BLiMP: The Benchmark of Linguistic Minimal Pairs for English

Alex Warstadt[1], Alicia Parrish[1], Haokun Liu[2], Anhad Mohananey[2], Wei Peng[2], Sheng-FuWang[1], Samuel R. Bowman[1,2,3]

[1]Department of Linguistics New York University [2]Department of Computer Science New York University [3]Center for Data Science New York University

{warstadt,alicia.v.parrish,haokunliu,anhad, weipeng,shengfu.wang,bowman}@nyu.edu

Abstract

We introduce The Benchmark of Linguistic Minimal Pairs (BLiMP),[1] a challenge set for evaluating the linguistic knowledge of language models (LMs) on major grammatical phenomena in English. BLiMP consists of 67 individual datasets, each containing 1,000 minimal pairs—that is, pairs of minimally different sentences that contrast in grammatical acceptability and isolate specific phenomenon in syntax, morphology, or semantics. We generate the data according to linguist-crafted grammar templates, and human aggregate agreement with the labels is 96.4%. We evaluate n-gram, LSTM, and Transformer (GPT-2 and Transformer-XL) LMs by observing whether they assign a higher probability to the acceptable sentence in each minimal pair. We find that state-of-the-art models identify morphological contrasts related to agreement reliably, but they struggle with some subtle semantic and syntactic phenomena, such as negative polarity items and extraction islands.

1 Introduction

Current neural networks for sentence processing rely on unsupervised pretraining tasks like language modeling. Still, it is an open question how the linguistic knowledge of state-of-the-art language models (LMs) varies across the linguistic phenomena of English. Recent studies (e.g., Linzen et al., 2016; Marvin and Linzen, 2018; Wilcox et al., 2018) have explored this question by evaluating LMs' preferences between *minimal pairs* of sentences differing in grammatical acceptability, as in Example 1. However, each of these studies uses a different set of metrics, and focuses on a small set of linguistic paradigms, severely limiting any possible big-picture conclusions.

(1) a. The cats annoy Tim. (*grammatical*)
b. *The cats annoys Tim. (*ungrammatical*)

We introduce the Benchmark of Linguistic Minimal Pairs (shortened to BLiMP), a linguistically motivated benchmark for assessing the sensitivity of LMs to acceptability contrasts across a wide range of English phenomena, covering both previously studied and novel contrasts. BLiMP consists of 67 datasets automatically generated from linguist-crafted grammar templates, each containing 1,000 minimal pairs and organized by phenomenon into 12 categories. Validation with crowdworkers shows that BLiMP faithfully represents human preferences.

We use BLiMP to study several pretrained LMs: Transformer-based LMs GPT-2 (Radford et al., 2019) and Transformer-XL (Dai et al., 2019), an LSTM LM trained by Gulordava et al. (2019), and an n-gram LM. We evaluate whether the LM assigns a higher probability to the acceptable sentence in each minimal pair to determine which grammatical distinctions LMs are sensitive to. This gives us indirect evidence about each model's linguistic knowledge and allows us to compare models in a fine-grained way. We conclude that current neural LMs appear to acquire robust knowledge of morphological agreement and some syntactic phenomena such as ellipsis and control/ raising. They show weaker evidence of knowledge about argument structure, negative polarity item licensing, and the semantic properties of quantifiers. All models perform at or near chance on extraction islands. Overall, every model we

[1] https://github.com/alexwarstadt/blimp.

Transactions of the Association for Computational Linguistics, vol. 8, pp. 377–392, 2020. https://doi.org/10.1162/tacl_a_00321
Action Editor: Mark Steedman. Submission batch: 1/2020; Revision batch: 3/2020; Published 7/2020.

evaluate falls short of human performance by a wide margin. GPT-2, which performs the best, performs 8 points below humans overall, though it does match or exceed human performance on specific phenomena.

In §6.3 we conduct additional experiments to investigate the effect of training size on the LSTM LM and Transformer-XL's performance on BLiMP. Although we see steady improvements in overall performance, we find that LMs learn phenomenon-specific distinctions at different rates. In §6.4 we consider alternative well-motivated evaluation metrics on BLiMP, but find that they do not differ drastically from our method of comparing LM probabilities for full sentences.

We conclude that whereas models like GPT-2 appear to have significant linguistic knowledge, this knowledge is concentrated in some specific domains of English grammar. We use BLiMP to uncover several linguistic phenomena where even state-of-the-art language models clearly lack human-like knowledge, and to bring into focus those areas of grammar that future studies evaluating LMs should investigate in greater depth.

2 Background and Related Work

2.1 Language Models

The objective of a language model is to give a probability distribution over the strings of a language. Both neural network and non-neural network architectures are used to build LMs, and neural models can be trained in a *self-supervised* setting without the need for labeled data. Recently, variants of neural language modeling have been shown to be a strong pretraining task for natural language processing tasks (Howard and Ruder, 2018; Peters et al., 2018; Radford et al., 2018; Devlin et al., 2019).

The last decade has seen two major paradigm shifts in the state of the art for language modeling. First, there was a movement from models based on local n-gram statistics (see Chen and Goodman, 1999) to neural sequence models such as LSTMs (Mikolov et al., 2010), which optimize on the task of predicting the next token. Subsequently, Transformer-based architectures employing self-attention (Vaswani et al., 2017) have outperformed LSTMs (e.g., Dai et al., 2019). Although these shifts have resulted in stronger LMs, perplexity on large benchmark datasets like WikiText-103 (Merity et al., 2016) has remained the primary performance metric, which cannot give detailed insight into these models' knowledge of grammar. Evaluation on benchmarks like GLUE (Wang et al., 2018, 2019a), which heavily adapt language models to perform downstream tasks, is more informative, but doesn't offer broad coverage of linguistic phenomena, and doesn't necessary reflect knowledge that is already present in the LMs.

2.2 Linguistic Knowledge of NNs

Many recent studies have searched for evidence that neural networks (NNs) learn representations that implicitly encode grammatical concepts. We refer to the ability to encode these concepts as *linguistic knowledge*. Some studies evaluate NNs' linguistic knowledge using probing tasks in which a classifier is trained to directly predict grammatical properties of a sentence (e.g., syntactic tree depth) or part of a sentence (e.g., part-of-speech) using only the NNs' learned representation as input (Shi et al., 2016; Adi et al., 2017; Conneau et al., 2018; Ettinger et al., 2018; Tenney et al., 2019). We follow a complementary approach that uses acceptability judgments to address the same question without the need for training data labeled with grammatical concepts. Acceptability judgments are the main form of behavioral data used in generative linguistics to measure human linguistic competence (Chomsky, 1965; Schütze, 1996).

One branch of this literature uses minimal pairs to infer whether LMs detect specific grammatical contrasts. Table 1 summarizes linguistic phenomena studied in this work. For instance, Linzen et al. (2016) look closely at minimal pairs contrasting subject-verb agreement. Marvin and Linzen (2018) expand the investigation to negative polarity item and reflexive licensing. However, these and related studies cover a limited set of phenomena, to the exclusion of well-studied phenomena in linguistics such as control and raising, ellipsis, quantification, and countless others. This is likely due to the labor-intensive nature of collecting such targeted minimal pairs.

A related line of work evaluates neural networks on acceptability judgments in a more domain-general way. Corpora of sentences and their grammaticality are collected for this purpose in a

Phenomenon	Relevant work
Anaphora/binding	Marvin and Linzen (2018), Futrell et al. (2018), Warstadt et al. (2019b)
Subj.-verb agreement	Linzen et al. (2016), Futrell et al. (2018), Gulordava et al. (2019), Marvin and Linzen (2018), An et al. (2019), Warstadt et al. (2019b)
Neg. polarity items	Marvin and Linzen (2018), Futrell et al. (2018), Jumelet and Hupkes (2018), Wilcox et al. (2019), Warstadt et al. (2019a)
Filler-gap/Islands	Wilcox et al. (2018), Warstadt et al. (2019b), Chowdhury and Zamparelli (2018, 2019), Chaves (2020), Da Costa and Chaves (2020)
Argument structure	Kann et al. (2019), Warstadt et al. (2019b), Chowdhury and Zamparelli (2019)

Table 1: Summary of related work organized by linguistic phenomena tested. All studies analyze neural networks using acceptability judgments on minimal pairs mainly in English. Some studies appear multiple times.

number of studies (Heilman et al., 2014; Lau et al., 2017; Warstadt et al., 2019b). The most recent and comprehensive corpus is CoLA (Warstadt et al., 2019b), containing 10k sentences covering a wide variety of linguistic phenomena provided as examples in linguistics papers and books. CoLA, which is included in the GLUE benchmark (Wang et al., 2018), has been used to track advances in the sensitivity of reusable sentence encoding models to acceptability. Current models like BERT (Devlin et al., 2019) and T5 (Raffel et al., 2019) now learn to give acceptability judgments that approach or even exceed individual human agreement with CoLA.

Although CoLA can provide evidence about phenomenon-specific knowledge of models, this method is limited by the need to train a supervised classifier on CoLA data prior to evaluation. This is because CoLA is designed for binary acceptability classification, and there is no generally accepted method for obtaining binary acceptability predictions from unsupervised models like LMs.[2] Warstadt and Bowman (2019) measure phenomenon-specific performance on CoLA for several pretrained sentence encoding models: an LSTM, GPT (Radford et al., 2018), and BERT. However, the use of supervision prevents making strong conclusions about the sentence encoding component, since it is not possible to distinguish what the encoder knows from what is learned through supervised training on acceptability data.

Evaluating LMs on minimal pairs avoids this problem, with the caveat that the LM probability of a sentence can only serve as a proxy for acceptability if confounding factors impacting a sentence's probability such as length and lexical content are controlled for. It is with these considerations in mind that we design BLiMP.

3 Data

BLiMP consists of 67 minimal pair paradigms, each with 1,000 sentence pairs in mainstream American English grouped into 12 categories.[3] We refer to minimal pair types as *paradigms* and categories as *phenomena*. Each paradigm is annotated for the unique contrast it isolates and the broader phenomena it is part of. We automatically generate the data from linguist-crafted grammar templates, and our automatic labels are validated with crowd-sourced human judgments.

Although each minimal pair type corresponds to exactly one paradigm, a particular fact about English grammar may be illustrated by multiple paradigms. For instance, the fact that certain determiners and nouns agree can be illustrated by keeping the determiner the same and changing the number marking of the noun as in the example in Table 2, or by keeping the noun the same and changing the determiner (e.g., *Rachelle had bought those chair.*). With completeness in mind, we include such complementary paradigms in BLiMP whenever possible.

[2] Though see Lau et al. (2017) for some promising proposals for normalizing LM probabilities to correlate with gradient acceptability.

[3] We choose English because it is the native language of the linguists who built the grammar templates, though in the long run, creating versions of BLiMP in additional languages would allow for coverage of more phenomena and expand BLiMP's range of usefulness. We assume 1,000 pairs is sufficient to limit random noise resulting from small sample sizes.

Phenomenon	N	Acceptable Example	Unacceptable Example
Anaphor agr.	2	*Many girls insulted themselves.*	*Many girls insulted herself.*
Arg. structure	9	*Rose wasn't disturbing Mark.*	*Rose wasn't boasting Mark.*
Binding	7	*Carlos said that Lori helped him.*	*Carlos said that Lori helped himself.*
Control/raising	5	*There was bound to be a fish escaping.*	*There was unable to be a fish escaping.*
Det.-noun agr.	8	*Rachelle had bought that chair.*	*Rachelle had bought that chairs.*
Ellipsis	2	*Anne's doctor cleans one important book and Stacey cleans a few.*	*Anne's doctor cleans one book and Stacey cleans a few important.*
Filler-gap	7	*Brett knew what many waiters find.*	*Brett knew that many waiters find.*
Irregular forms	2	*Aaron broke the unicycle.*	*Aaron broken the unicycle.*
Island effects	8	*Whose hat should Tonya wear?*	*Whose should Tonya wear hat?*
NPI licensing	7	*The truck has clearly tipped over.*	*The truck has ever tipped over.*
Quantifiers	4	*No boy knew fewer than six guys.*	*No boy knew at most six guys.*
Subject-verb agr.	6	*These casseroles disgust Kayla.*	*These casseroles disgusts Kayla.*

Table 2: Minimal pairs from each of the twelve linguistic phenomenon categories covered by BLiMP. Differences are underlined. *N* is the number of 1,000-example minimal pair paradigms within each broad category.

3.1 Data Generation Procedure

To create minimal pairs exemplifying a wide array of linguistic contrasts, we found it necessary to artificially generate all datasets. This ensures both that we have sufficient unacceptable examples, and that the data is fully controlled, allowing for repeated isolation of a single linguistic phenomenon (Ettinger et al., 2018). For each paradigm, we use a generation script to sample lexical items from a vocabulary of over 3,000 items according to a template specifying linear order of the phrases in the acceptable and unacceptable sentences in each minimal pair. Our data generation scripts are publicly available.[4] We annotate these lexical items with the morphological, syntactic, and semantic features needed to enforce selectional restrictions and create grammatical and semantically felicitous sentences.

All examples in a paradigm are structurally analogous up to the point required for the relevant contrast but may vary in some ways. For instance, the template for NPI licensing, illustrated in Table 2, specifies that an arbitrary verb phrase needs to be generated. Accordingly, the generation script samples from the entire set of verbs and generates the required arguments on-the-fly. Thus, the structure of the sentence then depends on whether the sampled verb is transitive, clause-embedding, raising, and so forth, but that same verb phrase and its arguments are used in both pairs in the paradigm.

This generation procedure is not without limitations, and despite the very detailed vocabulary we use, implausible sentences are occasionally generated (e.g., *Sam ran around some glaciers*). In these cases, though, both the acceptable and unacceptable sentences will be equally implausible given world knowledge, so any difference in the probability assigned to them is still attributable to the intended grammatical contrast.

3.2 Coverage

The paradigms covered by BLiMP represent well-established contrasts in English morphology, syntax, and semantics. Each paradigm is grouped into one of 12 phenomena, shown in Table 2. Examples of all 67 paradigms appear in Table 4 of the Appendix. The paradigms are selected with the constraints that they can be characterized using templates as described above and illustrated with minimal pairs of sentences equal in length[5] that differ in at most one vocabulary item.

Although this dataset has broad coverage, it is not exhaustive. It is not possible to include every

[4] `https://github.com/alexwarstadt/data_generation`.

[5] We define length as the number of entries from our lexicon. Some sentences in a pair contain different numbers of words because *visit* and *drop by* are each one lexical entry. Where discrepancies in number of words occur, they are generally randomly distributed across the grammatical and ungrammatical sentences in a paradigm.

grammatical phenomenon of English, and there is no agreed-upon set of core phenomena. However, we consider frequent inclusion of a phenomenon in a syntax/semantics textbook as an informal proxy for what linguists consider to be core phenomena. We survey several syntax textbooks (e.g., Sag et al., 2003; Adger, 2003; Sportiche et al., 2013), and find that nearly all of the phenomena in BLiMP are discussed in some source. Most of the topics that repeatedly appear in textbooks and can be represented with minimal pairs (e.g., agreement, control/raising, wh-extraction/islands, binding) are present in BLiMP.[6]

We characterize the 12 phenomena in BLiMP as follows[7]:

- ANAPHOR AGREEMENT: the requirement that reflexive pronouns like *himself* (a.k.a. anaphora) agree with their antecedents in person, number, gender, and animacy.

- ARGUMENT STRUCTURE: the ability of different verbs to appear with different types of arguments. For instance, different verbs can appear with a direct object, participate in the causative alternation, or take an inanimate argument.

- BINDING: the structural relationship between a pronoun and its antecedent. All paradigms illustrate aspects of Chomsky's (1981) Principle A. Because coindexation cannot be annotated in BLiMP, Principles B and C are not illustrated.

- CONTROL/RAISING: syntactic and semantic differences between various types of predicates that embed an infinitival VP. This includes control, raising, and *tough*-movement predicates.

- DETERMINER-NOUN AGREEMENT: number agreement between demonstrative determiners (e.g., *this/these*) and the associated noun.

- ELLIPSIS: the possibility of omitting expressions from a sentence. Because this is difficult to illustrate with sentences of equal length, our paradigms cover only special cases of noun phrase ellipsis that meet this constraint.

- FILLER-GAP: dependencies arising from phrasal movement in, for example, *wh*-questions.

- IRREGULAR FORMS: irregular morphology on English past participles (e.g., *broken*). We are unable to evaluate models on nonexistent forms like **breaked* because such forms are out of the vocabulary for some LMs.

- ISLAND EFFECTS: restrictions on syntactic environments where the gap in a filler-gap dependency may occur.

- NPI LICENSING: restrictions on the distribution of *negative polarity items* like *any* and *ever* limited to, for example, the scope of negation and *only*.

- QUANTIFIERS: restrictions on the distribution of quantifiers. We cover two such restrictions: superlative quantifiers (e.g., *at least*) cannot embed under negation, and definite quantifiers and determiners cannot be subjects in existential-*there* constructions.

- SUBJECT-VERB AGREEMENT: subjects and present tense verbs must agree in number.

3.3 Comparison to Related Resources

With a vocabulary of over 3,000 words, BLiMP has by far the most lexical variation of any related generated dataset. It includes verbs with 11 different subcategorization frames, including verbs that select for PPs, infinitival VPs, and embedded clauses. By comparison, datasets by Ettinger et al. (2018) and Marvin and Linzen (2018) use vocabularies of under 200 items. Other datasets of minimal pairs that achieve more lexical and syntactic variety use data-creation methods that limit empirical scope and control. Linzen et al. (2016) construct a dataset of minimal pairs for subject-verb agreement by changing verbs' number marking in a subset of English Wikipedia, but this approach does not generalize beyond agreement phenomena. Lau et al. (2017) construct minimal pairs by taking sentences from the BNC through round-trip machine translation. The resulting sentences contain a wider variety of

[6] In line with these textbooks, we rely on stereotyped gender-name pairings and contrasts not present in all English dialects (more detail provided in the Appendix).

[7] Our implementation of these phenomena is often narrower than the linguistic definition because of the particular constraints described above.

Model	Overall	Ana. Agr	Arg. Str	Binding	Ctrl. Rais.	D-N Agr	Ellipsis	Filler. Gap	Irregular	Island	NPI	Quantifiers	S-V Agr
5-gram	60.5	47.9	71.9	64.4	68.5	70.0	36.9	58.1	79.5	53.7	45.5	53.5	60.3
LSTM	68.9	91.7	73.2	73.5	67.0	85.4	67.6	72.5	89.1	42.9	51.7	64.5	80.1
TXL	68.7	94.1	69.5	74.7	71.5	83.0	77.2	64.9	78.2	45.8	55.2	69.3	76.0
GPT-2	80.1	99.6	78.3	80.1	80.5	93.3	86.6	79.0	84.1	63.1	78.9	71.3	89.0
Human	88.6	97.5	90.0	87.3	83.9	92.2	85.0	86.9	97.0	84.9	88.1	86.6	90.9

Table 3: Percentage accuracy of four baseline models and raw human performance on BLiMP using a forced-choice task. A random guessing baseline would achieve an accuracy of 50%.

grammatical violations, but it is not possible to control the nature or quantity of violations in the resulting sentences.

3.4 Data Validation

To verify that the generated sentences represent a real contrast in acceptability, we conduct human validation via Amazon Mechanical Turk.[8] Twenty separate validators rated five pairs from each of the 67 paradigms, for a total of 6,700 judgments. We restricted validators to individuals currently located in the US who self-reported as native speakers of English. To assure that our validators made a genuine effort on the task, each HIT included an attention check item and a hidden field question to catch bot-assisted humans. Validators were paid $0.25 for completing five judgments, which we estimate took 1-2 minutes. For each minimal pair, 20 individuals completed a forced-choice task mirroring the LMs' task; the human-determined acceptable sentence was calculated via majority vote of annotators. By this metric, we estimate aggregate human agreement with our annotations to be 96.4% overall. As a threshold of inclusion in BLiMP, the majority of validators needed to agree with BLiMP on at least 4/5 examples from each paradigm. Thus, all 67 paradigms in the public version of BLiMP passed this validation; only two additional paradigms were rejected on this criterion. We also estimate *individual* human agreement to be 88.6% overall using the approximately 100 annotations from each paradigm.[9] Table 3 reports individual human results (and model results) as a conservative measure of human agreement.

4 Models

GPT-2 GPT-2 (Radford et al., 2019) is a large-scale language model using the Transformer architecture (Vaswani et al., 2017). Our main experiments use GPT-2-large with 36 layers and 774M parameters.[10] The model is pretrained on Radford et al.'s WebText dataset, which contains 40GB of English text extracted from Web pages and filtered for quality. To our knowledge, WebText is not publicly available, so assuming an average of 5–6 bytes/chars per word, we estimate WebText contains about 8B tokens. We use `jiant`, a codebase for training and evaluating sentence understanding models (Wang et al., 2019b), to implement code for evaluating GPT-2 on BLiMP.[11]

Transformer-XL Transformer-XL (Dai et al., 2019) is another multilayer Transformer-based neural language model. We test the pretrained Transformer-XL Large model with 18 layers of Transformer decoders and 16 attention heads for each layer. The model is trained on WikiText-103 (Merity et al., 2016), a corpus of 103M tokens from English Wikipedia. Code for testing Transformer-XL on BLiMP is also implemented in `jiant`.

LSTM We include a long-short term memory (LSTM, Hochreiter and Schmidhuber, 1997) LM in our experiments. Specifically, we test a pretrained LSTM LM from Gulordava et al. (2019) on BLiMP. The model is trained on a 83M-token corpus extracted from English Wikipedia. To investigate the effect of training size on model performance (§6.3), we retrain a series of LSTM and Transformer-XL models with the same hyperparameters and the following training

[8] The full set of human judgments and a summary of the results for all 67 paradigms is in Table 4 in the Appendix.

[9] A few had to be excluded due to ineligible annotators.

[10] GPT-2-XL performs slightly worse on BLiMP; see §6.3.

[11] `https://github.com/nyu-mll/jiant/tree/blimp-and-npi/scripts/blimp`.

sizes: 64M, 32M, 16M, 8M, 4M, 2M, 1M, 1/2M, 1/4M, and 1/8M tokens. For each size, we train the model on five different random samples of the original training data, which has a size of 83M tokens.[12]

5-gram We build a 5-gram LM on the English Gigaword corpus (Graff et al., 2003), which consists of 3.1B tokens. To efficiently query n-grams we use an implementation[13] based on Heafield et al. (2013).[14]

5 Results and Discussion

An LM's overall accuracy on BLiMP is simply the proportion of the 67,000 minimal pairs in which the model assigns a higher probability to the acceptable sentence. We report the results for all models and human evaluation in Table 3. GPT-2 achieves the highest accuracy and the 5-gram model the lowest. All models perform well below estimated human accuracy (as described in § 3.4). The 5-gram model's poor performance—overall and on every individual category—indicates that BLiMP is likely not solvable from local co-occurrence statistics alone.

Because we evaluate pretrained models that differ in architecture and training data, we can only speculate about what drives these differences (though see § 6.3 for a controlled ablation study on the LSTM LM). The results seem to indicate that access to training data is the main driver of performance on BLiMP for the neural models we evaluate. This may explain why Transformer-XL and the LSTM LM perform similarly in spite of differences in architecture, as both are trained on approximately 100M tokens of Wikipedia text. Relatedly, GPT-2's advantage may come from the fact that it is trained on roughly two orders of magnitude more data. Possibly, LSTMs trained on larger datasets could perform comparably to GPT-2, but such experiments are impractical because of the inefficiency of training LSTMs at this scale.

5.1 Results and Discussion by Phenomenon

The results also give insight into how LM's linguistic knowledge varies by domain. Models generally perform best and closest to human level on morphological phenomena. For instance, GPT-2 performs within 2.1 points of humans on ANAPHOR AGR., DET.-NOUN AGR., and SUBJ.-VERB AGR.. The set of challenging phenomena is more diverse. ISLANDS are the hardest phenomenon by a wide margin. Only GPT-2 performs well above chance, and it remains 20 points below humans. Some semantic phenomena, specifically those involving NPI LICENSING and QUANTIFIERS, are also challenging overall. All models perform relatively poorly on ARG. STRUCTURE.

From these results we conclude that current SotA LMs robustly encode basic facts of English agreement. This does not mean that LMs will come close to human performance for all agreement phenomena. §6.1 discusses evidence that increased dependency length and the presence of agreement attractors of the kind investigated by Linzen et al. (2016) and Gulordava et al. (2019) reduce performance on agreement phenomena.

We find, in accordance with Wilcox et al. (2018), that LMs do represent long-distance wh-dependencies, but we also conclude that their representations differ fundamentally from humans'. Although some models approach human performance in ordinary filler-gap dependencies, they are exceptionally poor at identifying island violations overall. This finding suggests that they reliably encode long-distance dependencies in general, but not the syntactic domains in which these dependencies are blocked, though GPT-2 does perform well above chance on some paradigms of ISLAND EFFECTS. However, strong conclusions about how these models represent wh-dependencies are not possible using the forced-choice task compatible with BLiMP, and a complete assessment of syntactic islands is best addressed using a factorial design that manipulates both the presence of an island and an attempt to extract from it, as in Kush et al. (2018) or Wilcox et al. (2018).

In the semantic phenomena where models struggle (NPIs and QUANTIFIERS), violations are often attributed in semantic theories to a presupposition failure or contradiction arising from semantic composition or pragmatic reasoning (e.g., Chierchia, 2013; Ward and Birner, 1995; Geurts and Nouwen, 2007). These abstract

[12] https://github.com/sheng-fu/colorlessgreenRNNs.

[13] https://github.com/kpu/kenlm.

[14] https://github.com/anhad13/blimp_ngram.

semantic and pragmatic factors may be difficult for LMs to learn. Marvin and Linzen also find that LSTMs largely fail to recognize NPI licensing conditions. Warstadt et al. (2019a) find that BERT (which is similar in scale to GPT-2) recognizes these conditions inconsistently in an unsupervised setting.

The weak performance on ARG. STRUCTURE is somewhat surprising, since arguments and heads are usually—though not always—adjacent (e.g., subjects and direct objects are adjacent to the verb in default English word order). However, argument structure is closely related to semantic event structure (see Marantz, 2013), which may be comparatively difficult for LMs to learn. Also, judgments about argument structure are complicated by the possibility of coercing a frequently transitive verb to be intransitive and vice versa as well as the existence of secondary meanings of verbs with different argument structures (e.g., normally intransitive *boast* has a transitive use as in *The spa boasts 10 pools*), which might make this domain somewhat more difficult for LMs. Though even with these complications, humans detect the intended contrast 90% of the time. We note that the reported difficulty of these phenomena contradicts Warstadt and Bowman's (2019) conclusion that argument structure is one of the strongest domains for neural models. However, Warstadt and Bowman evaluate classifiers with supervision on CoLA, a large proportion of which is sentences related to argument structure.

Finally, we caution against interpreting positive results on a general phenomenon in BLiMP as proof of human-like knowledge. Although it is unlikely that GPT-2 could reach human performance on the SUBJ.-VERB AGR. paradigms without acquiring a concept of number marking that abstracts away from specific lexical items, it is difficult to rule out this possibility without accumulating different forms of evidence, for instance, by testing how it generalizes to nonce words. We take the paradigms in FILLER-GAP as a cautionary example (see Table 4). There are four paradigms that assess a model's sensitivity to the syntactic requirements of complementizer *that* versus a wh-word. We observe that all models more or less succeed when the unacceptable sentence lacks a necessary gap, but fail when it contains an illicit gap. These results suggest the models' ability to accurately detect a contrast in whether a gap is filled following a wh-word is not clearly based on a generalization about the relationship between that wh-word and its gap, as such a generalization should extend to the cases where the models currently fail to detect the correct contrast. More generally, conclusions about a model's knowledge of a particular grammatical concept can only be reached by considering several paradigms.

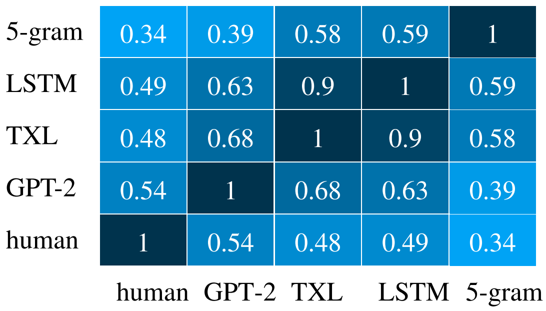

Figure 1: Heatmap showing the correlation between models' accuracies in each of the 67 paradigms.

5.2 Shallow Predictors of Performance

We also ask what factors besides linguistic phenomena affect model accuracy. Figure 2 shows how sentence length, perplexity (which does not depend on length), the probability of the good sentence (which does depend on length), and confidence affect model performance. The effect of perplexity is much weaker for GPT-2 than for other models, which indicates that it is probably more robust to sentences with non-stereotypical syntax or describing unlikely scenarios. GPT-2 is the only model where accuracy increases largely monotonically with confidence. A similar relationship holds between confidence and agreement in human acceptability judgments.

5.3 Correlation of Model and Human Performance

We examine the extent to which models and humans succeed at detecting contrasts for the same linguistic phenomena. Figure 1 shows the Pearson correlation between the four LMs and humans of their accuracies on the 67 paradigms. The neural models correlate moderately with humans, with GPT-2 correlating most strongly. The n-gram model's performance correlates with humans relatively weakly. Neural models correlate with each other more strongly, suggesting neural networks share some biases that are not human-like.

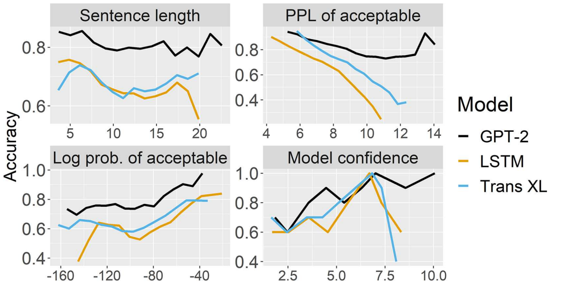

Figure 2: Models' performance on BLiMP as a function of sentence length, perplexity, log probability of the acceptable sentence, and model confidence (calculated as $|\log P(S_1) - \log P(S_2)|$).

Transformer-XL and LSTM's high correlation of 0.9 possibly reflects their similar training data.

6 Analysis

6.1 Long-Distance Dependencies

The presence of intervening material can lower the ability of humans to detect agreement dependencies (Bock and Miller, 1991). We study how intervening material affects the LMs' sensitivity to mismatches in agreement in BLiMP. First, we test for sensitivity to determiner-noun agreement with and without an intervening adjective, as in Example (2). The results are plotted in Figure 3. The n-gram model is the most heavily impacted, performing on average 35 points worse. This is unsurprising, since the bigram consisting of a determiner and noun is far more likely to be observed than the trigram of determiner, adjective, and noun. For the neural models, we find a weak but consistent effect, with all models performing on average between 5 and 3 points worse when there is an intervening adjective.

(2) a. Ron saw that man/*men.
b. Ron saw that nice man/*men.

Second, we test for sensitivity to mismatches in subject-verb agreement when an *attractor* noun of the opposite number intervenes. We compare attractors in relative clauses (3-b) and as part of a relational noun (3-c), following experiments by Linzen et al. (2016) and others. Again, we find that the n-gram model's performance is reduced significantly by this intervening material, suggesting the model is consistently misled by the presence of an attractor. All the neural models perform above chance with an attractor present, but GPT-2 and the LSTM perform 22 and 20 points worse when an attractor is present than when there is no attractor, while Transformer-XL's performance is reduced by only 5 points. Thus, we reproduce Linzen et al.'s finding that attractors significantly reduce LSTM LMs' sensitivity to mismatches in agreement and find evidence that this holds true of some Transformer LMs as well.

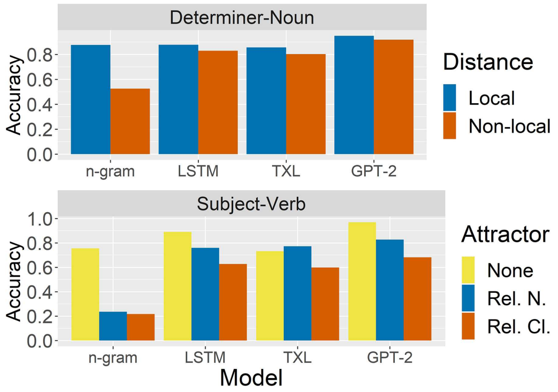

Figure 3: The effect of the locality of determiner-noun agreement (upper panel) and the type of agreement attractor (lower panel) on model performance.

(3) a. The sisters bake/*bakes.
b. The sisters who met Cheryl bake/*bakes.
c. The sisters of Cheryl bake/*bakes.

6.2 Regular vs. Irregular Agreement

In DET.-NOUN AGR. and SUBJ.-VERB AGR., we generate separate datasets for nouns with regular and irregular number marking, as in Example (4). All else being equal, only models with access to sub-word-level information should make any distinction between regular and irregular morphology.

(4) a. Ron saw that nice kid/*kids. (regular)
b. Ron saw that nice man/*men. (irregular)

In fact, Figure 4 shows that the two sub-word-level models GPT-2 and Transformer-XL show little effect of irregular morphology: They perform less than 1.3 points worse on irregulars than regulars. Their high overall performance suggests that they robustly encode number features without relying on segmental cues.[15]

[15] The LSTM LM, which has word-level tokens, averages 5.2 points worse on the irregular paradigms. This effect is not due to morphology, but rather to the higher proportion of out-of-vocabulary items among the irregular nouns, which include many loanwords such as *theses* and *alumni*.

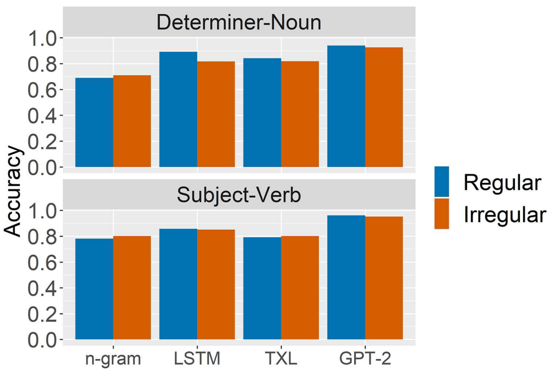

Figure 4: Models' performance on agreement phenomena between a determiner and noun and between a subject and verb, broken down by whether the noun/subject has a regular or irregular plural form

6.3 Training size and BLiMP performance

We use BLiMP to track how a model's representation of particular phenomena varies with the quantity of training data. Using different sized subsets of Gulordava et al.'s (2019) training data, we retrain the LSTM and Transformer-XL models and evaluate their performance on BLiMP. Figure 5 shows that different phenomena have notably different learning curves across different training sizes even if the full model trained on 83M tokens achieved equivalent accuracy scores. For example, the LSTM model ultimately performs well on both IRREGULAR and ANAPHOR AGR., but requires more training to reach this level of performance for ANAPHOR AGR. These learning curve differences show how BLiMP performance dissociates from perplexity on Wikipedia data, a standard measure of LM performance: Although perplexity decreases with more training data,[16] performance on different phenomena grows at varying rates.

We conjecture that there is a sigmoid relationship between the logarithm of training set size and BLiMP performance that appears to be roughly linear at this scale. We conduct linear regression analyses to estimate the rate of increase in performance in relation to the logarithm (base 2) of dataset size. For the LSTM LM, best-fit lines for phenomena on which the model had the highest accuracy have the steepest slopes: ANAPHOR AGR. (0.0623), DET.-NOUN AGR. (0.0426), and IRREGULAR (0.039). We see the shallowest slopes on phenomena with the worst performance: NPIs (0.0078) and ISLANDS (0.0036). For Transformer-XL, we observe a similar pattern: The steepest learning curves again belong to ANAPHOR AGR. (0.0545) and DET.-NOUN AGR. (0.0405), and the shallowest to NPIs (0.0055) and ISLANDS (0.0039). Based on these values, we estimate that if log-linear improvement continues, the LSTM LM and Transformer-XL should require well over 10^{20} tokens of training data to achieve human-like performance on these hardest phenomena.

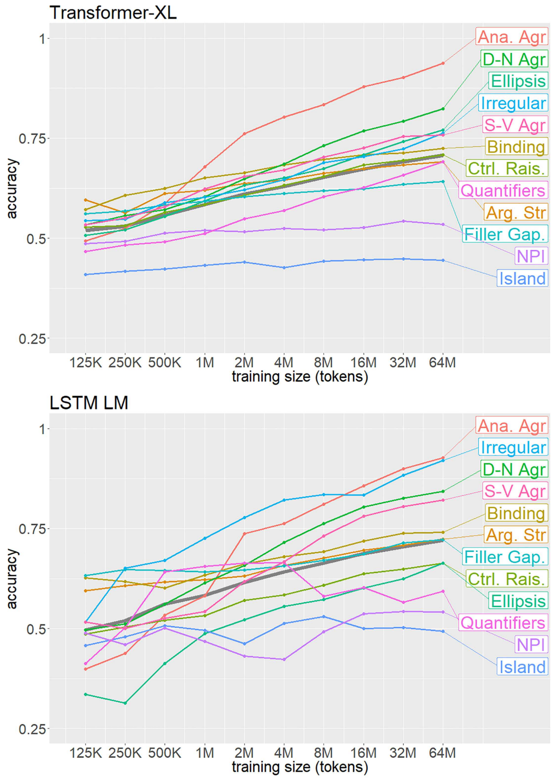

Figure 5: Transformer-XL (top) and LSTM LM (bottom) performance as a function of training size and phenomena in BLiMP. The gray line shows the average across all phenomena.

We also find that increasing model size (number of parameters) is unlikely to improve performance: We evaluate four pretrained versions of GPT-2 with 117 M to 1,558 M parameters trained on WebText. All models have overall BLiMP accuracy of 0.84 ± .01%, and standard deviation among the models on each of the 12 phenomena does not exceed 0.03. This finding bolsters our earlier conclusion in §5 that amount of training data has the biggest impact on BLiMP performance.

[16] Average perplexity on the Gulordava et al. (2019) test set: 595 at 0.125M, 212 at 1M, 92.8 at 8M, and 53 at 64M.

6.4 Alternate Evaluation Methods

There are several other methods one can use to measure an LM's preference between two minimally different sentences. So far, we have considered only the *full-sentence method*, advocated for by Marvin and Linzen (2018), which compares LM likelihoods of full sentences. In a followup experiment, we use two *prefix methods*, each of which has appeared in related prior work, that evaluate a model's preferences by comparing its prediction at a key point of divergence between the sentences. Subsets of BLiMP data are designed to be compatible with multiple methods, allowing us to conduct the first direct comparison. We find that all methods give broadly similar results when aggregating over a set of paradigms. We see no strong argument against evaluating solely using the full-sentence method, though some results diverge for specific paradigms.

One-Prefix Method In the *one-prefix method*, used by Linzen et al. (2016), a pair of sentences share the same initial portion of a sentence, but differ in a critical word that make them differ in grammaticality (e.g., *The cat **eats** mice* vs. *The cat **eat** mice*). The model's prediction is correct if it assigns a higher probability to the grammatical token given the shared prefix.

Two-Prefix Method In the *two-prefix method*, used by Wilcox et al. (2019), a pair of sentences differ in their initial string, and the grammaticality difference is only revealed when a shared critical word is included (e.g., *The cat **eats** mice* vs. *The cats **eats** mice*). For these paradigms, we evaluate whether the model assigns a higher probability to the critical word conditioned on the grammatical prefix than on the ungrammatical prefix.

The prefix methods differ from the full-sentence method in two key ways: (i) they require that the acceptability of the sentence be unambiguously predictable from the critical word, but not sooner, and (ii) they are not affected by predictions made by the LM following the critical word. These values do affect the full sentence method. For example, assuming that $P(\textit{are numerous}) \gg P(\textit{is numerous})$, a model could predict that *The cats **are** numerous* is more likely than *The cats **is** numerous* without correctly predicting that $P(\textit{are}|\textit{the cats}) > P(\textit{is}|\textit{the cats})$. Using prefix probabilities allows us to exclude models' use of this additional information and evaluate how the models perform when they have just enough information to judge grammaticality.

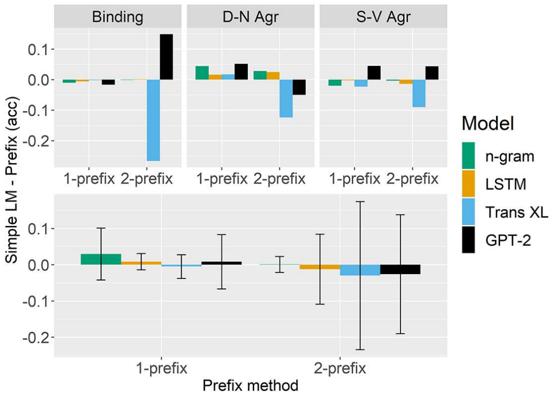

Figure 6: Comparison of models' performance on the simple LM method and the 1- and 2-prefix methods. The upper panels show results from three phenomena that are compatible with both 1-prefix and 2-prefix methods. The lower panel shows the averages and standard deviations across all phenomena.

Figure 6 shows that models have generally comparable accuracies across all three methods. However, there are some cases where we observe differences between these methods. For example, Transformer-XL performs much worse at BINDING, DET.-NOUN AGR., and SUBJ.-VERB AGR. in the simple LM method, suggesting that the probabilities Transformer-XL assigns to the irrelevant part at the end of the sentence very often overturn the observed preference based on probability up to the critical word. On the other hand, GPT-2 benefits from reading the whole sentence for BINDING phenomena, as its performance is better in the simple LM method than in the prefix method.

We conclude that with a sufficiently diverse set of paradigms, the various metrics under consideration will give similar results. Thus, it is not problematic that BLiMP relies only on the full-sentence method, and doing so allows BLiMP to include many paradigms not compatible with either prefix method. Nonetheless, prefix methods are still valuable for detailed analysis or for studies making direct comparison to psycholinguistic theories (e.g., Wilcox et al., 2018).

7 Conclusion and Future Work

We have shown ways in which BLiMP can be used as tool to gain evidence about both the overall and fine-grained linguistic knowledge of language models. Like the GLUE benchmark (Wang et al., 2018), BLiMP assigns a single overall score to an LM that summarizes its general sensitivity to minimal pair contrasts. It also provides a breakdown of LM performance by linguistic phenomenon, which can be used to draw more concrete conclusions about the kinds of grammatical features learned acquired by a given model. This kind of information is a linguistically motivated evaluation of LMs that can complement common metrics like perplexity.

Furthermore, the extent to which humans resemble data-driven learners like language models is debated in linguistics and cognitive science (see e.g., Chomsky, 1965; Reali and Christiansen, 2005). In some domains, we may require the aid of innate knowledge to acquire phenomenon-specific knowledge resembling that tested in BLiMP. By evaluating whether self-supervised learners like LMs acquire human-like grammatical acuity in a particular domain, we gather indirect evidence as to whether this phenomenon is a necessary component of humans' innate knowledge.

Another aim of BLiMP is to serve as a guide for future work on the linguistic evaluation of LMs. It is particularly interesting to better understand those empirical domains where current LMs appear to acquire some relevant knowledge, but still fall short of human performance. The results from BLiMP suggest that—in addition to relatively well-studied phenomena like filler-gap dependencies, NPIs, and binding—argument structure remains one area where there is much to uncover about what LMs learn. More generally, as language modeling techniques continue to improve, it will be useful to have large-scale tools like BLiMP to efficiently track changes in what these models do and do not know about grammar.

Acknowledgments

This material is based upon work supported by the National Science Foundation under grant no. 1850208. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This project has also benefited from support to SB by Eric and Wendy Schmidt (made by recommendation of the Schmidt Futures program), by Samsung Research (under the project *Improving Deep Learning using Latent Structure*), by Intuit, Inc., and by NVIDIA Corporation (with the donation of a Titan V GPU).

Appendix

The following contains examples from each of the 67 paradigms in BLiMP.

Caveats Some paradigms include non-transparent factors that may influence interpretation. We list here those factors that we are aware of:

- Several paradigms within ANAPHOR AGREEMENT and BINDING rely on stereotyped gender assignment associated with names (e.g., *Mary*). A model has to have at least a weak gender-name association in order to succeed on some paradigms in BLiMP. For example, we mark sentences like *Mary hugged themselves* and *Mary hugged himself* as unacceptable, and we never include possibilities like *Mary hugged themself.*

- To isolate certain phenomena, we had to rely on acceptability contrasts present in mainstream US and UK English but absent in many other dialects. For example, some speakers would accept the sentence *Suzy don't lie*, but we would mark this unacceptable based on mainstream US English judgments. BLiMP assesses models' knowledge of this specific dialect of English; in some cases it could *penalize* models that conform to a different dialect.

How to read this table:

- *Phenomenon* refers to the linguistic phenomenon as noted in Table 2. *UID* refers to the unique identifier used in the released dataset.

- Model and human performance are reported as percent accuracy. 'Human' uses the more conservative individual judgments (as opposed to majority vote, for which each paradigm would be either 100% or 80%).

- Each pair is marked for whether it is usable with a prefix method. All sentences are valid for the simple LM method.

- If a sentence has a checkmark (✓) under the *1pfx* column, the sentence can be used with the 1-prefix method in addition to the

Phenomenon	UID	5-gram	LSTM	TXL	GPT-2	Human	Acceptable Example	Unacceptable Example	1pfx	2pfx
Anaphor agreement	anaphor_gender_agreement	44	88	91	99	96	Katherine can't help **herself**.	Katherine can't help **himself**.	✓	
	anaphor_number_agreement	52	95	97	100	99	Many teenagers were helping **themselves**.	Many teenagers were helping **herself**.	✓	
Argument structure	animate_subject_passive	54	68	58	77	98	Amanda was respected by some **waitresses**.	Amanda was respected by some **picture**.	✓	
	animate_subject_trans	72	79	70	80	87	Danielle **visited** Irene.	The eye **visited** Irene.		✓
	causative	51	65	54	68	82	Aaron breaks the glass.	Aaron appeared the glass.		
	drop_argument	68	79	67	84	90	The Lutherans couldn't skate around.	The Lutherans couldn't disagree with.		
	inchoative	89	72	81	90	95	A screen was fading.	A screen was cleaning.		
	intransitive	82	73	81	90	86	Some glaciers are vaporizing.	Some glaciers are scaring.		
	passive_1	71	65	76	89	99	Jeffrey's sons are insulted by Tina's supervisor.	Jeffrey's sons are smiled by Tina's supervisor.		
	passive_2	70	72	74	79	86	Most cashiers are disliked.	Most cashiers are flirted.		
	transitive	91	87	89	49	87	A lot of actresses' nieces have toured **that** art gallery.	A lot of actresses' nieces have coped **that** art gallery.		✓
Binding	principle_A_c_command	58	59	61	100	86	A lot of actresses that thought about Alice healed **themselves**.	A lot of actresses that thought about Alice healed **herself**.	✓	
	principle_A_case_1	100	100	100	96	98	Tara thinks that **she** sounded like Wayne.	Tara thinks that **herself** sounded like Wayne.	✓	
	principle_A_case_2	49	87	95	73	96	Stacy imagines herself **praising** this actress.	Stacy imagines herself **praises** this actress.	✓	
	principle_A_domain_1	95	98	99	99	95	Carlos said that Lori helped **him**.	Carlos said that Lori helped **himself**.	✓	
	principle_A_domain_2	56	68	70	73	75	Mark imagines Erin might admire **herself**.	Mark imagines Erin might admire **himself**.	✓	
	principle_A_domain_3	52	55	60	82	83	Nancy could say every guy hides **himself**.	Every guy could say Nancy hides **himself**.		✓
	principle_A_reconstruction	40	46	38	37	78	It's herself who Karen criticized.	It's herself who criticized Karen.		
Control/ raising	existential_there_object_raising	84	66	76	92	90	William has declared **there** to be no guests getting fired.	William has obliged **there** to be no guests getting fired.		✓
	existential_there_subject_raising	77	80	79	89	88	There was bound to be a fish escaping.	There was unable to be a fish escaping.		
	expletive_it_object_raising	72	63	72	58	86	Regina wanted it to be **obvious** that Maria thought about Anna.	Regina forced it to be **obvious** that Maria thought about Anna.		✓
	tough_vs_raising_1	33	34	45	72	75	Julia wasn't fun to talk to.	Julia wasn't unlikely to talk to.		
	tough_vs_raising_2	77	93	86	92	81	Rachel was apt to talk to Alicia.	Rachel was exciting to talk to Alicia.		
Deter-miner-noun agr.	determiner_noun_agreement_1	88	92	92	100	96	Craig explored that **grocery store**.	Craig explored that **grocery stores**.	✓	
	determiner_noun_agreement_2	86	92	81	93	95	Carl cures those **horses**.	Carl cures that **horses**.		✓
	determiner_noun_agreement_irregular_1	85	82	88	94	92	Phillip was lifting this **mouse**.	Phillip was lifting this **mice**.	✓	
	determiner_noun_agreement_irregular_2	90	86	82	93	85	Those ladies walk through those **oases**.	Those ladies walk through that **oases**.		✓
	determiner_noun_agreement_with_adj_1	50	86	78	90	96	Tracy praises those lucky **guys**.	Tracy praises those lucky **guys**.	✓	
	determiner_noun_agreement_with_adj_2	53	76	81	96	94	Some actors buy these gray **books**.	Some actors buy this gray **books**.		✓
	determiner_noun_agreement_with_adj_irregular_1	55	83	77	88	85	This person shouldn't criticize this upset **child**.	This person shouldn't criticize this upset **children**.	✓	
	determiner_noun_agreement_with_adj_irregular_2	52	87	86	93	95	That adult has brought that purple **octopus**.	That adult has brought those purple **octopus**.		✓
Ellipsis	ellipsis_n_bar_1	23	68	65	88	92	Brad passed one big museum and Eva passed several.	Brad passed one museum and Eva passed several big.		
	ellipsis_n_bar_2	50	67	89	86	78	Curtis's boss discussed four sons and Andrew discussed five sick sons.	Curtis's boss discussed four happy sons and Andrew discussed five sick.		
Filler gap	wh_questions_object_gap	53	79	61	84	85	Joel discovered the vase that Patricia might take.	Joel discovered what Patricia might take the vase.		
	wh_questions_subject_gap	82	92	83	95	98	Cheryl thought about some dog that upset Sandra.	Cheryl thought about who some dog upset Sandra.		
	wh_questions_subject_gap_long_distance	86	96	86	88	85	Bruce knows that person that Dawn likes that argued about a lot of guys.	Bruce knows who that person that Dawn likes argued about a lot of guys.		
	wh_vs_that_no_gap	83	97	86	97	97	Danielle finds out that many organizations have alarmed Chad.	Danielle finds out who many organizations have alarmed Chad.		
	wh_vs_that_no_gap_long_distance	81	97	91	94	92	Christina forgot that all plays that win worry Dana.	Christina forgot who all plays that win worry Dana.		
	wh_vs_that_with_gap	18	43	42	56	77	Nina has learned who most men sound like.	Nina has learned that most men sound like.		
	wh_vs_that_with_gap_long_distance	20	14	17	56	75	Martin did find out what every cashier that shouldn't drink wore.	Martin did find out that every cashier that shouldn't drink wore.		
Irregular forms	irregular_past_participle_adjectives	79	93	91	78	99	The forgotten **newspaper article** was bad.	The forgot **newspaper article** was bad.		✓
	irregular_past_participle_verbs	80	85	66	90	95	Edward **hid** the cats.	Edward **hidden** the cats.	✓	
Island effects	adjunct_island	48	67	65	91	94	Who has Colleen aggravated before kissing Judy?	Who has Colleen aggravated Judy before kissing?		
	complex_NP_ island	50	47	58	72	80	Who hadn't some driver who would fire Jennifer's colleague embarrassed?	Who hadn't Jennifer's colleague embarrassed some driver who would fire?		
	coordinate_structure_constraint_complex_left_branch	32	30	36	42	90	What lights could Spain sell and **Andrea** discover?	What could Spain sell lights and **Andrea** discover?		✓
	coordinate_structure_constraint_object_extraction	59	71	74	88	91	Who will Elizabeth and Gregory cure?	Who will Elizabeth cure and Gregory?		
	left_branch_island_echo_question	96	32	63	77	91	David would cure what **snake**?	What would David cure **snake**?		✓
	left_branch_island_simple_question	57	36	36	82	99	Whose hat should Tonya wear?	Whose should Tonya wear hat?		
	sentential_subject_island	61	43	37	35	61	Who have many women's touring Spain embarrassed.	Who have many women's touring embarrassed Spain.		
	wh_island	56	47	20	77	73	What could Alan discover **he** has run around?	What could Alan discover **who** has run around?	✓	
NPI licensing	matrix_question_npi_licensor_present	1	2	1	67	98	Should Monica **ever** grin?	Monica should **ever** grin.		✓
	npi_present_1	47	54	61	55	83	Even these trucks have **often** slowed.	Even these trucks have **ever** slowed.	✓	
	npi_present_2	47	54	48	62	98	Many skateboards **also** roll.	Many skateboards **ever** roll.	✓	
	only_npi_licensor_present	57	93	80	100	92	Only Bill would **ever** complain.	Even Bill would **ever** complain.		✓
	only_npi_scope	30	36	45	85	72	Only those doctors who Karla respects **ever** conceal many snakes.	Those doctors who only Karla respects **ever** conceal many snakes.		✓
	sentential_negation_npi_licensor_present	93	100	99	89	93	Those banks had not **ever** lied.	Those banks had really **ever** lied.		✓
	sentential_negation_npi_scope	45	23	53	95	81	Those turtles that are boring April could not **ever** break those couches.	Those turtles that are not boring April could **ever** break those couches.		✓
Quantifiers	existential_there_quantifiers_1	91	96	94	99	94	There aren't many lights darkening.	There aren't all lights darkening.		
	existential_there_quantifiers_2	62	16	14	24	76	Each book is there disturbing Margaret.	There is each book disturbing Margaret.		
	superlative_quantifiers_1	45	63	84	84	91	No man has revealed more than five forks.	No man has revealed at least five forks.		
	superlative_quantifiers_2	17	83	85	78	85	An actor arrived at at **most** six lakes.	No actor arrived at at **most** six lakes.		✓
Subject-verb agr.	distractor_agreement_relational_noun	24	76	77	83	81	A sketch of lights **doesn't** appear.	A sketch of lights **don't** appear.	✓	
	distractor_agreement_relative_clause	22	63	60	68	86	Boys that aren't disturbing Natalie **suffer**.	Boys that aren't disturbing Natalie **suffers**.		✓
	irregular_plural_subject_verb_agreement_1	73	81	78	95	95	This goose **isn't** bothering Edward.	This goose **weren't** bothering Edward.	✓	
	irregular_plural_subject_verb_agreement_2	88	89	83	96	94	The woman **cleans** every public park.	The women **cleans** every public park.		✓
	regular_plural_subject_verb_agreement_1	76	89	73	97	95	Jeffrey **hasn't** criticized Donald.	Jeffrey **haven't** criticized Donald.	✓	
	regular_plural_subject_verb_agreement_2	81	83	85	96	95	The dress **crumples**.	The dresses **crumples**.		✓

Table 4: Examples of all 67 paradigms in BLiMP along with model performance and estimated human agreement.

simple LM method. The bolded word is the critical word—the probability of the two different critical words for the acceptable and unacceptable sentences can be compared based on the same 'prefix'.

- If a sentence has a checkmark (✓) under the *2pfx* column, the sentence can be used with the 2-prefix method in addition to the simple LM method. The bolded word is the critical word—the probability of that particular word can be compared based on the two different acceptable and unacceptable 'prefixes'.

References

Marantz, Alec. 2013. Verbal argument structure: Events and participants. *Lingua*, 30:152–168. Elsevier.

David Adger. 2003. *Core Syntax: A Minimalist Approach*. Oxford University Press Oxford.

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *Proceedings of ICLR Conference Track*. Toulon, France.

Aixiu An, Peng Qian, Ethan Wilcox, and Roger Levy. 2019. Representation of constituents in neural language models: Coordination phrase as a case study. *arXiv preprint arXiv:1909.04625*.

Kathryn Bock and Carol A. Miller. 1991. Broken agreement. *Cognitive Psychology*, 23(1):45–93.

Rui P. Chaves. 2020. What don't RNN language models learn about filler-gap dependencies? In *Proceedings of the Third Meeting of the Society for Computation in Linguistics (SCiL)*.

Stanley F. Chen and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394.

Gennaro Chierchia. 2013. *Logic in Grammar*. Oxford University Press.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press.

Noam Chomsky. 1981. *Lectures on Government and Binding*.

Shammur Absar Chowdhury and Roberto Zamparelli. 2018. RNN simulations of grammaticality judgments on long-distance dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 133–144.

Shammur Absar Chowdhury and Roberto Zamparelli. 2019. An LSTM adaptation study of (un) grammaticality. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 204–212.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single &!#* vector: Probing sentence embeddings for linguistic properties. In *ACL 2018-56th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 2126–2136.

Jillian K. Da Costa and Rui P. Chaves. 2020. Assessing the ability of transformer-based neural models to represent structurally unbounded dependencies. In *Proceedings of the Third Meeting of the Society for Computation in Linguistics (SCiL)*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988. Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Allyson Ettinger, Ahmed Elgohary, Colin Phillips, and Philip Resnik. 2018. Assessing composition in sentence vector representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1790–1801. Association for Computational Linguistics.

Richard Futrell, Ethan Wilcox, Takashi Morita, and Roger Levy. 2018. RNNs as psycholinguistic subjects: Syntactic state and grammatical dependency. *arXiv preprint arXiv:1809.01329*.

Bart Geurts and Rick Nouwen. 2007. 'At least' et al.: The semantics of scalar modifiers. *Language*, pages 533–559.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2003. English gigaword. *Linguistic Data Consortium, Philadelphia*, 4(1):34.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2019. Colorless green recurrent networks dream hierarchically. *Proceedings of the Society for Computation in Linguistics*, 2(1):363–364.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696.

Michael Heilman, Aoife Cahill, Nitin Madnani, Melissa Lopez, Matthew Mulholland, and Joel Tetreault. 2014. Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 174–180.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

Jaap Jumelet and Dieuwke Hupkes. 2018. Do language models understand anything? On the ability of LSTMs to understand negative polarity items. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 222–231.

Katharina Kann, Alex Warstadt, Adina Williams, and Samuel R. Bowman. 2019. Verb argument structure alternations in word and sentence embeddings. *Proceedings of the Society for Computation in Linguistics*, 2(1):287–297.

Dave Kush, Terje Lohndal, and Jon Sprouse. 2018. Investigating variation in island effects. *Natural Language & Linguistic Theory*, 36(3):743–779.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, 41(5):1202–1241.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *CoRR*, abs/1609.07843.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning, Technical report, OpenAI.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

Florencia Reali and Morten H. Christiansen. 2005. Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence. *Cognitive Science*, 29(6):1007–1028.

Ivan A. Sag, Thomas Wasow, and Emily M. Bender. 2003. *Syntactic Theory: A Formal Introduction*, 2nd edition. CSLI Publications.

Carson T. Schütze. 1996. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. University of Chicago Press.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534.

Dominique Sportiche, Hilda Koopman, and Edward Stabler. 2013. *An Introduction to Syntactic Analysis and Theory*, John Wiley & Sons.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *Proceedings of ICLR*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *33rd Conference on Neural Information Processing Systems*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Alex Wang, Ian F. Tenney, Yada Pruksachatkun, Katherin Yu, Jan Hula, Patrick Xia, Raghu Pappagari, Shuning Jin, R. Thomas McCoy, Roma Patel, Yinghui Huang, Jason Phang, Edouard Grave, Haokun Liu, Najoung Kim, Phu Mon Htut, Thibault F'evry, Berlin Chen, Nikita Nangia, Anhad Mohananey, Katharina Kann, Shikha Bordia, Nicolas Patry, David Benton, Ellie Pavlick, and Samuel R. Bowman. 2019b. `jiant` 1.2: A software toolkit for research on general-purpose text understanding models. `http://jiant.info/`.

Gregory Ward and Betty Birner. 1995. Definiteness and the English existential. *Language*, pages 722–742.

Alex Warstadt and Samuel R. Bowman. 2019. Linguistic analysis of pretrained sentence encoders with acceptability judgments. *arXiv preprint arXiv:1901.03438*.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretič, and Samuel R. Bowman. 2019a. Investigating BERT's knowledge of language: Five analysis methods with NPIs. In *Proceedings of EMNLP-IJCNLP*, pages 2870–2880.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019b. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221.

Ethan Wilcox, Peng Qian, Richard Futrell, Miguel Ballesteros, and Roger Levy. 2019. Structural supervision improves learning of non-local grammatical dependencies. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3302–3312.