Language acquisition: do children and language models follow similar learning stages?

Linnea Evanson
Meta AI Paris;
Laboratoire des systèmes perceptifs
École normale supérieure
PSL University
linnea.evanson8@gmail.com

Yair Lakretz[*]
Cognitive Neuroimaging Unit
CEA, INSERM
Université Paris-Saclay
NeuroSpin Center
yair.lakretz@gmail.com

Jean-Rémi King[*]
Meta AI Paris;
Laboratoire des systèmes perceptifs
École normale supérieure
PSL University
jeanremi@meta.com

Abstract

During language acquisition, children follow a typical sequence of learning stages, whereby they first learn to categorize phonemes before they develop their lexicon and eventually master increasingly complex syntactic structures. However, the computational principles that lead to this learning trajectory remain largely unknown. To investigate this, we here compare the learning trajectories of deep language models to those of children. Specifically, we test whether, during its training, GPT-2 exhibits stages of language acquisition comparable to those observed in children aged between 18 months and 6 years. For this, we train 48 GPT-2 models from scratch and evaluate their syntactic and semantic abilities at each training step, using 96 probes curated from the BLiMP, Zorro and BIG-Bench benchmarks. We then compare these evaluations with the behavior of 54 children during language production. Our analyses reveal three main findings. First, similarly to children, the language models tend to learn linguistic skills in a systematic order. Second, this learning scheme is parallel: the language tasks that are learned last improve from the very first training steps. Third, some – but not all – learning stages are shared between children and these language models. Overall, these results shed new light on the principles of language acquisition, and highlight important divergences in how humans and modern algorithms learn to process natural language.

1 Introduction

Language acquisition is marked by a series of successive stages (Dupoux, 2018; Kuhl, 2004; Werker, 2018). Within their first year of existence, humans infants successively acquire prosody contours (Mehler et al., 1988), phonetic categories (Werker and Tees, 1984; Kuhl et al., 1992; Mazuka et al., 2011) and frequent words (Tincoff and Jusczyk, 1999; Bergelson and Swingley, 2012). They then learn to produce basic syntactic structures (*e.g.* "The boy sang" or "The boy fell"), questions (*e.g.* "What sound does a cow make?") and nested syntactic structures (*e.g.* "The boy that I saw sang"), at approximately 12, 30, and 42 months, respectively (Friedmann et al., 2021). Even though some children may take slightly longer to learn than others, there is a set order in which children acquire various syntactic structures (Friedmann and Reznick, 2021).

Our understanding of the entire learning trajectory of children remains very coarse, however. This partly stems from the difficulty of measuring linguistic skills in young children. In babies, experimenters typically measure eye gaze and sucking rate while children process linguistic stimuli, as these reflexive behaviors are known to increase during surprising events. Such "implicit" approaches have successfully been used to assess whether non-speaking infants detect linguistic violations (Zamuner, 2006), distinguish lexical from grammatical words (Shi et al., 1999) or discriminate their native language from a foreign language (Mehler et al., 1988; Kuhl et al., 2006; Nazzi et al., 2000). In older children, linguistic skills can also be more explicitly measured from spontaneous speech and sentence repetition. For example, a recent study by Friedmann et al. (2021), to which we compare our work in this paper, quantified the extent to which 18 month to 6 year-old children produce variably complex syntactic structures. For both of these approaches, however, the measures from children at such early ages can be noisy and fragmented.

Interestingly, these issues do not apply to modern language models. Deep learning architectures trained to predict words from their proximal contexts have proved immensely effective at learning to process natural language (Radford et al., 2019; Devlin et al., 2019). Unlike humans, these algorithms can be easily probed during training, at any time point and rate, and with unlimited number of

[*] Equal Contribution

Findings of the Association for Computational Linguistics: ACL 2023, pages 12205–12218
July 9-14, 2023 ©2023 Association for Computational Linguistics

test stimuli, without interfering with their language acquisition (Jawahar et al., 2019; Manning et al., 2020; Bowman and Dahl, 2021). Furthermore, high-performing deep nets have been shown to implicitly (Lakretz et al., 2019; Gulordava et al., 2018) or explicitly learn to represent and use syntactic structures (Manning et al., 2020), as well as to use features such as concreteness and lexical class to learn language (Chang and Bergen, 2022). Finally, and importantly, these deep neural networks have recently been shown to represent lexical, syntactic and compositional representations similarly to the adult brain (Jain and Huth, 2018; Toneva and Wehbe, 2019; Caucheteux and King, 2022; Pasquiou et al., 2022, 2023; Caucheteux et al., 2023). Evidencing similar learning trajectories in children and language models could thus provide an invaluable framework to better understand the computational principles underlying language acquisition.

Here, we compare the trajectory of language acquisition between human children and modern language models. We focus on three main questions. First, do these models learn linguistic skills in a systematic order? Second, is this trajectory sequential or parallel? Third, is this trajectory similar to that of children? These hypotheses are illustrated in Figure 1.

Specifically, we train 48 GPT-2 architectures (Radford et al., 2019) from scratch, using a standard next-word prediction objective. We then evaluate, at each training step, their linguistic abilities with 96 semantic and syntactic probes curated from the BLiMP, Zorro and BIG-Bench benchmarks (Warstadt et al., 2020; Huebner et al., 2021; Srivastava et al., 2022). Finally, we compare a subset of these probes to the behavior of 54 children aged, between 18 months and 6 years (Friedmann et al., 2021).

2 Approach

2.1 Language models

We consider two main language models. First, we use a pretrained language model – GPT-2 – as provided by HuggingFace [1] and pretrained on 40 GB of data (Radford et al., 2019). Second, we separately train 48 versions of a 12-layer GPT-2 model from scratch. We train each model on WikiText103 (Merity et al., 2016) with a distinct random seed to set its initial parameters and data-loader. Each model is evaluated on all linguistic probes every 100 training steps. Further training details are provided in Appendix B.

2.2 Zero-shot linguistic probes

Zero-shot linguistic probes are sentences or phrases crafted to evaluate whether a model has learned a particular linguistic skill, without training or fine-tuning the model on that particular skill. In practice, a zero-shot probe consists of comparing the estimated probability of a grammatical sentence with that of a matched ungrammatical sentence. This two-alternative forced-choice approach can be compared to "acceptability judgements", classically used in linguistics (Warstadt et al., 2019).

We evaluate our models on 96 different linguistic probes, curated from three open source benchmarks, the details of which are presented in Appendix C.

Specifically, we compare the probability of each sentence in a grammatical/ungrammatical pair by evaluating the sum of the logarithm of the loss output by the softmax layer:

$$\sum_{i=0}^{n_g} log(f(X_g)_i) < \sum_{j=0}^{n_u} log(f(X_u)_j) \qquad (1)$$

with f the softmax layer of the language model, X_g and X_u the grammatical and ungrammatical sentences, respectively, and n_g and n_u, the number of tokens in the grammatical and ungrammatical sentences, respectively.

The accuracy of a given probe is the percentage of pairs where the estimated probability of the grammatical sentence is higher than that of the ungrammatical sentence.

2.3 Assessing learning trajectory

To evaluate whether the trajectory of language acquisition is shared across models, we rank the probes by their "acquisition time", *i.e.* the number of steps taken by a model to reach 90% of its final accuracy on a particular probe, for each model independently. We then assess the correlation of ranks between all pairs of the 48 models and take the average of these correlations. To estimate the statistical significance of this average correlation we redo this calculation for all possible model pairs after shuffling the ranks of one of the models in each pair. We repeat this permutation 1,000 times, getting 1,000 values for this shuffled correlation. If in all cases this shuffled correlation is lower than

[1] https://huggingface.co/gpt2

the true average correlation, then the order of acquisition time is shared across models with p $<$ 0.001.

2.4 Parallel versus Sequential learning

Language acquisition may be characterized by a "sequential" or a "parallel" learning scheme (Figure 1). "Sequential" learning designates the case where a complex skill does not start to be learned before simpler skills are mastered. By contrast, "Parallel" learning designates the case where all skills are acquired simultaneously, but at different speeds. The null hypothesis is that the order in which an agent learns linguistic skills varies across agents. To determine the learning scheme of language models, we consider whether the probes have a positive derivative in the first three checkpoints (parallel learning) or not (sequential learning), and whether they have statistically different learning rates (by performing a one-way ANOVA test) across the three groups.

2.5 Assessing linguistic skill from children's behavior

Friedmann et al. (2021) studied 54 Hebrew-speaking children between the ages of 18 - 71 months and investigated the emergence of 11 linguistic phenomena, which the authors propose to organize into three stages (details in Appendix A). For our analysis we select the following tests, one from each stage:

- Stage 1: Simple sentences in subject-verb (SV) order
- Stage 2: Wh Questions
- Stage 3: Relative Clauses

Data collection consisted of spontaneous speech samples produced by each child at home. Each sample was then manually annotated to detect the presence of each of the linguistic phenomena. A linguistic phenomenon was considered learned if and only if it was present in the speech sample. Speech samples had a mean length of 151 utterances per sample and standard deviation of 37. The aggregated data was made available directly in the original paper (under Creative Commons Attribution 4.0 International License), and here used for comparison with our language models. In Table 1 we show which probes in the models matched with these tests.

3 Results

We aim to compare the learning trajectories of deep language models to those observed in 54 children aged between 18 months and 6 years. For this, we trained variants of GPT-2 models (Radford et al., 2019) from 48 different random seeds with the WikiText103 dataset (Merity et al., 2016) and evaluated each model on 96 linguistic probes every 100 steps.

At the end of this training, 64 probes (66%) were achieved above chance level (50% accuracy) by all models. In comparison, a pretrained version of GPT-2 large (Radford et al., 2019) provided by Hugging Face[2], and trained on a much larger dataset[3], achieves above-chance performance on 93 of the 96 probes.

3.1 A systematic learning trajectory

For clarity, we focus on the learning dynamics of the probes that ultimately achieve above-chance performance in our training. Figure 2 lists all probes learned above chance level, ordered by their average acquisition time. We perform the permutation analysis outlined in 2.3, to evaluate whether the order of acquisition is shared between models, and find that their order of acquisition is correlated with $R = 0.743$ and p $<$ 0.001. These results suggest that there is a systematic learning trajectory among models.

3.2 Learning is parallel across linguistic tasks

Are these linguistic skills learned sequentially or in parallel (Figure 1)? To address this question, we evaluate whether each linguistic probe starts to improve from the very first training steps but with different rates (i.e. a “parallel” learning scheme) or, on the contrary, whether some probes only start to improve once others have reached a particular performance (i.e. a “sequential” learning scheme). As the individual learning trajectories of each probe were noisy, we group the 64 linguistic probes into three categories: early, middle and late acquisition time (Figure 3).

Overall, we observe parallel learning between the three groups: their performances all increase from the beginning of training: 95% of tests in all three groups have a positive derivative within the first three hundred steps. However, they have different learning rates, as evaluated with a one-way ANOVA test on the learning rate (i.e. change

[2] https://huggingface.co/tftransformers/gpt2-large

[3] 40 GB compared to the 181 MB of WikiText103

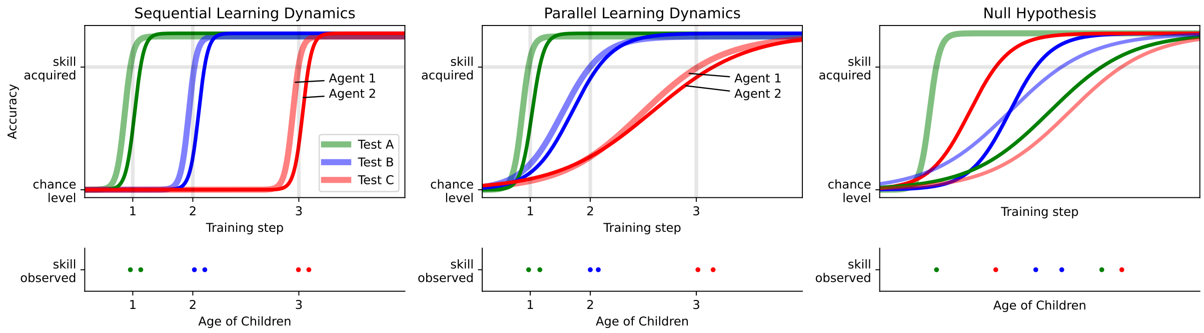


Figure 1: Hypotheses. Skill performance (y-axis) as a function of training (x-axis) illustrated on three tasks (colors) and two agents (model or children). Sequential learning implies that the learning of a complex skill (e.g. C, shown in red) does not start before simplest skills (e.g. A in green and B in blue) are fully learned. Parallel learning implies that all skills are acquired simultaneously, but at different speeds. Sequential and parallel learning may cross an arbitrary performance threshold at the same training step. The frequency at which we can probe artificial networks during learning is much greater than what is realistically possible in children, giving us the timescale granularity to distinguish sequential and parallel learning trajectories. We also present a null hypothesis, that artificial networks with different random seeds may learn skills in a different order.

Stage	Children	Language Model
1	Simple sentences in Subject-Verb (SV) order	SV agreement across simple sentences
2	Wh-questions	SV agreement in questions
3	Relative Clauses (RCs)	SV agreement across object RCs

Table 1: Linguistic phenomena in children selected for comparison with probes in the language models.

of accuracy over time) obtained in each group and in each model ($p < 10^{-23}$).

3.3 Comparison with children

Do these learning trajectories match the behavior of human children? For the three probes that correspond to the three stages identified in children's language acquisition (Table 1), we observe that the order in which these three probes are learned by the language models is the same as those of children (Figure 4). This effect is robust across random seeds: 46 of our 48 GPT-2 models follows this order, where chance level is $(1/3!)^{46} = 1.60e^{-36}$. For this subset of language abilities, models and children thus seem to acquire syntactic skills in a similar order.

3.4 Learning of theoretically-defined stages is parallel

In part 3.1, we showed that GPT-2 learns its language abilities in parallel. Does this learning scheme also characterize the three syntactic skills investigated in children? To address this question, we now look at the learning curves of the skills defined in Table 1, as well as an additional probe: Nounpp, as it can be separated into congruent and incongruent cases which is important for the analysis in section 3.5. Overall, we observe that these probes are all learned in parallel in the model (Figure 5A).

3.5 Models use both syntax and heuristics

Both an understanding of syntactic rules and a superficial heuristics can lead to above-chance performances on these probes. Indeed, in many sentences (*e.g.* The cat [subject] of the lady [attractor] is [verb] hungry.), the number of the verb is congruent with the number of the adjacent attractor, even if the two are not related syntactically. To verify that the GPT-2 models effectively learn the syntactic rules, we thus separately examine congruent and incongruent cases. Incongruent cases require knowledge of the syntax of the sentence as the correct verb number is different from the number of the attractor. Empirically, we observe that the models do not learn the incongruent case in stage three above chance level, and just barely reach chance level on the incongruent case in stage two (Figure **5B**), indicating that our models are using heuristics rather than syntactic rules to achieve high accuracy on the congruent cases (leading to above chance performance on the probe overall in Figure **5A**). On the contrary, the pretrained GPT-

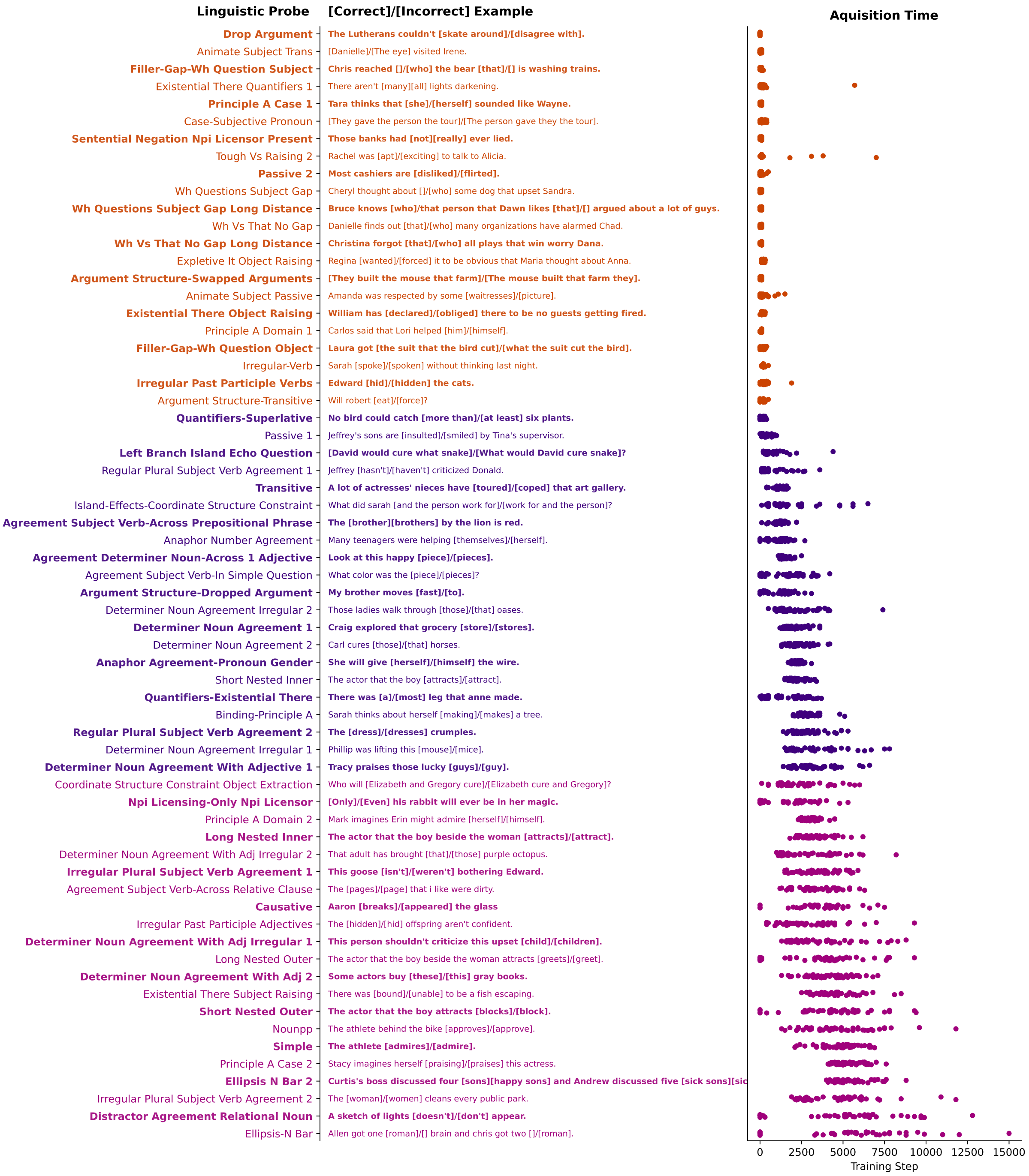


Figure 2: The performance of the models on each linguistic probe over time is smoothed using a moving average filter with window size of 6 checkpoints then the number of steps required to reach 90% of final accuracy (acquisition time) is calculated. Probes are ordered by increasing average acquisition time. Results shown for 48 models. Only probes which have final accuracy greater than chance (50%) are shown. This demonstrates that probes tend to be learned in the same order by all agents with R = 0.743, p < 0.001, disproving the null hypothesis.

2 large achieves above 75% accuracy also on the incongruent cases of these probes. Thus for the models trained on the WikiText103, syntax is only learned for stages one and two, and heuristics seem to explain the above chance accuracy in stage three. A larger training dataset is required to learn syntax and not only heuristics for the most difficult examples.

3.6 Impact of number biases in congruent and incongruent sentences

In previous work, it was found that a variety of language models have a bias towards plural English verbs, and several studies (Jumelet et al., 2019;

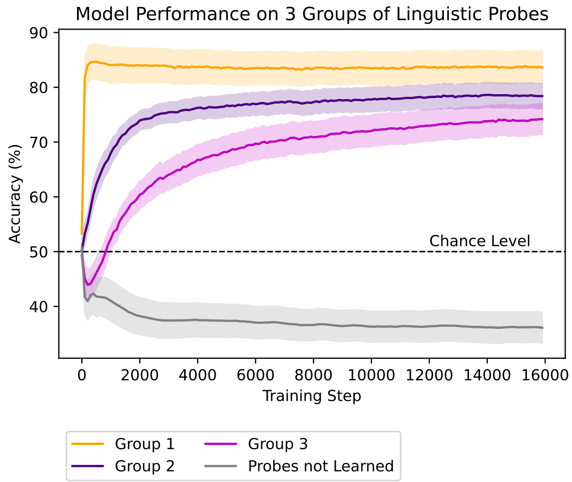


Figure 3: Linguistic probes grouped into 3 groups to avoid plotting one line per probe. The probes in each group correspond to the colours in Figure 2. Probes which have final accuracy less than chance (50%) are placed in their own group, and tend to zero likely due to biases towards plural verbs in English (*c.f.* 3.6). Shading is standard error of the mean across probes in the group. This figure demonstrates that linguistic skills are learned in parallel not in sequence.

Lakretz et al., 2021a,b) determined that LSTM-based models have a default number and gender prediction preference. To examine whether number bias has a significant effect on our analysis, we compare congruent sentences with only singular or only plural verbs and incongruent sentences with a plural or a singular verb. Accuracy on predicting plural verbs increases sharply from the start of the training and then drops. By contrast, the accuracy of singular cases first drops and then rises (Figure **5C**), indicating that the models are biased towards plural verbs at the beginning of training. This bias is overcome for the stage one probe but for stage two and three it remains throughout training. This explains the initial downward trend in Group 3 and why the unlearned probes tend toward zero in Figure 3.

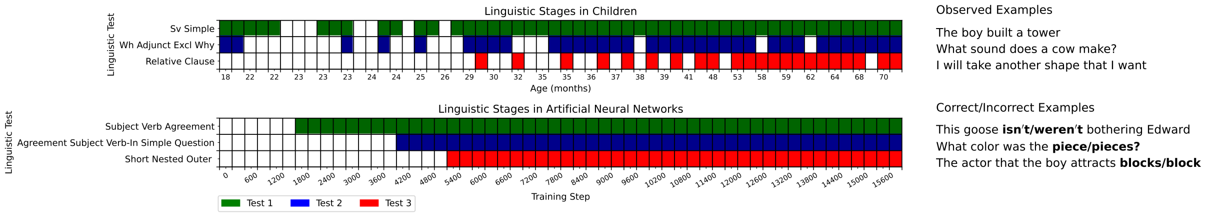
4 Discussion

The stages followed by children to acquire language has been the topic of intense research (Dupoux, 2018; Kuhl, 2004; Werker, 2018). While this learning trajectory is becoming clearer for sub-lexical representations (Dupoux, 2018), the acquisition of higher-level syntactic and semantic processing remains largely unclear. Here, we approach this long-lasting question through the lens of a deep language architecture, GPT-2 (Radford et al., 2019), to test whether this model follows a learning trajectory similar to children.

4.1 Language acquisition: similarities and differences between humans and GPT-2

First, we show that GPT-2 models tend to learn a battery of linguistic phenomena (Warstadt et al., 2020; Lakretz et al., 2019; Huebner et al., 2021) in a consistent order. It is the reliability of the acquisition trajectory that allows a direct comparison with the learning trajectory of children (Friedmann et al., 2021). However, this consistency in GPT-2 models may result from two non-mutually exclusive factors, that remain to be disentangled: either the acquisition time of each linguistic phenomenon relates to its frequency in natural language (e.g. Simple subject-verb-complement are more frequent in natural language than nested syntactic structures; Karlsson 2007), and/or it relates to their intrinsic complexity (e.g. sentences with nested structure require more operations to be composed than simple sentences). Future work systematically controlling for these relative frequencies is thus necessary to distinguish these two possibilities, and would build upon work by Weber et al. (2021) who found that less frequent linguistic phenomena can be learned from fewer examples, though later in training.

Second, we show that the order in which linguistic skills are acquired is similar between children and GPT-2 – at least on the syntactic phenomena that were evaluated in these two cohorts, and with the limitation of using number agreement as a proxy to whether the models acquire the corresponding syntactic structure. Similarly to children, GPT-2 models master subject-verb agreement in SV sentences before they master it in questions, or across nested center-embedded clauses (object-relative clauses). This result thus complements a series of studies comparing modern language models and humans. For example, a recent study showed that transformers trained on child-directed data can achieve comparable accuracy on linguistic probes to large pre-trained models (Huebner et al., 2021). Similarly, several studies have recently shown that the representations of GPT-2 become increasingly similar to those of the adult human brain during its training (Caucheteux and King, 2022). Finally, Lavechin et al. (2022) showed that models trained on audio in a self-supervised fashion learn phoneme and lexical abilities in a similar trajectory to children.



Observed Examples

The boy built a tower
What sound does a cow make?
I will take another shape that I want

Correct/Incorrect Examples

This goose **isn′t/weren′t** bothering Edward
What color was the **piece/pieces?**
The actor that the boy attracts **blocks/block**

Figure 4: Comparing language model performance on linguistic probes to children's performance. Example sentences observed in children were originally in Hebrew (Friedmann et al., 2021). Non-white indicates the phenomena is learned by the agent. The threshold for considering a probe learned by the model is performance above 55%.

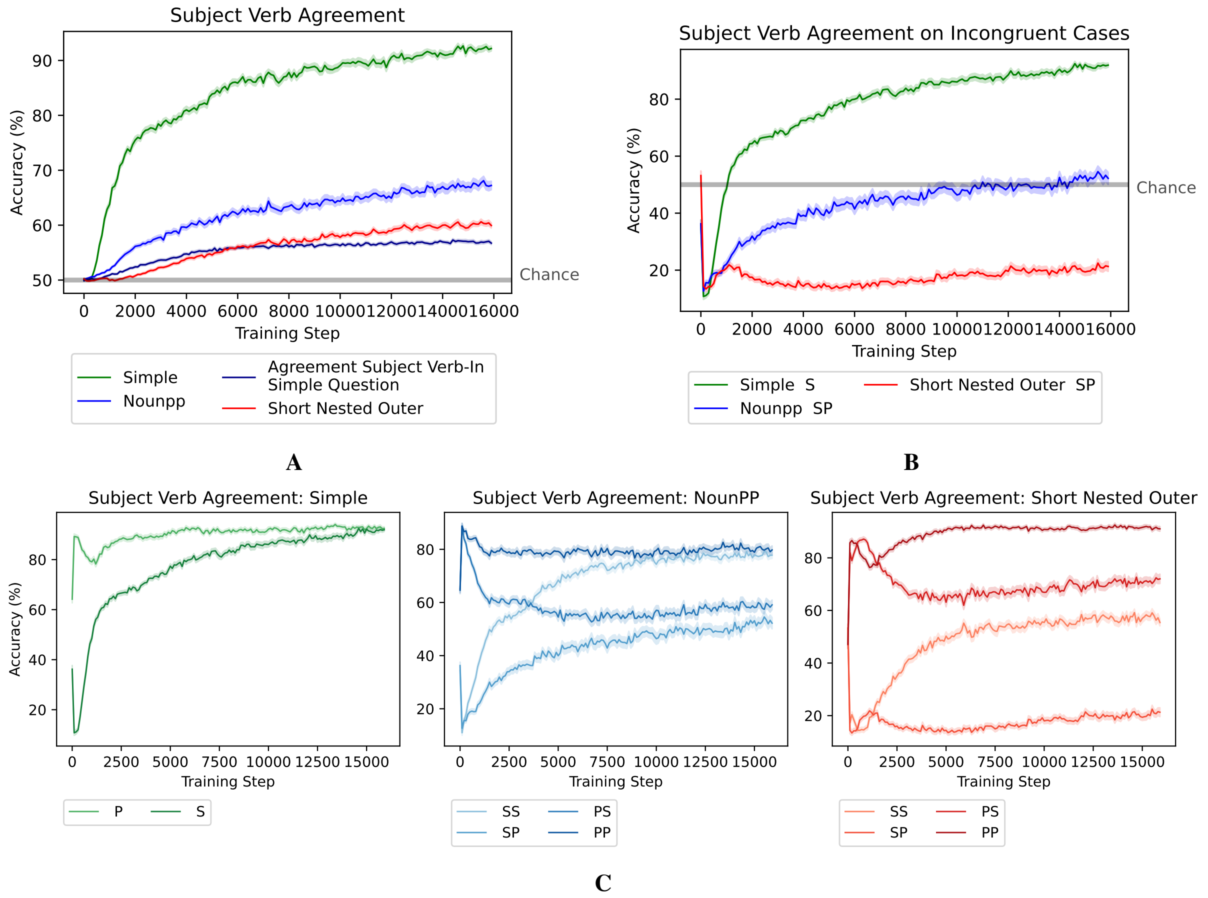


Figure 5: Subject verb agreement. Shading is standard error of the mean across seeds. S: Singular. P: Plural. **A**. Parallel learning is observed in the three stages defined by Friedmann et al. (2021), when results are averaged across congruent and incongruent cases. **B**. Subject verb agreement on incongruent cases which indicate whether the model understands syntax. The networks do not learn some syntax structures as the incongruent case of Short Nested Outer does not reach above chance level. **C**. Development trajectories of the bias towards plural number in English.

4.2 A work in progress

It is important to stress that significant work remains to be done before drawing any definitive conclusions about the similarities between language acquisition in humans and algorithms.

First, we only consider a single architecture (GPT-2, Radford et al. (2019)) with a unique textual corpus (WikiText103). Testing whether our results hold true independently of the model and training corpus remains an important milestone for future research.

Second, linguistic abilities are not tested with the same protocols in children and in the models: the models are explicitly tested on next word prediction, with a two-alternative forced-choice metric, whereas children were implicitly evaluated on their ability to spontaneously use specific syntactic structures during natural speech.

Third, there were only three linguistic features that were directly comparable between the model probes and the data in children, and all were syntactic. This leaves a significant margin of progress to modulate our conclusion, and investigate whether the lexicon, narrative structures, pragmatics and world-knowledge are acquired in the same order in humans and algorithms.

Fourth, the order in which some linguistic skills were learned by GPT-2 does not trivially fit with linguistic theory. For example, the probe “Simple”, which examines subject-verb agreement in a simple sentence, was one of the last probes to be learned by GPT-2 (it is part of group three in Figure 2). By contrast, "Wh Questions Subject Gap Long Distance" was among the first probes to be to be learned, even though it would be expected to be much harder than “Simple”. This unexpected result may be due to the way we approximate "Acquisition Time", namely, the moment when the probes reaches 90% of the final accuracy. Consequently, probes with very low final accuracy could end up with a shorter Acquisition Time, because noise may lead to crossing the 90% threshold relatively quickly.

Finally, we show that our models appear to use heuristics rather than a deep understanding of syntax for the most difficult linguistic probes (incongruent numbers between verbs and their attractors) and were biased towards plural English verbs. While our models learn only 66% of tasks to above chance level, a larger GPT-2 pretrained on considerably more texts successfully perform on 97% of the tasks, and has an accuracy above 75% on the incongruent examples, meaning this bias and reliance on heuristics could potentially be solved by training on a larger dataset.

In sum, additional work remains necessary to identify the exact elements of convergence and divergence between the acquisition of language in models and in children.

4.3 Fueling the debate between nativism versus empiricism

The present study fuels a long-lasting debate on the acquisition of language. While “empiricists“ argue that language can be acquired with a statistical approach (Clark, 2002; Kolodny et al., 2015; Chater and Christiansen, 2018; McCauley and Christiansen, 2019), “nativists“ maintain that this ability depends on a core and innate operation, specific to humans (Chomsky, 1959, 1971).

The present study shows how modern language models may contribute to resolving this debate, by systematically studying which components of a model (e.g. architecture) or properties of the training data (e.g., frequency of sentence structures) contribute to shape the trajectory of language acquisition. Claims about an innate Universal Grammar could be understood as an inductive bias of a language model, implemented in its architecture and dynamics, which tightly constrains learning trajectories across models. If this bias is hierarchical (rather than linear) then this could lead to learning trajectories that follow the structure of the syntactic tree, consistently with the hypothesis of three linguistic stages presented by Friedmann et al. (2021) in humans and what we find in this study in language models. Indeed, neural language models have been previously shown to have a weak inductive bias towards hierarchical processing (McCoy et al., 2020; Kharitonov and Chaabouni, 2020), which can partially explain our results.

This result echos the recent observation that syntactic trees spontaneously emerge in the middle layers of neural language models (Hewitt and Manning, 2019). Together, these elements thus suggest that modern neural networks provide fruitful models of language acquisition and could reconcile or settle the confronting theories of language acquisition (Warstadt and Bowman, 2022).

4.4 Conclusion

Overall, the similarities identified between children and GPT-2 suggest that there may be a small set of means by which to efficiently acquire language. This result is anything but trivial: humans and deep neural networks have extraordinarily different architectures, training, and language exposure. If generalized, this systematic learning trajectory would support the existence of an intrinsic hierarchy of linguistic structures that both machines and humans must climb, be that through inductive biases or properties of the training data, to master the faculty of language. And while these hypotheses remain open, the path to resolve them has never been clearer.

Acknowledgements

We would like to thank Dieuwke Hupkes, Naama Friedmann, Marco Baroni and the attendees of the EviL meetings for their comments and suggestions.

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 945304, for L.E for her work at PSL. This work was funded in part by FrontCog grant ANR-17-EURE-0017 for the work of L.E. and J.R.K. for their work at PSL.

References

Elika Bergelson and Daniel Swingley. 2012. At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, 109:3253–3258.

Samuel R. Bowman and George Dahl. 2021. What will it take to fix benchmarking in natural language understanding? *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. 2023. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*, pages 1–12.

Charlotte Caucheteux and Jean Rémi King. 2022. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1).

Tyler A. Chang and Benjamin K. Bergen. 2022. Word acquisition in neural language models. *Transactions of the Association for Computational Linguistics*, 10:1–16.

Nick Chater and Morten H Christiansen. 2018. Language acquisition as skill learning. *Current opinion in behavioral sciences*, 21:205–208.

Noam Chomsky. 1959. Review of verbal behavior. 35(1):26–58. Publisher: Linguistic Society of America.

Noam Chomsky. 1971. *Problems of Knowledge and Freedom*. New York,: W.W. Norton.

Alexander Clark. 2002. Unsupervised language acquisition: Theory and practice. *arXiv preprint cs/0212024*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova Google, and AI Language. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *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.

Emmanuel Dupoux. 2018. Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, 173:43–59.

Naama Friedmann, Adriana Belletti, and Luigi Rizzi. 2021. Growing trees: The acquisition of the left periphery. *Glossa: a journal of general linguistics*, 39(1).

Naama Friedmann and Julia Reznick. 2021. Stages rather than ages in the acquisition of movement structures: Data from sentence repetition and 27696 spontaneous clauses. *Glossa: a journal of general linguistics*, 39(1).

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.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138.

Philip Huebner. 2022. Unmasked. `https://github.com/phueb/UnMasked`. (Accessed 2023/05/24).

Philip A Huebner, Elior Sulem, Cynthia Fisher, and Dan Roth. 2021. BabyBERTa : Learning More Grammar With Small-Scale Child-Directed Language. *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646.

Shailee Jain and Alexander Huth. 2018. Incorporating context into language encoding models for fmri. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657.

Jaap Jumelet, Willem Zuidema, and Dieuwke Hupkes. 2019. Analysing neural language models: Contextual decomposition reveals default reasoning in number and gender assignment. *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 1–11.

Fred Karlsson. 2007. Constraints on multiple center-embedding of clauses. *Journal of Linguistics*, 43(2):365–392.

Eugene Kharitonov and Rahma Chaabouni. 2020. What they do when in doubt: a study of inductive biases in seq2seq learners. *arXiv:2006.14953*.

Oren Kolodny, Arnon Lotem, and Shimon Edelman. 2015. Learning a generative probabilistic grammar of experience: A process-level model of language acquisition. *Cognitive Science*, 39(2):227–267.

Patricia K. Kuhl. 2004. Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience*, 5:831–843.

Patricia K Kuhl, Erica Stevens, Akiko Hayashi, Toshisada Deguchi, Shigeru Kiritani, and Paul Iverson. 2006. Fast-track report infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science*, 9:13–21.

Patricia K. Kuhl, Karen A. Williams, Francisco Lacerda, Kenneth N. Stevens, and Bjorn Lindblom. 1992. Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, 255:606–608.

Yair Lakretz, Théo Desbordes, Dieuwke Hupkes, and Stanislas Dehaene. 2021a. Causal transformers perform below chance on recursive nested constructions, unlike humans. *arXiv:2110.07240*.

Yair Lakretz, Dieuwke Hupkes, Alessandra Vergallito, Marco Marelli, Marco Baroni, and Stanislas Dehaene. 2021b. Mechanisms for handling nested dependencies in neural-network language models and humans. *Cognition*, 213:104699. Special Issue in Honour of Jacques Mehler, Cognition's founding editor.

Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntax units in LSTM language models. *Association for Computational Linguistics*, pages 11–20.

Marvin Lavechin, Maureen De Seyssel, Hadrien Titeux, Hervé Bredin, Guillaume Wisniewski, Alejandrina Cristia, and Emmanuel Dupoux. 2022. Statistical learning bootstraps early language acquisition. *PsyArXiv*.

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.

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences of the United States of America*, 117:30046–30054.

Reiko Mazuka, Yvonne Cao, Emmanuel Dupoux, and Anne Christophe. 2011. The development of a phonological illusion: A cross-linguistic study with japanese and french infants. *Developmental Science*, 14:693–699.

Stewart M McCauley and Morten H Christiansen. 2019. Language learning as language use: A cross-linguistic model of child language development. *Psychological review*, 126(1):1.

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.

Jacques Mehler, Peter Jusczyk, Ghislaine Lambertz, Nilofar Halsted, Josiane Bertoncini, and Claudine Amiel-Tison. 1988. A precursor of language acquisition in young infants. *Cognition*, 29:143–178.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv:1609.07843*.

Thierry Nazzi, Peter W. Jusczyk, and Elizabeth K. Johnson. 2000. Language discrimination by english-learning 5-month-olds: Effects of rhythm and familiarity. *Journal of Memory and Language*, 43:1–19.

Alexandre Pasquiou, Yair Lakretz, John Hale, Bertrand Thirion, and Christophe Pallier. 2022. Neural language models are not born equal to fit brain data, but training helps. In *ICML 2022-39th International Conference on Machine Learning*, page 18.

Alexandre Pasquiou, Yair Lakretz, Bertrand Thirion, and Christophe Pallier. 2023. Information-restricted neural language models reveal different brain regions' sensitivity to semantics, syntax and context. *arXiv:2302.14389*.

Yada Pruksachatkun, Phil Yeres, Haokun Liu, Jason Phang, Phu Mon Htut, Alex Wang, Ian Tenney, and Samuel R. Bowman. 2020. jiant: A software toolkit for research on general-purpose text understanding models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 109–117, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *Semantic Scholar*. (Accessed 2023-05-04).

Rushen Shi, Janet F Werker, and James L Morgan. 1999. Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition*, 72:B11–B21.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv 2206.04615*.

Ruth Tincoff and Peter W. Jusczyk. 1999. Some beginnings of word comprehension in 6-month-olds. *Psychological Science*, 10:172–175.

Mariya Toneva and Leila Wehbe. 2019. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Warstadt and Samuel R Bowman. 2022. What artificial neural networks can tell us about human language acquisition. *arXiv:2208.07998*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng Fu Wang, and Samuel R. Bowman. 2020. Erratum: "blimp: The benchmark of linguistic minimal pairs for english". *Transactions of the Association for Computational Linguistics*, 8:867–868.

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

Lucas Weber, Jaap Jumelet, Elia Bruni, and Dieuwke Hupkes. 2021. Language modelling as a multi-task problem. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2049–2060, Online. Association for Computational Linguistics.

Janet F. Werker. 2018. Perceptual beginnings to language acquisition. *Applied Psycholinguistics*, 39:703–728.

Janet F. Werker and Richard C. Tees. 1984. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, 7:49–63.

Tania S. Zamuner. 2006. Sensitivity to word-final phonotactics in 9- to 16-month-old infants. *Infancy*, 10:77–95.

Appendix

A Tests in Children

Detailed description of tests available in children, in the three linguistic stages defined by (Friedmann et al., 2021):

- Stage 1: Subject-Verb Simple, Subject-Verb Unaccusative, Verb-Subject Unaccusative
- Stage 2: Root WH-Argument, WH-Adjunct Excluding Why, Preposed Adverb, Root y/n
- Stage 3: Why, Relative Clause, Topicalisation, Embedding

The probes chosen for comparison (stated in Table 1), were the only probes that matched well with one of the test available in children. In addition Nounpp was examined in the models, as it fits into the linguistic stage 2, and, as it is part of the BIG-Bench probes, could be separated into congruent and incongruent sentences.

B Model Training

To evaluate linguistic abilities of a high-performance language model, we first use the HuggingFace pretrained GPT-2 large which has 774M parameters and is trained on 40GB of data. This model has one-shot perplexity of 22 on WikiText103 (Radford et al., 2019).

Then, to evaluate how linguistic abilities vary with language acquisition, we separately trained 48 models (each with a distinct random seed which set the model's initial parameters and the seed of the dataloader) using the 12-layer GPT-2 architecture (Radford et al., 2019) provided by HuggingFace[4] on WikiText103 (Merity et al., 2016) with a learning rate of 10^{-5} and a batch size of 16 distributed over 8 GPUs, making a total batch size of 64 and a context size of 128. Training was stopped when then validation loss plateaued, reaching final perplexity of 28 after 10 epochs. This is lower perplexity than the one-shot performance of the HuggingFace pretrained 12-layer GPT-2 which was 37.5, which is logical as our model was trained specifically on this dataset.

In all cases we used the pretrained tokenizer which has vocabulary size of 50,257. All other parameters were the default training arguments for the transformer provided by HuggingFace. The HuggingFace architectures are publicly available under an MIT license, and WikiText103 is available under Creative Commons Attribution-ShareAlike License.

C Linguistic Probe Benchmarks

We use three different zero-shot benchmarks. The first benchmark, 'BLiMP' (The Benchmark of Linguistic Minimal Pairs for English) (Warstadt et al., 2020) contains 67 different probes, each in the form of 1,000 pairs of grammatical and ungrammatical sentences designed to isolate a specific linguistic phenomenon. Adult human performance on BLiMP is 96.4% (Warstadt et al., 2020). The second benchmark, 'Zorro' [5], was developed with a vocabulary frequent in child-directed corpora. Zorro contains 13 probes, each consisting of 2,000 pairs of sentences. Finally, the third benchmark is the Subject-Verb Agreement Task of BIG-

[4] https://huggingface.co/gpt2

[5] https://github.com/phueb/Zorro

Bench (Srivastava et al., 2022; Lakretz et al., 2019, 2021b; Linzen et al., 2016; Gulordava et al., 2018). We focus on the syntactic probes, namely: "Simple English" which contains 600 pairs, "NounPP" which contains 2,400 pairs, and "Short Nested Inner", "Short Nested Outer", "Long Nested Inner" and "Long Nested Outer" which each contain 4,096 pairs of grammatical and ungrammatical sentences.

Accuracy on a linguistic probe is evaluated with the Jiant (Pruksachatkun et al., 2020) and UnMasked method (Huebner, 2022). In practice, sentences are input to the model in batches of 300, with padding on the left to make all sentences the length of the longest sentence in the batch. The logit values of punctuation are discarded when estimating the probability of a sentence.

Zorro, Jiant and UnMasked are publicly available under the MIT License, BLiMP under a CC BY License, and BIG-Bench under the Apache License 2.0.

ACL 2023 Responsible NLP Checklist

A For every submission:

□ A1. Did you describe the limitations of your work?
Left blank.

□ A2. Did you discuss any potential risks of your work?
Left blank.

□ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

□ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B □ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
Left blank.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Left blank.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Left blank.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Left blank.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Left blank.

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C □ Did you run computational experiments?

Left blank.

□ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

□ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Left blank.

□ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Left blank.

□ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Left blank.

D □ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Left blank.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Left blank.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Left blank.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Left blank.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Left blank.