Learnability of Indirect Evidence in Language Models

Anonymous ACL submission

Abstract

What kinds of and how much data is necessary for language models to acquire grammatical knowledge to judge sentence acceptability? Recent language models still have much room for improvement in their data efficiency compared to humans. In this paper, we investigate whether language models efficiently use indirect data (*indirect evidence*), from which they infer sentence acceptability. In contrast, humans use indirect evidence efficiently, which is considered one of the inductive biases contributing to efficient language acquisition. To explore this question, we inject synthetic instances with newly coined *wug* words into pretraining data and explore the model's behavior on evaluation data that assess grammatical acceptability regarding those words. We prepare the injected instances by varying their levels of indirectness and quantity. Our experiments surprisingly show that language models do not acquire grammatical knowledge even after repeated exposure to instances with the same structure but differing only in lexical items from evaluation instances in certain language phenomena. Our findings suggest a potential direction for future research: developing models that use latent indirect evidence to acquire grammatical knowledge.

1 Introduction

Current language models, which have made significant progress in various tasks in recent years, are trained on large-scale data. For instance, recent large language models are trained on data thousands of times larger than the amount of data that children are exposed to acquire the same level of grammatical knowledge as adults (Warstadt et al., 2023). This implies that there is much room for improvement in their learning efficiency.

According to Pearl and Mis (2016), humans acquire language using *indirect* evidence, in addition to *direct* evidence, which is considered one of the inductive biases contributing to efficient language acquisition. As shown on the left side of Figure 1, when humans are exposed to the sentence "<wug> loves himself.", they can correctly judge the grammatical acceptability between "<wug> loves himself." and "* <wug> loves herself." Such observed sentences are referred to as *direct* evidence. Conversely, in the middle and right sides of the figure, we assume that humans are not exposed to such direct evidence. However, if they observe sentences from which they can make some inference for a correct judgment, such sentences are called *indirect* evidence. For example, humans can hypothesize that "him(self)" in "<wug> is helping himself." indicates <wug> or that the possessive pronoun "his" in "<wug> helped his friend." indicates <wug> has a male property.

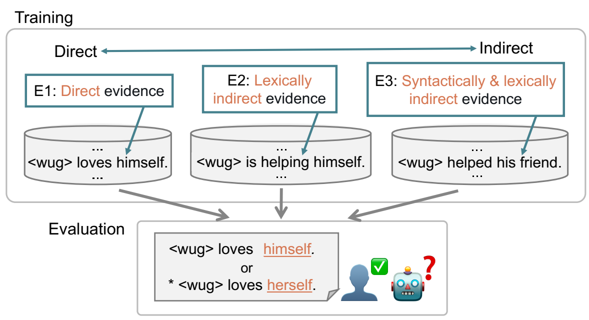

Figure 1: The indirectness of evidence. Direct evidence refers to instances identical to previously observed ones. Lexically indirect evidence targets the same linguistic knowledge but differs in lexical items. Syntactically & lexically indirect evidence is different in both their syntactical and lexical items.

However, whether language models acquire grammatical knowledge using indirect evidence remains unknown. Previous work has investigated the word frequency effect through few-shot learning or ablating sentences including target words from pretraining data (Wei et al., 2021; Yu et al., 2020), but they have not explored the learnability of indirect data in pretraining language models.

In this work, we investigate the degree of indirectness and the amount of data required for language models to induce linguistic generalization. To address this question, we train language models from scratch using pretraining data including indirect training instances. We then evaluate their linguistic generalization across seven different linguistic phenomena, such as anaphor agreement, transitivity, and subject-verb agreement. These phenomena require language models to understand the diverse properties and multiple parts of speech of specific words to judge their acceptability. To control the number of observed indirect training instances, we inject synthetic instances with newly coined words into pretraining data. Following Berko (1958), we refer to those words that do not appear in the original vocabulary and data as *wug* words.[1] We use varied synthetic data as additional indirect training instances, each differing in the degree of lexical and syntactic indirectness and in the number of observations.

We found that the language models generalize linguistic knowledge from training instances that are the same as correct evaluation instances, but their data efficiency varies across different linguistic phenomena. This variation is likely due to the number of words between the *wug* and the words that serve as cues for the model to learn its properties. We surprisingly observe that the language models do not acquire grammatical knowledge in certain phenomena even from instances that only differ in their lexical items. Syntactically indirect instances rarely induce the model's generalization. In a certain phenomenon, we observe that language models had drastically accelerated linear generalizations (Mueller et al., 2022; McCoy et al., 2020).

Given that distances might cause the inefficiency in language models, we conduct a detailed analysis of indirect instances with complicated interference, using anaphor gender agreement as a case study. We examine whether those instances affect the generalization, considering three factors related to attractors and distance. We find that when the language models are trained on the instances containing complicated interference, they stagnate in learning after sufficient observations.

Those findings from controlled and comprehensive experiments suggest that, at least in our small-scale settings, language models cannot generalize in a human-like manner even from the data with a degree of indirectness that seems intuitively manageable for humans, depending on language phenomena. This limitation indicates a direction for future studies: implementing a model that can use indirect evidence, which will lead to data-efficient language acquisition comparable to that of humans.[2]

2 Background

2.1 Evidence in Language Acquisition

In the field of language acquisition, the information used to learn grammatical knowledge is referred to as *evidence*. Positive (negative) evidence refers to information in data indicating what is acceptable (unacceptable) in a language, and it has been argued that humans use only positive evidence to acquire their language (Chomsky, 1993). Pearl and Mis (2016) further distinguishes indirect positive evidence from direct positive evidence. Direct positive evidence indicates the information that appears in the data observed by the learner and is used for learning, under the assumption that speakers' usage of it guarantees grammaticality (the left side of Figure 1). Indirect positive evidence, on the other hand, refers to a type of information that requires a learner to infer from the observed data what is grammatical in the language (the middle and right side of Figure 1). They argue that, in addition to direct positive evidence, indirect positive evidence potentially plays a significant role in efficient language acquisition. While the previous literature explores humans' capacity, it is still unknown whether language models induce linguistic generalization from such evidence.

2.2 Analysis of Language Models in Learning Grammatical Knowledge

NLP research has focused on how language models learn grammatical knowledge regarding the appearance of target lexical items in training data.

Yu et al. (2020) report that only a few examples suffice for learning grammatical knowledge of subject-verb agreement and reflexive agreement in few-shot learning. Wei et al. (2021) also analyze the frequency effect in BERT (Devlin et al., 2019) when learning subject–verb agreement. They find

[1]The original *wug* used in Berko (1958)'s work is not exactly same as our settings to create controlled instances. The details are discussed in Section 7.

[2]We will make our training and evaluation data publicly available.

that BERT can judge the agreement even for unseen subject–verb pairs, which is influenced by the frequency of target verb forms in the training data. The authors focus on the frequency effect of verb forms by removing sentences that contain verbs of interest from the pretraining corpora.

While the fingings from these studies imply strong generalizability in language models, they present several future research directions: (i) exploring a wider range of linguistic phenomena across various parts of speech, (ii) examining the model's learnability of lexically and syntactically indirect sentences, and (iii) investigating alternative learning paradigms beyond few-shot learning with pretrained models and pretraining models on ablated targeted sentences, to align more closely with human language acquisition processes and conduct more controlled experiments. In this study, we analyze the effect of evidence strength in learning grammatical knowledge by dissecting direct and indirect evidence into several levels of evidence strength, along with their frequency effect, with a wider variety of linguistic phenomena across various parts of speech.

While using artificial languages in analyzing language models is tackled by previous work (White and Cotterell, 2021; Ri and Tsuruoka, 2022), our approach is different in that we use a small number of artificial instances only at the token level by introducing a word *wug* to precisely investigate their effect in learning grammatical knowledge.

3 Our Motivations

We aim to clarify how many exposures to a word and what types of sentences containing the word are required for language models to accurately understand its properties to judge the acceptability of a sentence correctly. In this work, we employ newly coined words (*wugs*) to control injections in the pretraining corpus. The advantages include:

- Handling the occurrences of target lexical items may not fully remove the influence of those words from the pretraining corpus. To completely cancel out the effect of a lexical item, we need to remove all variants with the same stem form or subword, which can be intricate and have a risk of significantly distorting the natural distribution of the corpus.
- When automatically generating wug words, we can adequately control their frequency and evidence strength, including their tokenization. Since our aim here is to control the minimal information observable by the model, synthetic data enables the elimination of noises.
- Our approach is a type of data augmentation, which means that no modification of lexical items or sentences in corpora is required. Hence, this approach can be extended easily to other corpora and models.

4 Data

This section describes how we construct our evaluation and additional training instances.

Following targeted syntactic evaluation (Linzen et al., 2016; Marvin and Linzen, 2018; Warstadt et al., 2020), we use pairs of sentences that minimally differ in target words.

4.1 Evaluation Data

Linguistic Phenomena We employ the seven kinds of linguistic phenomena listed in Table 1. We selected them from the benchmark BLiMP (Warstadt et al., 2020)[3], based on whether understanding the properties of a single word is sufficient to correctly judge the linguistic phenomena. Because we introduce newly coined words *wug* in this work to investigate the number of observations necessary for generalization, we can only cover limited linguistic phenomena. We expect such phenomena as those related to island effects. As shown in Table 2, the phenomena targeted in this work vary in their properties crucial for accurately judging the evaluation data so that we can analyze model's behaviors from diverse perspectives.

Newly Coined Words *Wug* We employ the tag <wug#n> as a newly coined word to conduct controlled experiments using words that never appeared in the pretraining corpus. This approach does not entirely align with the policy in Berko (1958), which employed words like *wug* and *wuz* that are newly coined but phonologically natural in the target language by using actual subwords. One concerning issue with Berko (1958)'s policy is that the actual subwords can give model hints for correct grammatical judgement, for example by their occurrence in particular position. To eliminate such possible effect of actual subwords, we instead use the tag <wug#n>. We analyze the differences between conditions using tags and the original *wug*

[3]Appendix C shows which phenomena we specifically referenced from the BLiMP in this work.

Phenomena	*Evd*	Training instances	Evaluation instances
Anaphor gender agreement (ANA.GEN.AGR)	DE LexIE SynIE	<wug#n> has devoted herself <wug#n> is painting herself <wug#n> judges his work	<wug#n> has devoted herself *<wug#n> has devoted himself
Anaphor number agreement (ANA.NUM.AGR)	DE LexIE SynIE	the <wug#n> didn't see themselves the <wug#n> can reward themselves the <wug#n> loved its toy	the <wug#n> didn't see themselves *the <wug#n> didn't see itself
Transitive (TRANS.)	DE LexIE SynIE	some trees <wug#n>ed the car no street can <wug#n> the city every lion hunts what no prey can <wug#n>	some trees <wug#n>ed the car *some trees <wug#n>ed
Intransitive (INTRANS.)	DE LexIE SynIE	many rivers should <wug#n> each ethic might <wug#n> a man corrects that the answer will not <wug#n>	many rivers should <wug#n> *many rivers should <wug#n> dogs
Determiner-Noun agreement (D-N AGR)	DE LexIE SynIE	the senators use this <wug#n> a window will open this <wug#n> the <wug#n> sells the house	the senators use this <wug#n> *the senators use these <wug#n>
Subject-Verb agreement (V) (S-V AGR (V))	DE LexIE SynIE	the <wug#n> are leaving any traces the <wug#n> climb few ladders each key can open those <wug#n>	the <wug#n> are leaving any traces *the <wug#n> is leaving any traces
Subject-Verb agreement (S) (S-V AGR (S))	DE LexIE SynIE	the book <wug#n> a shelf every chocolate <wug#n> several bars cats that follows the leader <wug#n> the groups	the book <wug#n> a shelf *the books <wug#n> a shelf

Table 1: Linguistic phenomena and instances. The sentences starting with * are ungrammatical.

Phenomena	POS	Gen.	Num.	(In)Transitive	Long agr
ANA.GEN.AGR.	noun	✓	–	–	✓
ANA.NUM.AGR	noun	–	✓	–	✓
TRANS.	verb	–	–	✓	–
INTRANS.	verb	–	–	✓	–
D-N AGR	noun	–	✓	–	–
S-V AGR (V)	noun	–	✓	–	–
S-V AGR (S)	verb	–	✓	–	–

Table 2: The properties required to judge evaluation data. POS indicates part-of-speech. Gen./Num. indicates gender/number. Long agr. is whether a long agreement is required.

in Section 7. For number agreement, we added <wug#n> without any suffixes to these sentences, expecting the models to infer that <wug#n> is an inflected form based on the sentence structure in which they are embedded. We explore their effects in the model's generalization in Section 7 For the noun subject of S-V AGR (V) and ANA.NUM.AGR, we do not employ any quantifiers and determiners other than "the". This procedure is because quantifiers and determiners affect linguistic generalization, making it unclear which information the language models use as clues for judgment, the number of properties in verbs and reflexive pronouns or those in quantifiers and determiners. Due to the same reason, for the verb in S-V AGR (S), we only employ the present tense and do not employ any auxiliary verbs and tense suffixes. We ensured that <wug#n> remained the same word (i.e., the tag with the same id) in a pair, both grammatical and ungrammatical sentences, because we want the same occurrence of the *wug* in the training data. Otherwise, we compare the probability of ungrammatical sentences with zero *wug* with that of grammatical sentences with *wug*.

Data Generation with LLM To create varied degrees of and balanced corpus, we use GPT-4 Turbo in OpenAI API to generate the training and evaluation templates. To generate balanced training instances with different properties, we generate them separately based on concerning properties, (e.g., Female and male pronouns have the same percentage in ANA.GEN.AGR.). We prompts the GPT-4 to generate balanced, diverse and duplication sentences. We generate evaluation instances and training instances for indirect evidence (LexIE, SynIE) with three different prompts. Subsequently, we get DE by extracting the correct sentence in generated evaluation instances. We generate the setences with placeholders [WUG] and we replace [WUG] with the tag <wug#n>, where the index number i distinguishes the coined words (e.g., <wug#124>). The example of prompts and detailed procedures are shown in Appendix B.

4.2 Additional training instances

We define the following three dergrees of indirectness (DE, LexIE, and SynIE). The difficulty increases in the order of DE, LexIE, and SynIE:

Direct Evidence (DE) An instance that is the exact same as correct evaluation instances. We assume that the properties of *wug* in an evaluation instance are learned by referring to the training instance with the same syntactical and lexical items as the evaluation instance.

Lexically Indirect Evidence (LexIE) An instance that conveys the same syntactic structure as the evaluation instances but uses different lexical items. We assume that the properties of *wug* in an evaluation instance are learned by referring to training instances with the same usage but different lexical items from the evaluation instance.

Syntactically Indirect Evidence (SynIE) An instance that reveals the target linguistic feature with different syntactic and lexical items from evaluation instances. The properties of *wug* in an evaluation instance are learned by referring to the training instance with different syntactic and lexical items from the evaluation instance.

5 Experiments and Results

5.1 Settings

Pretraining Data We randomly sampled 675k sentences (16M words) from English Wikipedia articles and used them as pretraining data.[4] We inject additional training instances. The detailed preprocess and inject additional training instances are in Appendix D. We shuffled and deduplicated sentences and removed ones containing fewer than two words. The data was then lowercased, and periods were removed from the sentences.

Frequency of Injected Instances We compare the language models trained on the pretraining data injected indirect instances that appear n times ($n = 0, 1, 5, 25, 50, 75, 100$) for each instance.

Models We use BabyBERTa (Huebner et al., 2021), which is a minimal variant of RoBERTa (Liu et al., 2019). We modify some hyperparameters due to the pretraining data size. More detailed information is shown in Table 5. We train the tokenizer from scratch using the pretraining data, adding the tags to the vocabulary so that the tokenizer treats each tag as one token.

Evaluation Metrics We prepare 200 template pairs for each linguistic phenomenon. Each template has three different sets of tags, resulting in $200 \times 3 = 600$ pairs. We simply use the accuracy of choosing the grammatical sentence as our evaluation metric. As evaluation metrics, we use pseudo-likelihood[5] normalized by token length because we use evaluation sentences containing the sentence pair each of which has different token lengths. Note that normalization by token length may still result in token-biases (Ueda et al., 2024).

5.2 Main Results

We review the main results by answering our research questions: (i) What degree of and how much data do language models need to acquire grammatical knowledge to judge the acceptability of a sentence? (ii) Are observations showing similar trends in broader categories of linguistic phenomena? The results are shown in Figure 2.

Direct Evidence As for DE, increasing the number of observations generally contributed to linguistic generalization in language models. However, the extent of improvement varied across different linguistic phenomena. In ANA.GEN.AGR and ANA.NUM.AGR, the score increased more gradually, particularly between 25 and 75 occurrences, compared to the other agreement phenomena. This difference might be due to anaphor agreement, which often involves a longer distance between the target words and the words with properties necessary for correct judgment. We thoroughly examine the effects of distance and attractors in Section 6.

Lexically Indirect Evidence In about a half of the phenomena, D-N AGR, S-V AGR (V), ANA.NUM.AGR, and INTRANSITIVE, LexIE induces generalization more slowly but steadily than DE. However, in the remaining half of the phenomena, the language models do not acquire grammatical knowledge necessary to correctly judge acceptability. This result is surprising because LexIE differs only in lexical items from a correct sentence in the evaluation and shares the same syntactical structure. This trend cannot be explained by the properties of Table 2.

Syntactically and Lexically Indirect Evidence In most of the phenomena, SynIE does not induce generalization; the increase in the number of observations did not aid models' generalization but only resulted in a prolonged learning time. In TRANSITIVE, the accuracy of SynIE drastically decreases

[4] Retrieved from https://github.com/phueb/BabyBERTa.

[5] We use the source code in https://github.com/babylm/evaluation-pipeline-2023.

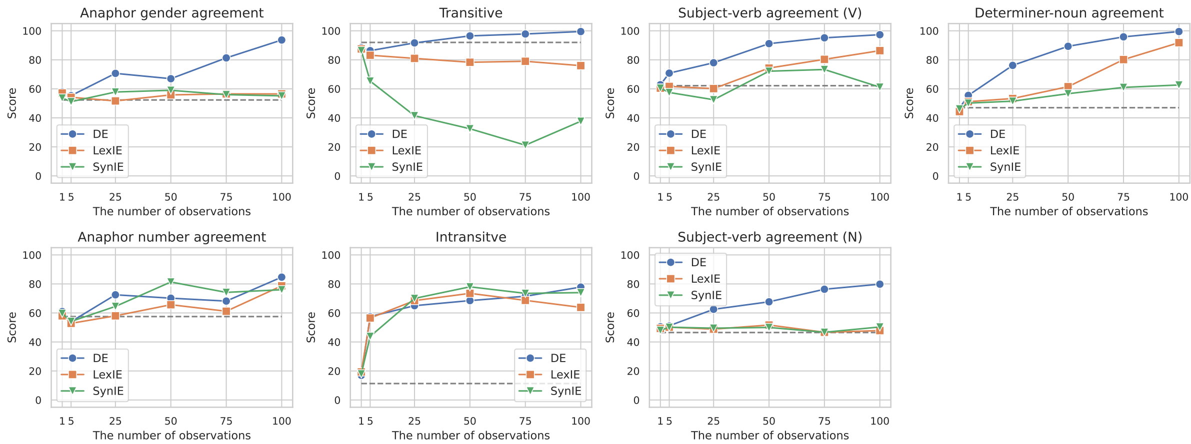

Figure 2: The results (accuracy; %) of experiments for language phenomena and evidence. The gray dot lines indicate the model's scores trained on pretraining data without any additional instances (n=0).

inversely with the number of observations. This interesting phenomenon is likely due to the heuristics of the language model. The final word in the training instances (see Table 1) is the coined word <wug#n>, whereas, whereas it is a actual direct onject noun in the correct evaluation sentences. This suggests that the language model might exhibit linear generalization (Mueller et al., 2022; McCoy et al., 2020), which differs from the human-like hierarchical generalization. It is most likely that they just judged the correctness using whether some words follow the coined words, even though the *wug* should be recognized as a transitive verb because the relative pronoun "what" is its object. This implies that instances requiring complicated hierarchical inference may impair generalization.

Overall Our findings mainly suggest that indirect positive evidence does not sufficiently induce linguistics generalization in language models, especially SynIE, while direct evidence induces it. Wei et al. (2021) find that their results support the Reduce Error Hypothesis (Ambridge et al., 2015), where high-frequency words are learned better. The results in our work also support the hypothesis in DE, but in LexIE and SynIE, not all linguistic phenomena support it.

6 Analysis with More Indirect Instances

In Section 5, DE induced the model's linguistic generalization but its data efficiency varies by linguistic phenomena. For anaphor agreement, the models' learning are more apt to stagnate in 25 – 75 observations compared to other phenomena (See the figure for anaphor agreement in Table 2). This stagnation might be caused by the longer distance between the *wug* and the reflexives, whereas the relevant items are adjacent to each other in other phenomena such as TRANSITIVE. To corroborate this negative effect of long-distance on learning, we employ more indirect agreement instances to investigate whether the long-distance hinders linguistic generalization on ANA.GEN.AGR in language models.

The difficulty of long-distance agreement is caused by attractors and distance (Linzen et al., 2016). Agreement attractors indicate the intervening words that distract the learner from judging the correct agreement (Giulianelli et al., 2018). When language models judge the gender agreement, they would check if the word "<wug#n>" corresponds to the gender to the reflexive. *Distance* refers to the number of the words intervening between the antecedent "<wug#n>" and "herself". *Attractor* indicates the competing words (e.g., "man" in the case of AT1 in Table 2) that distract learners from judging the agreement.

The language models' grammatical knowledge concerning long-distance dependencies has been investigated in previous studies (Giulianelli et al., 2018; Li et al., 2023), and these studies argue that the models can indeed acquire the knowledge of long-distance agreement. However, the overall results on anaphor agreement in this study suggest that further investigation is required to reveal the relationship between models' performance and the distance of items relevant for correct judgment. For this purpose, we conduct a fine-grained analysis using synthetic sentences varying the distance between *wugs* and reflexive pronouns.

Interf.	Evd.	Training instances
Attractor type (AT)	DE	<w> loves herself
	AT0	<w> **helping the child** loves herself
	AT1	<w> **helping the man** loves herself
	AT2	<w> **helping him** loves herself
Attractor number (AN)	DE	<w> loves herself
	AT1	<w> **helping the man** loves herself
	AN0	<w> **helping the man to see the dad** loves herself
	AN1	<w> **helping the man for the king to see the dad** loves herself
	AN2	<w> **helping the man for the son of the king to see the dad** loves herself
Distance (DT)	DE	<w> loves herself
	AT0	<w> **helping the child** loves herself
	DT0	<w> **who helps the child** loves herself
	DT1	<w> **whose cat helps the child** loves herself
	DT2	<w> **whose cat helps the child who finds the teachers** loves herself

Table 3: Interference types and training instances used in the analysis. <w> corresponds to <wug#n>.

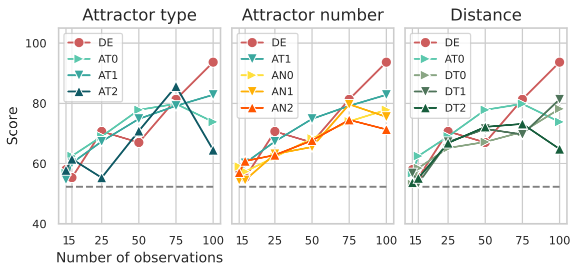

Figure 3: Models' scores for more indirect instances.

6.1 Target Phenomena

We compare the models trained on the corpus with additional instances, from the perspective of the attractor type, attractor number, and distance as below. Table 3 lists all kinds of training instances compared in this analysis.

To create the instances, we use GPT-4 to generate nouns differing in gender and number, and sample the designated number of items from these generated items. For female and male nouns, we collect 100 nouns each. From the generated items, we first select 25 nouns for each gender. Then, we create both the singular and plural forms of the selected words and double them to create minimal pairs. The prompt is shown in Appendix B. Additionally we also collect 100 neutral nouns. The verb that we newly employ is collected from LexIE in ANA.GEN.AGR to avoid duplication.

Attractor Type (AT) We investigate whether attractors downgrade the linguistic generalization in ANA.GEN.AGR and how their distract strength affects the models' acquisition of anaphor agreement. DE indicates the indirect instances examined in Section 5, which does not have any attractors and works as a baseline here. AT0 includes neutral common nouns, while AT1 employs common opposite gender nouns, and AT2 uses opposite gender proper nouns. We assume that the magnitude of attractors' interference follows the order AT0 < AT1 < AT2, given that the more similar their properties are to reflexives, the more distracting they will be.

Attractor Number (AN) We examine whether the number of attractors affects the model's acquisition. We use the gender common nouns as attractors. DE works as a baseline because it has no attractors. We expect that the more attractors there are, the more difficult it is to generalize correctly.

Distance (DT) We analyze the effect of distance on model's acquisition. We assume that the more distance intervening between *wug* and reflexive, the more difficult it is to judge sentence acceptability. We use neutral nouns there to explore the effect of the number of words genuinely.

6.2 Results

As shown in Figure 3, After 100 observations in all viewpoints, SynIE, with the shortest distance and no attractors, got the highest scores, while in midway observations this tendency does not happen. The most difficult instances in each interference lead to the language model's lowest score, after their 100 observations. AT2, including an opposed pronoun as an attractor, particularly shows unstable generalization. We expected that the instances with longer distances and more attractors, more strongly interfere with the models' generalization, but this tendency is not clearly shown in this experiment. To the question of whether the instances with long-distance agreement induce linguistic generalization, these results answer that with the larger number of observations, the model's generalization relatively stagnates.

7 Discussion: Considering *Wug* Creation

In this work, we use to newly coined words that do not appear in the original vocabulary, following Berko (1958). Still, our used *wug* has some gap

N	*wug* methods	Phenomena		
		ANA. NUM. AGR	D-N AGR	S-V AGR (V)
0	*tag*	57.5	47.0	62.2
	tag w/ morph.	59.0	80.5	83.3
	wug_v1	81.3	89.5	86.7
	wug_v2	81.2	91.2	86.0
	wug_v3	81.5	88.7	85.0
25	*tag*	72.5	76.2	78.0
	tag w/ morph.	94.0	99.5	91.3
	wug_v1	92.3	87.7	90.2
	wug_v2	81.2	87.7	88.5
	wug_v3	90.5	87.5	86.5

Table 4: Models' scores calculated by the language models that are trained on the pretraining data with indirect instances of different *wug* creation methods. *N* denotes the number of observations.

from the original one. In the original *wug* test, they use the words that do not exist in the language but conform to the phonological rule in the language, In contrast, we use the tag <wug#n> as *wug* in those experiments. Since the original *wug* is more phonologically natural, and the subwords are in the existing vocabulary, the original setting is closer to the environment of human language acquisition. On the other hand, to conduct controlled experiments on the number of instances that the model observed, the setting might not be suitable because this is far from the settings where a certain word is never encountered. We used the tag <wug#n>. In this section, we compare our method (*tag* method) and the original method (*wug* method) to explore the difference in their impact on the model's linguistic generalization.

***Wug* Generation** We create *wug* using pseudoword generator Wuggy.[6] We choose 1,200 nouns from sample data taken from the one billion-word Corpus of Contemporary American English (COCA).[7] To create *wug*-like words, we use the nouns to output four pseudo words for one noun and randomly select one pseudo noun. We prepare $200 \times 3 = 600$ pseudo words, each 200 of which are used separately (*wug_v1*–*wug_v3*) because we expect that different *wug*s have different subwords and they can show different results. [8] We use those pseudo nouns instead of the tag in the same way as in the previous experiments.

Settings We target three phenomena, ANA.NUM.AGR, D-N AGR, and S-V AGR (V), the *wug* of which is considered as common nouns. No inflectional morphemes are added to plural common nouns in the *tag* method while the morphemes are added to plural common nouns in the *wug* method. For ablation, we prepare the tag with inflectional morphemes (*tag w/ morph.* method), which employs the tag <wug#n> same as the *tag* method but uses inflectional morphemes same as the *wug* method. We compare the models trained on the pretraining data with the *tag* method, the *wug* methods, and *tag w/ morph.* method. Other settings are the same as Section 5.

Results Figure 4 shows the scores of the *tag*, *tag w/ morph.*, and three sets of *wug*. In the *wug* and *tag w/ morph.* methods, the language models correctly judge the acceptability of sentences, mostly more than 80–90%, surprisingly with the data that includes zero additional instances. This result is probably because language models determine whether a word is singular or plural, based on whether an inflection morpheme "s" follows it, even if the word is novel. This occurs with both novel words and novel subword combinations, but the impact is greater with the latter, comparing the two methods. In addition, despite our expectation that different subword combinations show different results, we observed no large score variances among the three vocabulary sets except for 25 times in ANA.NUM.AGR. From those results, we found a trade-off between the settings plausible for human language acquisition and strictly controlled settings. We prioritized the latter in this work, but the direction to the former is also a good setting depending on the research questions.

8 Conclusion

We investigate the degree of indirectness and the amount of data required to induce human-like linguistic generalization in language models. We found that language models do not induce human-like linguistic generalization even with a degree of indirectness that seems intuitively manageable for humans, depending on language phenomena. This limitation indicates a direction for future studies: implementing a model that can use indirect evidence, which will lead to data-efficient language acquisition comparable to that of humans.

[6] https://github.com/WuggyCode/wuggy

[7] Downloaded from https://www.wordfrequency.info/samples/words_219k.txt

[8] On the other hand, for *tag* and *tag w/ morph.*, we show the results of only one model, because the different *tags* <wug#n> have the same parameters and they actually show the same results.

Limitations

We recognize the following limitations in this study:

Linguistic Knowledge by Function Words We generate synthetic instances only for linguistic phenomena concerning content words such as nouns and verbs. We avoid generating new function words (e.g., new *wh*-word as a relative pronoun).

Nonce Sentence We have not dug into the difference between natural sentences and nonce sentences (Gulordava et al., 2018; Wei et al., 2021) that are grammatical but completely meaningless because we create additional training and evaluation instances with LLM, which tends to generate naturally plausible sentences. Nonce sentences are less plausible in human language acquisition but exclude semantic selectional-preferences cues (Gulordava et al., 2018; Goldberg, 2019). According to Section 7, there can be a trade-off between training language models in experimental settings that closely resemble natural human language acquisition and those that are strictly controlled. Future work can determine whether nonce sentences with indirect evidence differently affect linguistic generalization in language models.

Zero Observations While adding the tags <wug#n> into the vocabulary, their parameters in language models are randomly initialized. When the language models never observe the sentences including the tag while training, their parameters still remain initialized, which may lead to different results in language models. To confirm this effect, we compare the language model with the default standard deviation of the initializer for all weight matrices (std=0.02) to that with one-tenth standard deviation (std=0.002), using three kinds of seeds. Table 6 in Appendix E shows that the deviation of scores in the model used one-tenth std are smaller. This finding implies that a smaller std would contribute to the stability of the results. However, too small std may pose a risk of negatively impacting the training process. We thus use default std in the current work.

Limited Model Size and Pretraining Data We use a small-scale language model and pretraining data in this work because we aim to find the differences from human inductive biases as much as possible. It is uncertain that the same trends as our work will appear in models of any size. Whether scaling laws apply to indirect data in accelerating model generalization would be an interesting future work.

Ethics Statement

There might be a possibility that the texts we used (Wikipedia) and the sentences generated by large language models are socially biased, despite their popular use in the NLP community.

References

Ben Ambridge, Evan Kidd, Caroline F. Rowland, and Anna L. Theakston. 2015. The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, 42(2):239–273.

Jean Berko. 1958. The child's learning of english morphology. *WORD*, 14(2-3):150–177.

Noam Chomsky. 1993. *Lectures on Government and Binding*. De Gruyter Mouton, Berlin, New York.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248, Brussels, Belgium. Association for Computational Linguistics.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online. Association for Computational Linguistics.

Bingzhi Li, Guillaume Wisniewski, and Benoît Crabbé. 2023. Assessing the capacity of transformer to abstract syntactic representations: A contrastive analysis based on long-distance agreement. *Transactions of the Association for Computational Linguistics*, 11:18–33.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach.

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, Brussels, Belgium. Association for Computational Linguistics.

R. Thomas McCoy, Robert Frank, and Tal Linzen. 2020. Does syntax need to grow on trees? sources of hierarchical inductive bias in sequence-to-sequence networks. *Transactions of the Association for Computational Linguistics*, 8:125–140.

Aaron Mueller, Robert Frank, Tal Linzen, Luheng Wang, and Sebastian Schuster. 2022. Coloring the blank slate: Pre-training imparts a hierarchical inductive bias to sequence-to-sequence models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1352–1368, Dublin, Ireland. Association for Computational Linguistics.

Lisa S. Pearl and Benjamin Mis. 2016. The role of indirect positive evidence in syntactic acquisition: A look at anaphoric "one". *Language*, 92(1):1–30.

Ryokan Ri and Yoshimasa Tsuruoka. 2022. Pretraining with artificial language: Studying transferable knowledge in language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7302–7315, Dublin, Ireland. Association for Computational Linguistics.

Naoya Ueda, Masato Mita, Teruaki Oka, and Mamoru Komachi. 2024. Token-length bias in minimal-pair paradigm datasets. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16224–16236, Torino, Italia. ELRA and ICCL.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Jason Wei, Dan Garrette, Tal Linzen, and Ellie Pavlick. 2021. Frequency effects on syntactic rule learning in transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 932–948, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jennifer C. White and Ryan Cotterell. 2021. Examining the inductive bias of neural language models with artificial languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 454–463, Online. Association for Computational Linguistics.

Charles Yu, Ryan Sie, Nicolas Tedeschi, and Leon Bergen. 2020. Word frequency does not predict grammatical knowledge in language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4040–4054, Online. Association for Computational Linguistics.

A Hyperparameters

Hyperparameters in our work are listed in Table 5.

B Prompts

The example of prompts is in Figure 4.

C Linguistic phenomena

We employ seven linguistic phenomena, following (Warstadt et al., 2020), to create training/evaluation instances. The linguistic phenomenon "transitive" is from "causative", "intransitive" is from "drop_arguement", "determiner-noun agreement" is from "determiner_noun_agreement_2", "subject-verb agreement (V)" is from "regular_plural_subject_verb_agreement_1", and "subject-verb agreement (S)" is from "regular_plural_subject_verb_agreement_2".

```
Create 400 minimal sentence pairs, containing a grammatical and an ungrammatical sentence, following the template pair and rules.
    Template pair:
    [WUG] <singular transitive verb> herself.
    [WUG] <singular transitive verb> himself.
    Rules:
    - You must include the lemma of <singular transitive verb> with a different initial letter and different final letter from the previous ones.
    - Always use the female proper noun [WUG] with bracket[] and uppercase.
    - You must include various auxiliary verbs and tenses in <singular transitive verb> with a different initial letter and different final letter from the previous ones.
    - You often include negations in <singular transitive verb> if previous pairs did not contain ones.
    - Do not include adverbs.
    - Generate 400 pairs including numbering that starts from 1 and ends at 400.
    Example:
    [WUG] will hurt herself.
    *[WUG] will hurt himself.
```

Figure 4: Prompts used to create evaluation examples.

Model	architecture	roberta-base
	vocab size	9,600
	hidden size	512
	heads	8
	layers	8
	dropout	0.1?
	layer norm eps	1e-12?
	initializer range	0.02
Optimizer	algorithm	AdamW
	learning rates	2e-4
	betas	(0.9, 0.999)
	weight decay	0.0
Scheduler	type	linear
	warmup updates	24,000
Training	gradient accumulation	4
	epoch	18
	batch size	16
	line by line	true
	NGPU	1

Table 5: Hyperparameters of the language models.

D Data generation

D.1 Pretraining Data

We aim to pretrain the language models for 18 epochs while controlling the number of occurrences of target instances. To achieve this, we concatenate the pretraining data 18 times consecutively and randomly select where to inject each additional training instance.

D.2 Creating data with LLM

The GPT-4 sometimes inconsistently generates sentences with hallucination; it generates the same sentence repeatedly and sometimes stops generating midway. To generate lexically diverse instances as many instances as possible, we prompt GPT-4 to avoid using the same lemma as in the previous instance. To get appropriate instances, we prompt the GPT-4 to generate double the number of instances[9], and then select the designated number of instances, avoiding duplicates. We adjust the percentage of sentences with negation words to 10–50%. The balanced instances resulted in containing 100 female and 100 male instances in ANA.GEN.AGR, 34 female singular and 33 male singular, 34 singular and 100 plural instances in ANA.NUM.AGR, 200 instances each in TRANSITIVE and INTRANSITIVE, 50 this, 50 that, 50 these and 50 those in D-N AGR. 100 singular and 100 plural each in S-V AGR.

E Different Seeds

The scores of language models with different seeds and the standard deviation of the initializers are listed in Table 6.

[9]The number of instances generated based on the prompt can vary. Sometimes the output meets the specified quantity, while other times it may be fewer, potentially even less than half of the requested amount. If not enough instances are generated, we input instances from three steps earlier and generate additional instances to meet the requirements.

phenomena	seed	std 0.02	std 0.002
ANA.GEN.AGR	1	52.3	56.5
	2	51.0	53.5
	3	50.5	56.5
ANA.NUM.AGR	1	57.5	62.8
	2	62.3	67.7
	3	59.3	62.7
TRANSITIVE	1	92.0	90.7
	2	89.0	90.7
	3	89.7	88.7
INTRANSITIVE	1	11.3	11.5
	2	12.3	11.8
	3	14.3	12.7
D-N AGR	1	47.0	47.3
	2	49.0	50.7
	3	46.3	48.7
S-V AGR (V)	1	62.2	53.2
	2	52.0	54.3
	3	55.0	56.7
S-V AGR (S)	1	46.5	49.2
	2	48.3	50.7
	3	52.3	48.3

Table 6: Scores of langauge models with different seeds and standard deviation of the initializers.