen, J., & Randall, B. (2014). The Centre for Speech, Language and the Brain (CSLB) concept property norms. *Behavior Research Methods*, *46*(4), 1119–1127. https://doi.org/10.3758/s13428-013-0420-4

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of North American chapter of the association for computational linguistics: Human language technologies* (pp. 4171–4186). Association for Computational Linguistics.

Doumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, *115*(1), 1–43. https://doi.org/10.1037/0033-295X.115.1.1

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. https://doi.org/10.1162/tacl_a_00298

Ettinger, A., Elgohary, A., Phillips, C., & Resnik, P. (2018). Assessing composition in sentence vector representations. In E. Bender, L. Derczynski, & P. Isabelle (Eds.), *Proceedings of the 27th international conference on computational linguistics* (pp. 1790–1801). Association for Computational Linguistics.

Evans, J. S. B. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*, 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Făgărăşan, L., Vecchi, E. M., & Clark, S. (2015). From distributional semantics to feature norms: Grounding semantic models in human perceptual data. In M. Purver, M. Sadrzadeh, & M. Stone (Eds.), *Proceedings of the 11th international conference on computational semantics* (pp. 52–57). Association for Computational Linguistics.

Frazier, L. (1987). Sentence processing: A tutorial review. In M. Coltheart (Ed.), *Attention and performance XII* (pp. 559–585). Lawrence Erlbaum.

Glass, A. L., Holyoak, K. J., & Kiger, J. I. (1979). Role of antonymy relations in semantic judgments. *Journal of Experimental Psychology: Human Learning and Memory*, *5*(6), 598–606. https://doi.org/10.1037/0278-7393.5.6.598

Glass, A. L., Holyoak, K. J., & O'Dell, C. (1974). Production frequency and the verification of quantified statements. *Journal of Verbal Learning and Verbal Behavior*, *13*(3), 237–254. https://doi.org/10.1016/S0022-5371(74)80061-7

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, *109*(1), 75–90. https://doi.org/10.1037/0033-295X.109.1.75

Golinkoff, R. M., Hoff, E., Rowe, M. L., Tamis-LeMonda, C. S., & Hirsh-Pasek, K. (2019). Language matters: Denying the existence of the 30-million-word gap has serious consequences. *Child Development*, *90*(3), 985–992. https://doi.org/10.1111/cdev.13128

Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2018). Semantic projection: Predicting high-level judgment recovering human knowledge of multiple, distinct object features from word embeddings. arXiv preprint arXiv:1802.01241.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244. https://doi.org/10.1037/0033-295X.114.2.211

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, *14*(6), 1006–1033. https://doi.org/10.1177/1745691619861372

Hampton, J. A. (1982). A demonstration of intransitivity in natural categories. *Cognition*, *12*(2), 151–164. https://doi.org/10.1016/0010-0277(82)90010-5

Hampton, J. A. (1984). The verification of category and property statements. *Memory & Cognition*, *12*(4), 345–354. https://doi.org/10.3758/BF03198294

Hertwig, R., Herzog, S. M., Schooler, L. J., & Reimer, T. (2008). Fluency heuristic: A model of how the mind exploits a by-product of information retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(5), 1191–1206. https://doi.org/10.1037/a0013025

Hill, F., Reichart, R., & Korhonen, A. (2015). SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665–695. https://doi.org/10.1162/COLI_a_00237

Hinton, G. E. (1981). Implementing semantic networks in parallel hardware. In G. E. Hinton & J. A. Anderson (Eds.), *Parallel models of associative memory* (pp. 161–187). Lawrence Erlbaum.

Hinton, G. E. (1986). Learning distributed representations of concepts. In R. G. M. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neurobiology* (pp. 46–61). Clarendon Press.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, *110*(2), 220–264. https://doi.org/10.1037/0033-295X.110.2.220

Jawahar, G., Sagot, B., & Seddah, D. (2019). What does BERT learn about the structure of language? In P. Nakov & A. Palmer (Eds.), *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 3651–3657). Association for Computational Linguistics.

Jones, M. N., Gruenenfelder, T. M., & Recchia, G. (2018). In defense of spatial models of semantic representation. *New Ideas in Psychology*, *50*(4), 54–60. https://doi.org/10.1016/j.newideapsych.2017.08.001

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, *55*(4), 534–552. https://doi.org/10.1016/j.jml.2006.07.003

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*(1), 1–37. https://doi.org/10.1037/0033-295X.114.1.1

Jones, M. N., Willits, J. A., & Dennis, S. (2015). Models of semantic memory. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *Oxford handbook of computational and mathematical psychology* (pp. 232–254). Oxford University Press.

Kahneman, D. (2011). *Thinking, fast and slow*. Penguin Books.

Kassner, N., Krojer, B., & Schütze, H. (2020). Are pretrained language models symbolic reasoners over knowledge? In R. Fernández & T. Linzen (Eds.) *Proceedings of the 24th conference on computational natural language learning* (pp. 552–564). Association for Computational Linguistics.

Lake, B. M., & Murphy, G. L. (in press). Word meaning in minds and machines. *Psychological Review*.

Lakoff, G. (1982). *Categories and cognitive models*. Cognitive Science Program, Institute of Cognitive Studies, University of California at Berkeley.

Landauer, T. K., & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics*, *4*, 151–171. https://doi.org/10.1146/annurev-linguistics-030514-125254

Lewandowsky, S., & Farrell, S. (2010). *Computational modeling in cognition: Principles and practice*. SAGE Publications.

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., & Hinton, G. (2020). Backpropagation and the brain. *Nature Reviews Neuroscience*, *21*(6), 335–346. https://doi.org/10.1038/s41583-020-0277-3

Linzen, T., & Baroni, M. (2020). Syntactic structure from deep learning. *Annual Review of Linguistics*, *7*, 195–212. https://doi.org/10.1146/annurev-linguistics-032020-051035

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.

Loftus, E. F., & Scheff, R. W. (1971). Categorization norms for fifty representative instances. *Journal of Experimental Psychology*, *91*(2), 355–365. https://doi.org/10.1037/h0031944

Lu, H., Chen, D., & Holyoak, K. J. (2012). Bayesian analogy with relational transformations. *Psychological Review*, *119*(3), 617–648. https://doi.org/10.1037/a0028719

Lu, H., Wu, Y. N., & Holyoak, K. J. (2019). Emergence of analogy from relation learning. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(10), 4176–4181. https://doi.org/10.1073/pnas.1814779116

Ma, W. J., & Peters, B. (2020). *A neural network walks into a lab: Towards using deep nets as models for human behavior*. arXiv preprint. https://arxiv.org/abs/2005.02181

Malt, B. C., & Smith, E. E. (1984). Correlated properties in natural categories. *Journal of Verbal Learning and Verbal Behavior*, *23*(2), 250–269. https://doi.org/10.1016/S0022-5371(84)90170-1

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. https://doi.org/10.1016/j.jml.2016.04.001

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(48), 30046–30054. https://doi.org/10.1073/pnas.1907367117

Marcus, G. (2020). *The next decade in AI: Four steps towards robust artificial intelligence*. arXiv preprint arXiv:2002.06177. https://arxiv.org/abs/2002.06177

McClelland, J. L., Hill, F., Rudolph, M., Baldridge, J., & Schütze, H. (2020). Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(42), 25966–25974. https://doi.org/10.1073/pnas.1910416117

McClelland, J. L., & Rumelhart, D. E., & the PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2: Psychological and Biological Models* (ISBN 978-0262631105). MIT Press.

McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition*, *6*(4), 462–472. https://doi.org/10.3758/BF03197480

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*(4), 547–559. https://doi.org/10.3758/BF03192726

McRae, K., de Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, *126*(2), 99–130. https://doi.org/10.1037/0096-3445.126.2.99

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality* [Conference session]. Advances in neural information processing systems.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41. https://doi.org/10.1145/219717.219748

Murphy, G. L. (1988). Comprehending complex concepts. *Cognitive Science*, *12*(4), 529–562. https://doi.org/10.1207/s15516709cog1204_2

Murphy, G. L. (2004). *The big book of concepts*. MIT Press.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407. https://doi.org/10.3758/BF03195588

Nematzadeh, A., Meylan, S. C., & Griffiths, T. L. (2017). Evaluating vector-space models of word representation, or, the unreasonable effectiveness of counting words near other words. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th annual conference of the cognitive science society* (pp. 859–864). Lawrence Erlbaum.

Nosofsky, R. M., Sanders, C. A., & McDaniel, M. A. (2018). Tests of an exemplar-memory model of classification learning in a high-dimensional natural-science category domain. *Journal of Experimental Psychology: General*, *147*(3), 328–353. https://doi.org/10.1037/xge0000369

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198524496.001.0001

Osherson, D. N., & Smith, E. E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition*, *9*(1), 35–58. https://doi.org/10.1016/0010-0277(81)90013-5

Padó, U., Crocker, M. W., & Keller, F. (2009). A probabilistic model of semantic plausibility in sentence processing. *Cognitive Science*, *33*(5), 794–838. https://doi.org/10.1111/j.1551-6709.2009.01033.x

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1532–1543). Association for Computational Linguistics.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., & Miller, A. (2019). Language models as knowledge bases? In S. Padó & R. Huang (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* (pp. 2463–2473). Association for Computational Linguistics.

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, *116*(1), 129–156. https://doi.org/10.1037/a0014420

Porada, I., Suleman, K., & Cheung, J. C. K. (2019). Can a gorilla ride a camel? Learning semantic plausibility from text. In S. Ostermann, S. Zhang, M. Roth, & P. Clark (Eds.), *Proceedings of the first workshop on commonsense inference in natural language processing* (pp. 123–129). Association for Computational Linguistics.

Raaijmakers, J. G., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*(2), 93–134. https://doi.org/10.1037/0033-295X.88.2.93

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*. https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Ratcliff, R., & McKoon, G. (1982). Speed and accuracy in the processing of false statements about semantic information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*(1), 16–36. https://doi.org/10.1037/0278-7393.8.1.16

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., … Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, *22*(11), 1761–1770. https://doi.org/10.1038/s41593-019-0520-2

Richie, R., & Bhatia, S. (in press). Similarity judgment within and across categories: A comprehensive model comparison. *Cognitive Science*. https://doi.org/10.31234/osf.io/5pa9r

Richie, R., Zou, W., & Bhatia, S. (2019). Predicting high-level human judgment across diverse behavioral domains. *Collabra. Psychology*, *5*(1), Article 50. https://doi.org/10.1525/collabra.282

Rips, L. J., Shoben, E. J., & Smith, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior*, *12*(1), 1–20. https://doi.org/10.1016/S0022-5371(73)80056-8

Rips, L. J., Smith, E. E., & Medin, D. L. (2012). Concepts and categories: Memory, meaning, and metaphysics. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 177–209). Oxford University Press.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, *8*, 842–866. https://doi.org/10.1162/tacl_a_00349

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. MIT Press. https://doi.org/10.7551/mitpress/6161.001.0001

Rogers, T. T., & McClelland, J. L. (2008). Précis of semantic cognition: A parallel distributed processing approach. *Behavioral and Brain Sciences*, *31*(6), 689–714. https://doi.org/10.1017/S0140525X0800589X

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104*(3), 192–233. https://doi.org/10.1037/0096-3445.104.3.192

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605. https://doi.org/10.1016/0010-0285(75)90024-9

Roth, E. M., & Mervis, C. B. (1983). Fuzzy set theory and class inclusion relations in semantic categories. *Journal of Verbal Learning and Verbal Behavior*, *22*(5), 509–525. https://doi.org/10.1016/S0022-5371(83)90310-9

Rumelhart, D. E., & McClelland, J. L., & the PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations* (ISBN 978-0262680530). MIT Press. https://doi.org/10.7551/mitpress/5236.001.0001

Rumelhart, D. E., & Todd, P. M. (1993). Learning and connectionist representations. In D. E. Meyer & S. Kornblum (Eds.), *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience* (pp. 3–30). MIT Press.

Salazar, J., Liang, D., Nguyen, T. Q., & Kirchhoff, K. (2020). Masked language model scoring. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 2699–2712). Association for Computational Linguistics.

Saxe, A., Nelli, S., & Summerfield, C. (2021). If deep learning is the answer, what is the question? *Nature Reviews Neuroscience*, *22*, 55–67. https://doi.org/10.1038/s41583-020-00395-8

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(23), 11537–11546. https://doi.org/10.1073/pnas.1820226116

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3–22. https://doi.org/10.1037/0033-2909.119.1.3

Smith, E. E., & Osherson, D. N. (1984). Conceptual combination with prototype concepts. *Cognitive Science*, *8*(4), 337–361. https://doi.org/10.1207/s15516709cog0804_2

Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, *81*(3), 214–241. https://doi.org/10.1037/h0036351

Sperry, D. E., Sperry, L. L., & Miller, P. J. (2019). Reexamining the verbal environments of children from different socioeconomic backgrounds. *Child Development*, *90*(4), 1303–1318. https://doi.org/10.1111/cdev.13072

Steyvers, M. (2010). Combining feature norms and text data with topic models. *Acta Psychologica*, *133*(3), 234–243. https://doi.org/10.1016/j.actpsy.2009.10.010

Tenney, I., Das, D., & Pavlick, E. (2019). BERT Rediscovers the classical NLP pipeline. In In P. Nakov & A. Palmer (Eds.) *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 4593–4601). Association for Computational Linguistics.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352. https://doi.org/10.1037/0033-295X.84.4.327

Unger, L., & Fisher, A. V. (2021). The emergence of richly organized semantic knowledge from simple statistics: A synthetic review. *Developmental Review*, *60*, Article 100949. https://doi.org/10.1016/j.dr.2021.100949

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*(3), 550–592. https://doi.org/10.1037/0033-295X.108.3.550

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language*, *50*(3), 289–335. https://doi.org/10.1016/j.jml.2003.10.003

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In U. von Luxburg, I. Guyon, & S. Bengio (Eds.) *Proceedings of the 31st international conference on neural information processing systems* (pp. 6000–6010). Curran Associates.

Whitten, W. B., II, Suter, W. N., & Frank, M. L. (1979). Bidirectional synonym ratings of 464 noun pairs. *Journal of Verbal Learning and Verbal Behavior*, *18*(1), 109–127. https://doi.org/10.1016/S0022-5371(79)90604-2

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, J., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Brew, J. (2019). *HuggingFace's transformers: State-of-the-art natural language processing*. ArXiv, arXiv-1910.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., … Dean, J. (2016). *Google's neural machine translation system: Bridging the gap between human and machine translation*. arXiv preprint arXiv:1609.08144.

Yao, L., Mao, C., & Luo, Y. (2019). KG-BERT: BERT for knowledge graph completion. arXiv preprint arXiv:1909.03193.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Zadeh, L. A. (1982). A note on prototype theory and fuzzy sets. *Cognition*, *12*(3), 291–297. https://doi.org/10.1016/0010-0277(82)90036-1

Zhang, Z., Wu, Y., Zhao, H., Li, Z., Zhang, S., Zhou, X., & Zhou, X. (2020). Semantics-aware BERT for language understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(5), 9628–9635. https://doi.org/10.1609/aaai.v34i05.6510

Zou, W., & Bhatia, S. (2021). Judgment errors in naturalistic numerical estimation. *Cognition*, *211*, Article 104647. https://doi.org/10.1016/j.cognition.2021.104647

Received November 13, 2020
Revision received June 18, 2021
Accepted June 19, 2021 ■