The Syntactic Productivity of Large Language Models

Anonymous ACL submission

Abstract

Do Large Language Models (LLMs) produce output that exhibits syntactic productivity similar to human language? Although recent work has focused on quantifying the lexical, ngram or templatic novelty of LLMs with respect to their training data, we posit the problem is formally equivalent to a major issue in child language research where conclusions must be drawn about the underlying grammar solely on the basis of a child's production data. We apply a mathematically rigorous and independently validated measure of Syntactic Productivity–the combinatorial diversity of Determiner-Noun (D×N) pairs used to measure young children's developing grammars–to four OpenAI LLMs whose training data is inaccessible. We find children, their caretakers and professional writers show the statistical hallmark of Syntactic Productivity but LLM-generated texts do not (Figure 1).

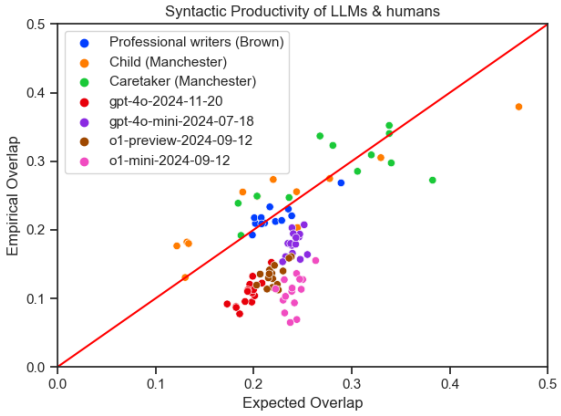

Figure 1: Syntactic productivity measure (overlap; Section 2) of human language corpora (children, their caretakers, and professional writers) and 4 LLM-generated corpora from the OpenAI API. Each point indicates a corpus. Human corpora show measures comparable to the expectations under a fully productive grammar (Section 3) but LLM corpora show significantly lower measures of productivity (Section 4). The red reference line indicates a perfect match between the two.

1 Introduction

The success of LLMs has spurred significant research to characterize their capacity to represent linguistic structures in comparison to human language users.

A prominent approach has focused on the development and use of benchmarks to probe for specific linguistic properties in LLMs. These range from extracting structures from internal representations (e.g., Hewitt and Manning, 2019; Tenney et al., 2019; McCoy et al., 2020; Tucker et al., 2021; Papadimitriou et al., 2021), to building tasks inspired by psycholinguistic processing studies (e.g., Chowdhury and Zamparelli, 2018; Wilcox et al., 2018; Hu et al., 2020), to classic acceptability rating tasks that theoretical linguists use to infer grammatical knowledge (e.g., Linzen et al., 2016; Warstadt et al., 2020; Huebner et al., 2021; Sinclair et al., 2022). The evaluation paradigms typically rely on benchmarking LLMs against fixed datasets, which either require the LLM to receive task-specific fine-tuning, or require researchers to carefully engineer prompts that adapt tasks into formats that LLMs can interpret and perform well on (Scao and Rush, 2021). However, this reliance on prompt engineering introduces evaluation challenges, as model performance can be significantly impacted by the choice of problem framing (Mishra et al., 2022), choice and order of in-context examples (Zhao et al., 2021), the token and sentence distribution in the prompt (Min et al., 2022), and whether the model is of large enough size to learn new priors in-context (Wei et al., 2023), among other factors.

While the benchmarking approach has provided valuable insights into LLMs' linguistic capacity, they are by design limited to the specific structural properties identified by the researcher and

may provide an insufficiently representative coverage of linguistic phenomena (McCoy et al., 2019; Vázquez Martínez, 2021; Wang et al., 2022; Guest and Martin, 2023; Vázquez Martínez et al., 2023). The rise of generative AI models necessitates the development of evaluation methods for open-ended LLM output (Chang et al., 2024).

Implicit in the open-ended evaluation paradigm is the assumption that, if a text exhibits certain linguistic properties (e.g. subj-verb agreement), then its source must have learnt that property. Yet, if an LLM simply copies its training data, its output does not provide clear evidence for linguistic abstraction (McCoy et al., 2023). Recent work has thus attempted to quantify the rate of ngram novelty (McCoy et al., 2023; Merrill et al., 2024) and structural diversity (Shaib et al., 2024) of text generated by LLMs relative to their pretraining data. But, how can one evaluate closed-source LLMs or even open-weight LLMs whose training data is undisclosed?

In this paper, we introduce a novel approach to LLM evaluation with specific focus on syntactic productivity that is agnostic to the model's training data. Our approach draws inspiration from the study of child language, where researchers frequently need to assess a learner's underlying grammar based solely on a corpus of their language production. The Syntactic Productivity evaluation thus compares the open-ended output of LLMs to itself by calculating an expected productivity threshold based on the statistical profile of natural human language. We demonstrate how this can be done without access to any training data by applying the Syntactic Productivity evaluation to four OpenAI LLMs and compare their performance to how children, their caretakers and professional writers perform on the same test.

Our contributions are as follows:

- We adapt the method of Syntactic Productivity drawn from the study of child language (Section 3) for application to AI-generated text.[1]

- We validate our method by applying it to a subset of the child and caretaker speech in the CHILDES database (MacWhinney, 2000) as well as the Brown Corpus (Kučera and Francis, 1967). Our implementation replicates prior findings in the literature (Section 4): Young children are indeed as syntactically productive (Figure 2a) as their caretakers (Figure 2b) and as professional writers (Figure 3).

- We generate a large body of continuous narrative text using the models available in the OpenAI API as of `14.02.2025`. We make our datasets publicly available[2] so that further analyses of the LLMs' outputs do not need to incur similar token-generation expenses.

- We find that narrative text generated by LLMs fails to show the statistical properties of productivity (Figure 4), whereas the humans' predicted and empirical overlap scores are statistically indistinguishable from each other across all contexts (Figure 1; Section 5).

2 Measuring productivity in child language

The defining feature of language is its infinite productivity, as new words and sentences can always be generated. A revealing method for uncovering productivity, as shown in the celebrated Wug test (Berko, 1958), is to provide the language learner with novel input and assess whether appropriate output forms can be generated. However, such experimental approaches have certain task-related complications that limit their applications. For example, while children learn the English past tense suffix (-ed) before age 3 as shown by occasional over-regularization errors (e.g., *goed*; Kuczaj 1977), not even first graders consistently produced -ed on the Wug test (Berko, 1958) as children often struggle learning and using a novel word in an artificially induced setting. Comprehension studies also carry extra cognitive demands. Even 4-year-olds fail to completely accurately distinguish the temporal reference of "was" and "is" in an experimental setting (Valian, 2006).

Hence, the investigation of early child language has often focused on children's naturalistic production, which is least subject to performance constraints while also providing the most accessible type of acquisition data. In particular, the combination of determiners (D) and nouns (N), or `D×N` for short, has been a major focus in child language research (Pine and Martindale, 1996; Valian et al.,

[1] We will update this footnote with a link to the GitHub repository in the deanonymized version.

[2] We will update this footnote with a link to the dataset in the deanonymized version.

2009; Pine et al., 2013). This is because determiners, especially singular determiners *the* and *a*,[3] are highly frequent and thus well represented in child language. Despite its simplicity, D×N fully exhibits the hallmark of syntactic productivity: Any singular noun used with *the* can also be used with *a*. A simple metric, dubbed *overlap* (Pine and Lieven, 1997), has been widely used to quantify productivity: the proportion of singular nouns used with both *the* and *a* out of those used with either (Equation 1). The overlap value is bounded between 0 and 1: A higher value would be stronger evidence for productivity, but as we will see shortly, this intuition needs to be qualified.

$$\text{empirical} = \frac{1}{|N|} \sum_{n \in N} \mathbb{1}\left[\forall_{d \in D}\, C_{\text{d}\times n} > 0\right] \quad (1)$$

Many previous studies of D×N focus on the comparison of overlap values in children and their caretaker's language. However, any corpus of caretaker language is only a small sample of a learner's input data. Moreover, adults talk more and have larger vocabularies than children, so it has been difficult to develop "fair" comparisons across samples. A statistical test for syntactic productivity (Yang, 2013; Goldin-Meadow and Yang, 2017) sidesteps these issues. This test calculates the expected value of D×N overlap in a corpus under the assumption that D×N is fully productive i.e., statistically independent.

3 A Statistical Test for Productivity

The test builds on two key statistical properties of language, one universal and the other specific to D×N in English. First, the test assumes that the frequencies of words, especially open class words such as nouns, follow Zipf's or inverse power law distribution (Zipf, 1949; Baroni, 2009). As such, if a corpus contains $|N|$ unique nouns in D×N combinations, the noun with rank r has the expected probability (p_r):

$$p_r = \frac{1}{r^a H_{n,a}}$$

$$\text{where } H_{n,a} = \sum_{i=1}^{n} \frac{1}{i^a}$$

$H_{n,a}$ is the generalized harmonic number with a as the exponent of inverse power law. In most cases a is approximately 1 following Zipf's original formulation but deviation from 1 can be accommodated in the calculation.

Second, it is observed that in D×N combinations, nouns tend to have a "favorite" determiner that combines far more frequently with it than the other. For example, *bathroom* greatly favors *the* over *a* but for *bath*, the reverse is true. This imbalance, referred to as *bias* (b), is defined as follows:

$$b = \frac{\sum_{n \in N} \max(C_{\text{the}\times n},\ C_{\text{a}\times n})}{\sum_{n \in N} (C_{\text{the}\times n} + C_{\text{a}\times n})} \quad (2)$$

where $C_{\text{the/a}\times n}$ is the frequency of *the/a* combined with noun n. The bias value is not part of the grammar *per se* nor does it require learning: It is unlikely that children track the frequency of bodily functions ("the bathroom") or hygienic practices ("a bath"). Rather, the bias value is the vagaries of life reflected in language use. As *bath* and *bathroom* illustrate, not all nouns have the same favorite determiners. Situational factors may also skew the bias: a pediatrician will have more balanced use for *the* and *a* for the noun *baby* than the parent of a newborn. Nevertheless, as we show in Section 4, the bias value in aggregate is remarkably stable across samples of English at $b = 0.82$.

Taken together, these two statistical properties greatly enhance the applicability of the test. For a corpus, one only needs S, the total number of D×N combinations, and N, the number of unique singular nouns. Once the exponent of Zipf's Law is obtained from frequencies of the N nouns, one can compute the expected overlap value of each noun in the set (Equation 3).

$$\begin{aligned} E_r &= 1 - (1 - p_r)^S \\ &- [(b * p_r + 1 - p_r)^S - (1 - p_r)^S] \\ &- [((1 - b) * p_r + 1 - p_r)^S - (1 - p_r)^S] \end{aligned} \quad (3)$$

Taking the mean over all nouns in the sample (Equation 4; the full derivation is given in (Yang, 2013, Supporting Information) yields the average expected overlap, which is compared to the empirical value calculated in Equation 1.

$$E[S] = \frac{1}{N} \sum_{r=1}^{N} E_r \quad (4)$$

If there are no statistically significant differences between the `empirical` and `expected` overlap values, one can conclude that the D×N combinations

[3] The phonological variant *an* is treated as *a* as it is an independent developmental process.

are in fact consistent with a fully productive grammar.

The syntactic productivity test is not limited to determiners and nouns but can be applied to any two combinatorial categories, as long as the closed class category has only two members and the open class category frequency can be approximated by Zipf's Law. Moreover, it can be applied to detect both the presence and absence of productivity. For example, Goldin-Meadow and Yang (2017) adapted the test to the combinatorial structure of homesign, the gestural system created by deaf children in the absence of sign language input. The test finds that homesign combinations are fully productive, providing independent evidence for traditional behavioral analysis. On the other hand, the test has been applied to the ASL sign combinations produced by Nim Chimpsky (Yang, 2013). Results show that Nim's sign combinations show considerably less diversity than would be expected under a fully productive system, again supporting conclusions based on frame-by-frame sign analyses (Terrace et al., 1979).

We focus on the measure of D×N overlap specifically because it is so well represented in children's speech, and therefore studied in child language acquisition. D×N combinations enable us to draw a robust comparison between young children and LLMs, whereas pairs of closed and open class categories that are acquired later would immediately preclude the children's speech data from comparison.

4 Syntactic Productivity in Humans

The first set of human language analyses is based on the Manchester corpus (Theakston et al., 2001). There are 12 dyads of typically developing children and their caretakers, and the transcripts are based on regular recording sessions between age 2 and 3. The Manchester Corpus is the largest longitudinal record of English language development for this age group and has been frequently used in child language acquisition research.

Following previous work (Pine et al., 2013), a D×N combination is extracted if D is *the* or *a* and N is a singular noun that immediately follows D or with one non-noun intervening word. Data extraction used the spaCy dependency parser (Honnibal and Johnson, 2015) which also provides POS tagging of the transcripts. The statistical conclusions of our study remain unchanged if we use the POS annotation provided in CHILDES.

We found that in the Manchester Corpus, the nouns in both child and caretaker language show excellent fit for the original Zipf's Law with an average exponent of $a = 1.03$. Furthermore, as noted earlier, D×N combinations in English are heavily biased toward one of the two determiners. The bias value estimated from the Corpus of Contemporary American English (COCA; Davies, 2009) based on Equation 2 is $b = 0.82$. Remarkably, the bias value across the 12 dyads of children and caretakers is almost identical (mean = 0.814, sd = 0.03), and there is no significant difference between the bias value in child language samples and caretaker language samples (paired t-test $p = 0.612$). Thus in all studies we have used the universal bias value $b = 0.82$ for expected overlap calculation. These values of a and b were used to calculate the expected overlap value. The results are shown in Figure 2, with Figure 1 putting them in context with all other syntactic productivity samples tested. There is no statistically significant difference between expected and empirical values in the Manchester Corpus for children (paired t-test: $p = 0.334$) nor for caretakers (paired t-test: $p = 0.733$).

The second set of human language analyses is based on the Brown Corpus (Kučera and Francis, 1967), a collection of professional print materials across a wide range of genres. To make suitable comparisons with the Manchester Corpus, we grouped successive files in the Brown Corpus into 12 samples. The D×N combinations were extracted with spaCy following the method used for the Manchester Corpus. The nouns in each sample do not follow the canonical Zipf's Law with exponent of 1. Rather, the average exponent of the Brown Corpus samples is 0.771. We believe that this is due to the nature of the Brown Corpus, where each file is a relatively short document about a particular topic. Collectively, the most frequent nouns in each sample are much closer to each other in frequency. By contrast, the speakers in the dialog samples in the Manchester Corpus had more focused and extensive conversations about fewer topic nouns. For the Brown corpus analysis, we used the exponent $a = 0.771$ along with the universal bias value $b = 0.82$ to calculate the expected overlap value in comparison to empirical values. Figure 1 summarizes the results with additional details in Figure 3. Once again, the difference between the `expected` and the `empirical` overlap values is not statistically significant (paired t-test

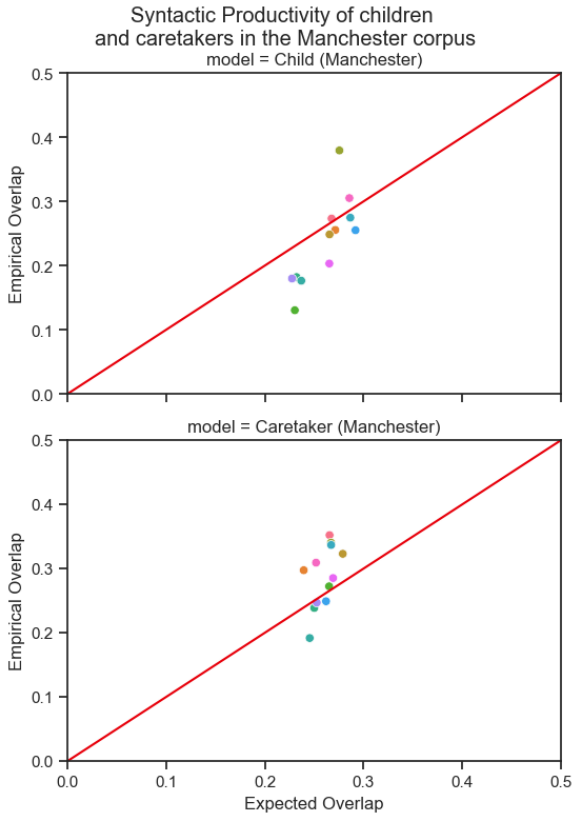

Figure 2: Scatter plot of expected and empirical productivity measure (D×N overlap) for the 12 children and their corresponding caretakers from the Manchester Corpus (Theakston et al., 2001). No statistically significant difference is found (paired t-test $p = 0.334$ children and $p = 0.771$ for caretakers).

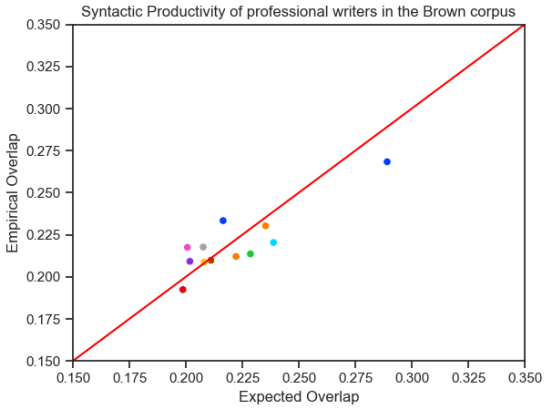

Figure 3: Scatter plot of expected and empirical productivity measured (D×N overlap) for 12 sections of the Brown corpus (Kučera and Francis, 1967). No statistically significant difference is found (paired t-test $p = 0.562$).

$p = 0.586$).

Note the Syntactic Productivity test is not limited to D×N but is applicable to any rule that combines a two-member closed class category with an open class category. To further establish the robustness of the test, we extracted the Manchester Corpus verb lemmas inflected with either -ed or -ing from the Manchester corpus: the overlap measures the proportion inflected with both. Note that -ed and -ing are not fully interchangeable due to irregular verbs (e.g. **goed*), . Thus, the empirical overlap for verb lemmas over -ed and -ing must be *lower* than the expected value, the latter of which is computed on the assumption of full interchangeability. Indeed, across the 24 dyad samples, the empirical values are significantly lower than the expected values (paired t-test $p < 0.001$). However, once the irregular verbs are removed, the empirical overlap value of verb lemmas for -ed and -ing are not significantly different from the expected value across the 24 dyad samples (paired t-test $p = 0.852$) because the two suffixes are indeed fully interchangeable for regular verbs.

Taken together, the analyses of human language illustrate the robustness of the test for detecting both true positives of productivity such as adult usage in Manchester and Brown as well as true negatives, such as the counterfactual application to verbal inflection. Next, we examine whether LLMs constitute a true positive or a true negative of syntactic productivity.

5 LLMs Fail Productivity Test

To evaluate the syntactic productivity of LLMs, we need a sample of text from each model whose raw count of D×N pairs (S) and unique nouns (N) is comparable to that of the human data we use as a baseline. While we would most easily obtain AI-generated text from previously generated detection tasks, these generally consist of short documents between 200 and 500 tokens (Kim et al., 2024a). We therefore generate multiple long-form texts of at least 15K tokens with each of the four most advanced OpenAI models available to us as of `02.15.2025`, listed in Table 1.

For each model, we compose a set of 15 `NARRATIVE_TOPICS` spanning different genres (e.g., a science fiction story, an academic job talk, an economics survey, among others), each with three more follow up topics that keep the discourse coherent. To prompt the models, we constructed a list of `NARRATIVE_TEMPLATES` that can be filled in with each of the 15 topics and follow ups. We additionally included a `SYSTEM_PROMPT` that instructs

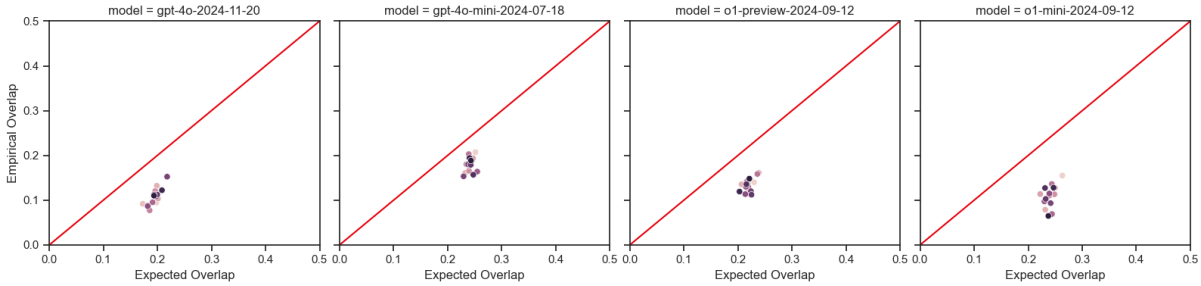

Figure 4: Scatter plot of expected and empirical productivity measures (D×N overlap) for 15 samples of narrative texts generated by OpenAI models. The empirical values of overlap are considerably lower than the expected values under full productivity (paired t-test, $p < 0.001$).

model	N	S
gpt-4o-2024-11-20	353	687
gpt-4o-mini-2024-07-18	608	1982
o1-mini-2024-09-12	407	1200
o1-preview-2024-09-12	563	1408

Table 1: The mean size of the AI-generated narrative text measured in D×N combinations.

the model to write as coherently and in as much detail as possible in order to pass the Turing test. This yielded 15 long-form narratives for each of the four OpenAI models. We summarize the relevant statistics of the generated narratives, mainly S and N, in Table 1.

As in the human analyses, D×N combinations are extracted from LLM texts using spaCy. Empirical analysis shows that on average, the inverse power law exponent of the nouns across 60 texts is $a = 0.745$ – analogous to that in the Brown corpus — which is used in the expected overlap calculation. We used the human universal bias value $b = 0.82$ to calculate the expected value of D×N overlap for the LLMs in comparison to the empirical values. The results are summarized in Figure 1 with additional details in Figure 4. The expected values are significantly higher than the empirical values (paired t-test $p < 0.001$ for all four models). The LLM text showed a higher average bias value (0.92) than human texts but $b = 0.82$ still resulted in expected values significantly higher than the empirical values ($p < 0.05$ for all four models). We thus conclude that unlike human language learners and users, LLMs do not generate D×N combinations in a fully productive way.

6 Related and Future Work

The present work falls in line with current efforts to uncover human knowledge of language encoded in LLMs using their textual output as a proxy for their underlying grammars. This includes efforts to quantify novelty in terms of lexical sequences—how often a model produces n-grams not seen in its training data (McCoy et al., 2023; Merrill et al., 2024)—as well as the syntactic templates in generated text (Shaib et al., 2024). These approaches typically require access to the model's training corpus or high-quality estimates thereof, limiting their applicability to closed-source systems or models trained on undisclosed data. The Syntactic Productivity method evaluates the generalization of closed-class elements across open-class categories without requiring any knowledge of the model's training data. Training data notwithstanding, the method maintains sensitivity to memorized retrieval rather than productive generation. In other words, LLMs that rely more on retrieval from the training data will exhibit lower Syntactic Productivity by our measure.

A parallel line of research seeks to characterize the nature of LLM-generated text as it compares to naturalistic human text (e.g., Muñoz-Ortiz et al., 2024; Zanotto and Aroyehun, 2024). Although notable progress has been made toward the identification of an "AI-signature", such as lexical overrepresentation (Juzek and Ward, 2024), or reduced diversity of discourse motifs (Kim et al., 2024b), current methods cannot reliably detect it. Automated AI-text detection methods operate at unadvisable False Positive rates, perform poorly on out-of-sample data, and are susceptible to adversarial attacks (Dugan et al., 2024, 2025). On the other hand, a high proportion of human participants are also at chance when it comes to identifying

the source (human or not) of a given text (Jannai et al., 2023; Jones and Bergen, 2024; Clark et al., 2021). Encouragingly, recent studies also suggest that the level of exposure an individual has had to AI-generated text significantly improves their ability to identify it (Dugan et al., 2023; Russell et al., 2025).

LLMs' open-ended outputs are thus distinct in ways that are characterizable with linguistic methods. It follows that understanding the differences between LLMs' and humans' knowledge of language in a principled manner has scientific implications for our understanding of language as a computational construct, as well as practical implications for its identification in real-world use.

7 Discussion

Our results point to a significant difference in syntactic productivity between humans and LLMs as it pertains to D×N combinations, but it is difficult to ascertain the nature of such discrepancies. If LLMs are relying on memorization and retrieval of fixed D×N combinations from the training data, it is a mathematical fact that the overlap in D×N combinations will always be less than or equal to the overlap in the training data. That is because there will always be a nonzero probability that, when the D×N combinations are fixed, a given N is stochastically sampled with one determiner but not the other (Yang, 2013). This behavior would be reflected in LLMs' over reliance on the memorization of lexically specific combinations (Juzek and Ward, 2025).

Additional phenomena of syntactic productivity can be investigated as the test can be applied to any combinatorial process that meets the criterion of statistical independence and full interchangeability.

8 Limitations

While the productivity test can be applied to many combinatorial processes, it has two inherent limitations. First, the closed class category can only have two members (e.g., *the* and *a* in D). Adding more members (e.g., *this* and *that*) makes the mathematical formulation intractable. Second, the test assumes that the categories combine in fully interchangeable and thus statistically independent ways. While processes such as those studied in the present paper can be characterized as such, this is not the case for all rules in language, at least not in a way than lends readily to the test. For example, not all transitive verbs can passivize ("John resembles Bill" cannot be passivized as "*Bill was resembled by John"), not all dative verbs can appear in both the double object construction (*tell* but not *say*) and the *to*-dative construction (*tell* but not *ask*).

Despite observing a categorical distinction between human and AI-generated text, it would be impractical to apply the Syntactic Productivity test to the AI-text detection task. Firstly, the test of Syntactic Productivity assumes the text under evaluation comes from some connected discourse. The Zipfian distribution of nouns hinges upon discourse being centered around a particular topic. For example, if we scrambled all the sentences across the Manchester Corpus and then reran the Syntactic Productivity test, participants' `empirical` scores would exceed the `predicted` D×N overlap.

More practically, the volume of the text needed to achieve statistically significant results is modest but not trivial. For each of 60 samples generated by the OpenAI models, we needed roughly 1,000 lines of text after significant efforts to supply coherent prompts and keep the models both on topic and stop them from repeating text they had already generated. In a setting where one may want to find out whether the source of a particular text was AI or human, 1000+ lines of text are rare to come by, unless the text in question were a whole novel. Therefore, the utility of our test as a tool for AI-text detection is currently quite limited.

Finally, we acknowledge the limitations of our prompting and text generation methods. We wrote all prompt topics by hand in order to ensure diversity of theme and genre. More diversity, more prompt topics, or perhaps more followups to the topics could have been collected, with or without the assistance of AI, to ensure more generalizable conclusions. Yet the cost incurred to produce the final dataset exceeded $500, the total budget allotted to the project. We make our data publicly available in the hopes that it be useful to other researchers who study linguistic phenomena in long-form AI-generated text.

References

Marco Baroni. 2009. Chapter 37: Distributions in text. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics: An International Handbook*, pages 803–822. Mouton de Gruyter.

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

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.*, 15(3):39:1–39:45.

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

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. 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 7282–7296, Online. Association for Computational Linguistics.

Mark Davies. 2009. The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14(2):159–190. Publisher: John Benjamins.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. RAID: A Shared Benchmark for Robust Evaluation of Machine-Generated Text Detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12463–12492, Bangkok, Thailand. Association for Computational Linguistics.

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, Sherry Shi, and Chris Callison-Burch. 2023. Real or Fake Text?: Investigating Human Ability to Detect Boundaries between Human-Written and Machine-Generated Text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):12763–12771. Number: 11.

Liam Dugan, Andrew Zhu, Firoj Alam, Preslav Nakov, Marianna Apidianaki, and Chris Callison-Burch. 2025. GenAI content detection task 3: Cross-domain machine generated text detection challenge. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 377–388, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Susan Goldin-Meadow and Charles Yang. 2017. Statistical evidence that a child can create a combinatorial linguistic system without external linguistic input: Implications for language evolution. *Neuroscience and Biobehavioral Reviews*, 81(Part B):150 – 157.

Olivia Guest and Andrea E. Martin. 2023. On Logical Inference over Brains, Behaviour, and Artificial Neural Networks. *Computational Brain & Behavior*, 6(2):213–227.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. 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.

Daniel Jannai, Amos Meron, Barak Lenz, Yoav Levine, and Yoav Shoham. 2023. Human or Not? A Gamified Approach to the Turing Test. *arXiv preprint*. ArXiv:2305.20010 [cs].

Cameron R. Jones and Benjamin K. Bergen. 2024. People cannot distinguish GPT-4 from a human in a Turing test. *arXiv preprint*. ArXiv:2405.08007 [cs].

Tom S. Juzek and Zina B. Ward. 2024. Why Does ChatGPT "Delve" So Much? Exploring the Sources of Lexical Overrepresentation in Large Language Models. *arXiv preprint*. ArXiv:2412.11385 [cs].

Tom S Juzek and Zina B. Ward. 2025. Why does ChatGPT "delve" so much? exploring the sources of lexical overrepresentation in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6397–6411, Abu Dhabi, UAE. Association for Computational Linguistics.

Zae Myung Kim, Kwang Lee, Preston Zhu, Vipul Raheja, and Dongyeop Kang. 2024a. Threads of subtlety: Detecting machine-generated texts through discourse motifs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5449–5474, Bangkok, Thailand. Association for Computational Linguistics.

Zae Myung Kim, Kwang Lee, Preston Zhu, Vipul Raheja, and Dongyeop Kang. 2024b. Threads of Subtlety: Detecting Machine-Generated Texts Through Discourse Motifs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5449–5474, Bangkok, Thailand. Association for Computational Linguistics.

Stan A. Kuczaj. 1977. The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior*, 16(5):589–600.

Henry Kučera and W. Nelson Francis. 1967. *Computational analysis of present-day American English*. Brown University Press, Providence.

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.

Brian MacWhinney. 2000. *The CHILDES project: Tools for analyzing talk*, 3rd edition. Lawrence Erlbaum, Mahwah, NJ.

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.

R. Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. 2023. How Much Do Language Models Copy From Their Training Data? Evaluating Linguistic Novelty in Text Generation Using RAVEN. *Transactions of the Association for Computational Linguistics*, 11:652–670. Place: Cambridge, MA Publisher: MIT Press.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

William Merrill, Noah A. Smith, and Yanai Elazar. 2024. Evaluating n-gram novelty of language models using rusty-DAWG. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14459–14473, Miami, Florida, USA. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? *arXiv preprint*. ArXiv:2202.12837 [cs].

Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. 2022. Reframing instructional prompts to GPTk's language. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 589–612, Dublin, Ireland. Association for Computational Linguistics.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting Linguistic Patterns in Human and LLM-Generated News Text. *Artificial Intelligence Review*, 57(10):265.

Isabel Papadimitriou, Ethan A. Chi, Richard Futrell, and Kyle Mahowald. 2021. Deep subjecthood: Higher-order grammatical features in multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2522–2532, Online. Association for Computational Linguistics.

Julian M Pine, Daniel Freudenthal, Grzegorz Krajewski, and Fernand Gobet. 2013. Do young children have adult-like syntactic categories? Zipf's law and the case of the determiner. *Cognition*, 127(3):345–360.

Julian M Pine and Elena VM Lieven. 1997. Slot and frame patterns and the development of the determiner category. *Applied Psycholinguistics*, 18(2):123–138.

Julian M Pine and Helen Martindale. 1996. Syntactic categories in the speech of young children: The case of the determiner. *Journal of child language*, 23(2):369–395.

Jenna Russell, Marzena Karpinska, and Mohit Iyyer. 2025. People who frequently use ChatGPT for writing tasks are accurate and robust detectors of AI-generated text. *arXiv preprint*. ArXiv:2501.15654 [cs].

Teven Le Scao and Alexander M. Rush. 2021. How Many Data Points is a Prompt Worth? *arXiv preprint*. ArXiv:2103.08493 [cs].

Chantal Shaib, Yanai Elazar, Junyi Jessy Li, and Byron C Wallace. 2024. Detection and measurement of syntactic templates in generated text. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6416–6431, Miami, Florida, USA. Association for Computational Linguistics.

Arabella Sinclair, Jaap Jumelet, Willem Zuidema, and Raquel Fernández. 2022. Structural persistence in language models: Priming as a window into abstract language representations. *Transactions of the Association for Computational Linguistics*, 10:1031–1050.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Herbert S Terrace, Laura-Ann Petitto, Richard J Sanders, and Thomas G Bever. 1979. Can an ape create a sentence? *Science*, 206(4421):891–902.

Anna Theakston, Elena Lieven, Julian Pine, and Caroline Rowland. 2001. The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of child language*, 28:127–52.

Mycal Tucker, Peng Qian, and Roger Levy. 2021. What if this modified that? syntactic interventions with counterfactual embeddings. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 862–875, Online. Association for Computational Linguistics.

Virginia Valian. 2006. Young children's understanding of present and past tense. *Language Learning and Development*, 2(4):251–276.

Virginia Valian, Stephanie Solt, and John Stewart. 2009. Abstract categories or limited-scope formulae? The case of children's determiners. *Journal of Child Language*, 36(4):743–778.

Héctor Javier Vázquez Martínez. 2021. The acceptability delta criterion: Testing knowledge of language using the gradience of sentence acceptability. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 479–495, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Héctor Javier Vázquez Martínez, Annika Heuser, Charles Yang, and Jordan Kodner. 2023. Evaluating neural language models as cognitive models of language acquisition. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 48–64, Singapore. Association for Computational Linguistics.

Tianlu Wang, Rohit Sridhar, Diyi Yang, and Xuezhi Wang. 2022. Identifying and mitigating spurious correlations for improving robustness in NLP models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1719–1729, Seattle, United States. 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.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023. Larger language models do in-context learning differently. *arXiv preprint*. ArXiv:2303.03846 [cs].

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

Charles Yang. 2013. Ontogeny and phylogeny of language. *Proceedings of the National Academy of Sciences*, 110(16):6324–6327. Publisher: Proceedings of the National Academy of Sciences.

Sergio E. Zanotto and Segun Aroyehun. 2024. Human Variability vs. Machine Consistency: A Linguistic Analysis of Texts Generated by Humans and Large Language Models. *arXiv preprint*. ArXiv:2412.03025 [cs].

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate Before Use: Improving Few-Shot Performance of Language Models. *arXiv preprint*. ArXiv:2102.09690 [cs].

George Kingsley Zipf. 1949. *Human Behavior And The Principle Of Least Effort*. Addison-Wesley Press.

A Raw overlap statistics for models and humans tested

speaker	N	S	bias	r	empirical	predicted
Gail	291	741	0.879892	2.546392	0.182131	0.131964
Gail_mot	775	3012	0.862882	3.886452	0.238710	0.184154
Dominic	115	277	0.895307	2.408696	0.130435	0.130193
Dominic_mot	492	3637	0.814407	7.392276	0.272358	0.382441
Becky_mot	551	3060	0.840523	5.553539	0.323049	0.280750
Becky	354	1281	0.846214	3.618644	0.248588	0.203689
Liz_mot	566	2474	0.871059	4.371025	0.249117	0.203448
Liz	294	1135	0.881057	3.860544	0.255102	0.188859
Carl	398	3425	0.780146	8.605528	0.379397	0.470363
Carl_mot	473	3048	0.823491	6.443975	0.340381	0.338281
Joel_mot	756	2981	0.862798	3.943122	0.191799	0.187063
Joel	323	899	0.904338	2.783282	0.176471	0.121682
Ruth_mot	638	3688	0.818330	5.780564	0.285266	0.305849
Ruth	187	646	0.801858	3.454545	0.203209	0.244713
Aran	364	1499	0.824550	4.118132	0.255495	0.243752
Aran_mot	985	7073	0.823272	7.180711	0.297462	0.340261
Anne	300	1055	0.822749	3.516667	0.273333	0.219962
Anne_mot	673	5265	0.855461	7.823180	0.352155	0.338199
John_mot	674	3269	0.814010	4.850148	0.336795	0.267622
John	324	1479	0.814064	4.564815	0.274691	0.277488
Nicole_mot	753	3735	0.853280	4.960159	0.247012	0.236090
Nicole	189	458	0.879913	2.423280	0.179894	0.133672
Warren	367	2025	0.800000	5.517711	0.305177	0.329656
Warren_mot	747	4858	0.828119	6.503347	0.309237	0.319939

Table A4: Types (N), tokens (S), determiner bias score, token/type ratio (r), predicted and observed (empirical) raw overlap values for 12 children and their corresponding caretakers in the Manchester Corpus.

	N	S	bias	r	empirical	predicted
0	1774	4743	0.820000	2.673600	0.233400	0.216500
1	1872	5152	0.820000	2.752100	0.212100	0.222200
2	1845	5226	0.820000	2.832500	0.213600	0.228600
3	1715	4195	0.820000	2.446100	0.192400	0.198700
4	1845	4598	0.820000	2.492100	0.209200	0.201800
5	1926	5033	0.820000	2.613200	0.209800	0.211100
6	1991	4948	0.820000	2.485200	0.217500	0.200700
7	2030	5228	0.820000	2.575400	0.217700	0.207700
8	1902	4897	0.820000	2.574700	0.208700	0.208100
9	1479	4347	0.820000	2.939100	0.220400	0.238800
10	1505	5450	0.820000	3.621300	0.268400	0.289100
11	1863	5442	0.820000	2.921100	0.230300	0.235300

Table A4: Types (N), tokens (S), determiner bias score, token/type ratio (r), predicted and observed (empirical) raw overlap values for 12 sections of the Brown corpus.

model	trial	N	S	bias	r	empirical	predicted
gpt-4o-2024-11-20	1	304	567	0.820000	1.865132	0.095395	0.137723
gpt-4o-2024-11-20	2	407	814	0.820000	2.000000	0.108108	0.153538
gpt-4o-2024-11-20	3	413	975	0.820000	2.360775	0.152542	0.185942
gpt-4o-2024-11-20	4	300	524	0.820000	1.746667	0.086667	0.127352
gpt-4o-2024-11-20	5	345	650	0.820000	1.884058	0.110145	0.160066
gpt-4o-2024-11-20	6	368	781	0.820000	2.122283	0.122283	0.173534
gpt-4o-2024-11-20	7	391	783	0.820000	2.002558	0.112532	0.157004
gpt-4o-2024-11-20	8	347	661	0.820000	1.904899	0.112392	0.156128
gpt-4o-2024-11-20	9	363	718	0.820000	1.977961	0.132231	0.162352
gpt-4o-2024-11-20	10	294	479	0.820000	1.629252	0.091837	0.120483
gpt-4o-2024-11-20	11	297	531	0.820000	1.787879	0.077441	0.136122
gpt-4o-2024-11-20	12	357	688	0.820000	1.927171	0.120448	0.153388
gpt-4o-2024-11-20	13	391	763	0.820000	1.951407	0.094629	0.143709
gpt-4o-2024-11-20	14	366	747	0.820000	2.040984	0.103825	0.151978
gpt-4o-2024-11-20	15	352	618	0.820000	1.755682	0.088068	0.127715

Table A4: Types (N), tokens (S), determiner bias score, token/type ratio (r), predicted and observed (empirical) raw overlap values for `gpt-4o-2024-11-20`.

model	trial	N	S	bias	r	empirical	predicted
gpt-4o-mini-2024-07-18	1	717	2363	0.820000	3.295676	0.193863	0.245099
gpt-4o-mini-2024-07-18	2	616	1860	0.820000	3.019481	0.180195	0.236519
gpt-4o-mini-2024-07-18	3	589	1804	0.820000	3.062818	0.179966	0.241821
gpt-4o-mini-2024-07-18	4	604	2294	0.820000	3.798013	0.163907	0.262553
gpt-4o-mini-2024-07-18	5	564	1723	0.820000	3.054965	0.177305	0.229549
gpt-4o-mini-2024-07-18	6	537	1688	0.820000	3.143389	0.165736	0.245672
gpt-4o-mini-2024-07-18	7	547	1551	0.820000	2.835466	0.160878	0.223957
gpt-4o-mini-2024-07-18	8	666	2158	0.820000	3.240240	0.193694	0.249588
gpt-4o-mini-2024-07-18	9	691	2375	0.820000	3.437048	0.189580	0.260209
gpt-4o-mini-2024-07-18	10	492	1667	0.820000	3.388211	0.207317	0.255604
gpt-4o-mini-2024-07-18	11	529	1742	0.820000	3.293006	0.156900	0.230636
gpt-4o-mini-2024-07-18	12	705	2430	0.820000	3.446809	0.194326	0.256132
gpt-4o-mini-2024-07-18	13	609	2048	0.820000	3.362890	0.178982	0.255989
gpt-4o-mini-2024-07-18	14	613	1758	0.820000	2.867863	0.153344	0.224515
gpt-4o-mini-2024-07-18	15	610	1954	0.820000	3.203279	0.188525	0.227921

Table A4: Types (N), tokens (S), determiner bias score, token/type ratio (r), predicted and observed (empirical) raw overlap values for `gpt-4o-mini-2024-07-18`.

model	trial	N	S	bias	r	empirical	predicted
o1-preview-2024-09-12	1	544	1345	0.820000	2.472426	0.128676	0.203507
o1-preview-2024-09-12	2	725	2313	0.820000	3.190345	0.158621	0.233397
o1-preview-2024-09-12	3	603	1503	0.820000	2.492537	0.116086	0.199715
o1-preview-2024-09-12	4	530	1212	0.820000	2.286792	0.141509	0.202243
o1-preview-2024-09-12	5	586	1575	0.820000	2.687713	0.139932	0.221396
o1-preview-2024-09-12	6	549	1315	0.820000	2.395264	0.136612	0.203304
o1-preview-2024-09-12	7	665	1938	0.820000	2.914286	0.160902	0.232420
o1-preview-2024-09-12	8	554	1165	0.820000	2.102888	0.135379	0.179700
o1-preview-2024-09-12	9	461	935	0.820000	2.028200	0.119306	0.178230
o1-preview-2024-09-12	10	486	1188	0.820000	2.444444	0.148148	0.205791
o1-preview-2024-09-12	11	570	1490	0.820000	2.614035	0.112281	0.211588
o1-preview-2024-09-12	12	640	1645	0.820000	2.570312	0.120313	0.211472
o1-preview-2024-09-12	13	516	1188	0.820000	2.302326	0.129845	0.198295
o1-preview-2024-09-12	14	590	1354	0.820000	2.294915	0.113559	0.188407
o1-preview-2024-09-12	15	420	954	0.820000	2.271429	0.135714	0.195233

Table A4: Types (N), tokens (S), determiner bias score, token/type ratio (r), predicted and observed (empirical) raw overlap values for `o1-preview-2024-09-12`.

model	trial	N	S	bias	r	empirical	predicted
o1-mini-2024-09-12	1	466	1121	0.820000	2.405579	0.113734	0.202470
o1-mini-2024-09-12	2	381	1011	0.820000	2.653543	0.078740	0.214553
o1-mini-2024-09-12	3	419	1790	0.820000	4.272076	0.155131	0.294557
o1-mini-2024-09-12	4	353	1091	0.820000	3.090652	0.127479	0.250204
o1-mini-2024-09-12	5	415	1313	0.820000	3.163855	0.113253	0.253715
o1-mini-2024-09-12	6	390	1104	0.820000	2.830769	0.110256	0.232185
o1-mini-2024-09-12	7	376	1038	0.820000	2.760638	0.069149	0.248325
o1-mini-2024-09-12	8	411	1220	0.820000	2.968370	0.136253	0.246390
o1-mini-2024-09-12	9	508	1717	0.820000	3.379921	0.127953	0.259590
o1-mini-2024-09-12	10	324	917	0.820000	2.830247	0.064815	0.214257
o1-mini-2024-09-12	11	374	1077	0.820000	2.879679	0.114973	0.238506
o1-mini-2024-09-12	12	421	1108	0.820000	2.631829	0.097387	0.223555
o1-mini-2024-09-12	13	396	1239	0.820000	3.128788	0.093434	0.242729
o1-mini-2024-09-12	14	476	1226	0.820000	2.575630	0.102941	0.216957
o1-mini-2024-09-12	15	393	1027	0.820000	2.613232	0.127226	0.223354

Table A4: Types (N), tokens (S), determiner bias score, token/type ratio (r), predicted and observed (empirical) raw overlap values for `o1-mini-2024-09-12`.