

From Babbling to Fluency: Evaluating the Evolution of Language Models in Terms of Human Language Acquisition

Qiyuan Yang^{1*}, Pengda Wang^{2*}, Luke D. Plonsky³, Frederick L. Oswald², and Hanjie Chen¹

¹Department of Computer Science, Rice University

²Department of Psychological Sciences, Rice University

³Department of English-Applied Linguistics, Northern Arizona University
{qy28, pw32, fo3, hc86}@rice.edu; luke.plonsky@nau.edu

Abstract

We examine the language capabilities of language models (LMs) from the critical perspective of human language acquisition. Building on classical language development theories, we propose a three-stage framework to assess the abilities of LMs, ranging from preliminary word understanding to complex grammar and complex logical reasoning.¹ Using this framework, we evaluate the generative capacities of LMs using methods from linguistic research. Results indicate that although recent LMs outperform earlier models in overall performance, their developmental trajectory does not strictly follow the path of human language acquisition. Notably, in generation tasks, LMs are more similar to human performance in areas where information is easier to extract from the corpus, such as average word length, clauses, and auxiliary verbs. Newer LMs did not exhibit significant progress in terms of specific dimensions, such as clauses and auxiliary verbs, where the variation across corpora is relatively limited. Register theory offers a plausible explanation for these observations, suggesting that the linguistic features of the training data have a substantial impact on the models' abilities.

1 Introduction

Since the advent of early natural language processing (NLP) systems such as ELIZA (Weizenbaum, 1966) and SHRDLU (Winograd, 1971) in the 1950s, researchers have been striving to develop computer programs to understand human language. With continuous technological advancements, we have witnessed the rise of language models (LMs), which have achieved unprecedented success in language understanding and language generation (e.g., Gemini, Anil et al., 2023; GPT-4, Achiam et al.,

¹Code and dataset are available at <https://github.com/ericyang1029/Language-Acquisition>

*Equal contribution

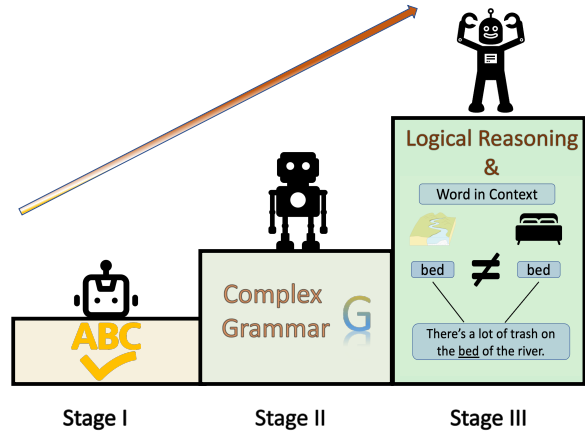


Figure 1: Three-Stage Anatomy of Language Acquisition.

2023; Llama 3, Dubey et al., 2024). These models not only handle complex contexts and generate coherent, human-like text; they also exhibit emergent reasoning abilities and a plausible degree of creativity.

As the capabilities of LMs continue to grow, so does the need for comprehensive evaluations of their performance. To date, this need has produced a series of benchmark studies that evaluate the capabilities of LMs across various language tasks, such as text classification (Sun et al., 2023), natural language inference (NLI) (Ravichander et al., 2019), and question answering (Kwiatkowski et al., 2019), with the goal of comparing different models, identifying their limitations in terms of these tasks, and providing guidance for future model development. However, most existing benchmarks, such as GLUE (Wang et al., 2019), SuperGLUE (Wang et al., 2020) and MMLU (Hendrycks et al., 2021), while thoroughly evaluating models on specific language tasks, overlook the focus of our current paper: i.e., understanding model capabilities in terms of the developmental stages of human language acquisition (Goldberg, 2005). Similar to how humans acquire language through extensive exposure to

spoken or written words as they develop, LMs are similarly trained on large collections of text. Both humans and LMs build their language skills by repeatedly encountering language, gradually forming and refining stable patterns and associations. Insights from previous studies on the stages of human language development could offer valuable reference points for understanding this process in terms of LMs.

As one of the unique abilities of humans, the acquisition of language has long been a key area of research in psycholinguistics. During the process of language acquisition, humans go through multiple stages, from imitation and rule learning to complex contextual understanding (Goldberg, 2005). These stages bear some resemblance to the way current LMs are trained. For instance, LMs learn the statistical patterns and grammatical rules of language through training on large-scale data, similar to how infants develop language abilities by receiving a vast amount of input through listening and speaking. If we design theory-driven tests based on the human language acquisition process to evaluate the capabilities of LMs, it could help us better understand the nature, potential, and limitations of LMs in their development.

Our work draws on classical theories of human language development to assess LMs in terms of a three-stage human language development framework (Chomsky, 2014; Loban, 1976; Pinker, 2003), as shown in Figure 1. The first stage involves developing basic language understanding, similar to early language acquisition in infants. At this stage, we evaluate the model’s ability to recognize vocabulary, grasp syntax, and perform simple reasoning. In the second stage, the focus shifts to mastering complex grammar and semantics, where the model demonstrates a deeper understanding of language rules and logical relationships between sentences. The third stage assesses advanced language abilities, evaluating the model’s capacity for complex reasoning and logical analysis.

We further investigate another theory: register theory in linguistics, which posits that different language use scenarios influence the form and structure of language (Halliday, 1977; Matthiessen, 1993). This theory offers insights into the extent to which models’ abilities depend on the linguistic features encountered in specific situations, referred to as registers. In LMs, the training corpus will reflect some registers but not others, which can raise general questions or concerns about the generaliz-

ability and biases contained in any given corpus.

We evaluated 15 LMs from 2019 to 2024, excluding instruction fine-tuned or chat versions, with varying parameter sizes (see §4.1). Our findings include: (1) newer LMs generally outperform older ones, though performance varies by task; (2) LMs do not follow human language acquisition patterns but rather reflect changes in architecture and training data; (3) for easily accessible information such as average word length, clauses, and auxiliary verbs, LMs show little improvement over time. Recent models have demonstrated minimal progress in these areas due to limited variation across corpora. Overall, register theory, which focuses on data, better explains model differences than human developmental processes.

2 Related Works

LMs are computational systems designed to understand and generate text in human language. Over time, advancements in LMs, particularly in pre-trained models like GPT (Radford et al., 2019) and BERT (Devlin et al., 2019), have significantly improved performance across various NLP tasks. Large LMs, which leverage vast amounts of data and computational power, can capture more intricate nuances in language (Bommasani et al., 2022; Wei et al., 2022a), improving its generative capabilities involving masked token or next-token predictions.

These models are typically fine-tuned for specific tasks after pre-training, further enhancing their adaptability and versatility in practical applications (Gururangan et al., 2020). As noted previously, systematically evaluating the performance of LMs is critical as they grow in their complexity and diversity (Srivastava et al., 2023). Benchmarking not only provides a standardized way to compare different models; it also highlights areas where improvements are needed, guiding future advancements in the field.

There are many benchmarks that evaluate LMs’ abilities. Some focus on specific aspects, whereas others cover a broad range of tasks. For instance, the SST2 dataset (Socher et al., 2013) measures text classification and the TriviaQA dataset (Joshi et al., 2017) focuses on question answering. Additionally, comprehensive benchmark suites like GLUE (Wang et al., 2019), SuperGLUE (Wang et al., 2020), and MMLU (Hendrycks et al., 2021) assess multitask language understanding across a wide

range of topics and tasks. However, these benchmarks do not provide insights about a model’s capabilities in terms of human language acquisition, such as in the three-stage framework we provided. Insights from previous studies on the stages of human language development may offer valuable reference points for evaluating models’ performance.

Previous studies have demonstrated that models can learn hierarchical syntactic structures and exhibit sensitivity to various linguistic phenomena, even when trained with the amount of data that humans typically encounter (Millière, 2024; Wilcox et al., 2024). Assessing these models through the lens of human language development can provide further insights and deepen our understanding of LMs’ capabilities.

Human language development is a gradual, stage-based process. In the following section (§3), we will provide a more detailed description of this process, along with a breakdown of language capabilities at each developmental stage.

3 Psycholinguistics View Framework and Datasets

Psycholinguistics explores the cognitive processes behind language acquisition, focusing on how humans gradually develop language abilities. We primarily focus on research related to the various stages of language development.

Previous research has established that language development follows a relatively stable trajectory, with several key stages identifiable along the way. For example, Gesell et al. (1946) found that the development of spoken language demonstrates consistent growth, as reflected in metrics such as the average number of words per communication unit, the number of clauses per unit, and the elaboration between subjects and verbs.

Similarly, Templin’s (1957) analysis of subordinate clause usage also underscores these stages, showing that eight-year-old children use subordinate clauses significantly more often than three-year-olds, marking a pivotal point in language acquisition. And Gesell et al. (1946) indicated that the development of spoken language shows a relatively stable growth trend. For example, the average number of words per communication unit (C-Unit), the number of clauses in each communication unit, and the amount of elaboration between subjects and verbs all continue to increase.

3.1 Framework

Combining the findings above with those of Watts (1944); O’Donnel et al. (1967); Paul (2007) and the summary of Loban (1976), we can roughly divide the overall process of language development into three stages:

Stage I (Ages 0-6): At this stage, children primarily focus on understanding vocabulary, and simple syntactic structures begin to emerge. They gradually learn to use pronouns and verbs and become able to distinguish between the present and past tense. Although language expression remains relatively simple at this age, the use of compound sentences increases, especially those that express conditionality and causality. Using words like “why,” “because,” and “if,” children begin to engage in preliminary causal reasoning, though this ability is not yet fully developed.

Stage II (Ages 6-12): During this stage, the development of language gradually moves towards more complex grammatical structures. They begin to master finer syntactic elements, such as predicate-argument structures, prepositional phrases, subordinate clauses, and the use of active and passive voice. Their semantic understanding also advances, enabling them to grasp the implied meanings of words (e.g., “run” implies “movement”) and handling negation through pre-pending or appending particles to the stem of a word (i.e., morphological negation, refers to the process of creating a negative form of a word by adding a prefix, such as when “possible” becomes “impossible.” This involves using prefixes like “un-,” “in-,” or “im-” to change the meaning of the original word to its opposite).

In addition, during this stage, children develop the ability to recognize named entities, quantifiers, and complex concepts such as factuality, symmetry, and redundancy.

Stage III (Above age 12): At this stage, children’s language abilities are reflected not only in the complexity of their verbal expression but, more significantly, in their use of logical reasoning and abstract thinking. They begin to engage in spatial reasoning, deductive reasoning, and syllogistic analysis, which allows them to use language with greater precision and rigor. Additionally, they become adept at resolving ambiguities in words with multiple meanings and demonstrate a marked improvement in reading comprehension skills.

3.2 Datasets

Within each stage we just introduced, we compile several datasets and introduce them in the following section.* For an overview of the datasets, please refer to Table 2 in Appendix E and see Table 5 in Appendix E for the example of each dataset.

3.2.1 Stage I

one-word understanding: To assess the LM’s understanding of individual vocabulary items, we selected examples from publicly accessible vocabulary sample tests (Test, 2024; EnglishTestsOnline.com, 2024) and randomly extracted frequently used vocabulary with brief examples from Oxford_Learner’s_Dictionary (2024).

In this task, LMs will be asked to answer simple multiple-choice questions. They will need to choose one of the four choices (a word or phrase) that makes the most sense in the given context.

agent-action-object (AAO): To test whether LMs possess the knowledge to decide whether it is reasonable to take an action on the object, we chose the “subject-verb-trans” set from BLiMP (Warstadt et al., 2023) as our AAO dataset.

In this task, LMs will be provided two sentences that have minimal differences (one or two words), where one of the two sentences is grammatically correct, and the other is not. LMs will be asked to distinguish between correct and incorrect sentences.

bc-if-why: We select examples containing the words {because, if, why} from the Multi-Genre Natural Language Inference (MNLI) dataset (Williams et al., 2018), to test the models’ preliminary expressiveness in terms of conditionality and causality.

Following the same format in the MNLI dataset, we let the models perform a three-class classification task. Given premise and hypothesis, models will need to classify them into {entailment, neutral, contradiction}.

3.2.2 Stage II

Grammar-comp: To evaluate complex grammatical structures, we included more comprehensive and diverse grammatical types (e.g. quantifiers, belief verbs) in this task from MNLI (Williams

et al., 2018). We also exclude instances containing participial words that are not typically mastered at this stage. We keep the same task setup as in “bc-if-why” in Stage I.

BLiMP-comp: To minimize the influence of inference on grammar tasks in addition to MNLI, we extract minimal pair tasks from BLiMP (Warstadt et al., 2023), which includes a wide range of grammatical phenomena, from subject-verb-agreement to syntactic structure. We select those subsets with human average performance of at least 80% accuracy as tests. The format is the same as the AAO task.

CoLA (Warstadt et al., 2018): Unlike the other two tasks in this stage, models are required to classify a sentence as either grammatically correct or incorrect, assigning it to one of two classes: True or False, respectively.

3.2.3 Stage III

WiC: The WiC dataset (Pilehvar and Camacho-Collados, 2019) focuses on words that have multiple meanings. We used it to test the models’ ability to probe both the context of the sentences and different definitions of the word when those exist.

In this task, two sentences will be given, where each has one word in common, but they may or may not have the same meanings. Models will need to judge whether this word has the same meaning or not under these two contexts.

ReClor: This dataset (Yu et al., 2020) is composed of complex logical reasoning questions. We used it to test whether the models possess complex language abilities, including word understanding, grammatical accuracy, inference, and reasoning.

During this task, models will do multiple-choice questions. Provided with a context and a question, models are expected to choose the most suitable answers to the question from one of four choices.

4 Experimental Setup

In this section, we introduce the LMs we tested (§4.1), the testing methods for different tasks performed by the LMs (§4.2), as well as the evaluation method (§4.3).

4.1 Models

We investigated 15 LMs in total (excluding instruction fine-tuned or chat versions) over a broad time period (2019 to 2024) and with varying model parameter sizes.

*Note that we filter the training dataset and restrict the average C-Unit. In some cases (e.g., bc-if-why), because there is not a sufficient number of filtered examples from its evaluation set, we randomly split off 20% of the training dataset for validation. For datasets that do not require filtering, the evaluation sets are provided.

These include GPT-2 (gpt-2-large, gpt-2-xl; Radford et al., 2019), RoBERTa (RoBERTa-base, RoBERTa-large; Liu et al., 2019), ALBERT (ALBERT-xlarge, ALBERT-xxlarge; Lan et al., 2019), Google T5 (T5-3b, T5-large; Raffel et al., 2020), OPT (opt-1.3b, opt-2.7b; Zhang et al., 2022), Llama2 (Llama-2-7b-hf), Mistral (Mistral-7B-v0.3; Jiang et al., 2023), Llama3 8B (Meta-Llama-3-8B), and Gemma2 (gemma-2-2b, gemma-2-9b).

4.2 Testing Methods

We use three different strategies to test the performance of LMs because of the specific task design of certain datasets (e.g., classification), and LMs’ architecture differences.

Classification Task: In this type of task, sentences are given as inputs to models. Models will output a class label (e.g., {0, 1} for two-class classification, {0, 1, 2} for three-class classification).

Minimal Pair Task and Vocabulary Task: In these two kinds of tasks, we will either calculate the loss for decoder-only models or compare the probability distributions of the masked token through Masked Token Prediction (MLM) (BERT-style) or Span Predictions (T5). Please refer to Appendix C.1 for details on the format.

Reading Comprehension Task: For this task, we select either the available chat versions or the instruction-fine-tuned versions of our chosen models, as these can be prompted to answer questions in a designated format. In addition to the normal prompt, we also apply the zero-shot CoT (Wei et al., 2022b) and one-shot ICL (Brown et al., 2020) to determine whether any further improvement in the performance of the LMs can be obtained.

Generation Task: The chat and instruction-fine-tuned versions of the models are prompted with instructions for ten different topics, taken from GRE public issue writing prompts (Educational Testing Service). Sample essays with full scores are sourced from (Yu, 2024) to compare with the performance of the LMs on this task.

4.3 Evaluation Method

We report accuracy as our main performance metric, as in the original formulation, because most of our testing data is balanced. CoLA dataset (Warstadt et al., 2018) also uses the Matthews correlation coefficient (see C.2).

Normalized Accuracy: Although the NLI task has a baseline accuracy of 0.33 (random

guess), tasks with four choices, such as one-word understanding, have a baseline accuracy of 0.25. Therefore, it is unreasonable to compare them solely on their original accuracy. We have therefore normalized each metric by the following formula:

$$Normalized_Accuracy = \frac{A - R}{1 - R}$$

where A is the observed accuracy, R is the accuracy of a random guess. This formula is the same as Cohen’s kappa coefficient for rating tasks, which takes random rater agreement into account (Cohen, 1960).

5 Experimental Results

We first analyzed whether the LMs’ overall developmental trends between the years 2019 and 2024 were consistent with the developmental trajectory of human language (§5.1). On this basis, we further explored three core questions: (1) Did scale matter? (2) Did architecture matter? (3) Did data matter? Finally, we conducted a comprehensive and in-depth evaluation of the models’ generative abilities from a linguistic perspective (§5.2).

5.1 Overall Trends in Language Models’ Development

Here, we focused on the overall development trends of LMs, and whether these models mimic the process of human language acquisition. As noted previously, just as humans learn language from an early age by being exposed to a large amount of spoken or written language, LMs are trained on vast text corpora. Both humans and LMs develop language abilities through repeated exposure to language, forming patterns and associations over time. Previous research on the stages of human language development may serve as a reference.

As mentioned earlier, these datasets have been divided into tasks based on theories of human language development. We anticipated that certain LMs would exhibit stronger performance in the early stages of language acquisition but show more modest results in the later stages. Further, if these stages of human language development hold for the development of LMs, then if an LM achieves relatively good results in the third stage, then it should also demonstrate corresponding success in the first and second stages on which the third stage depends. Despite this theoretical motivation, the experimental results did not support this hypothesis.

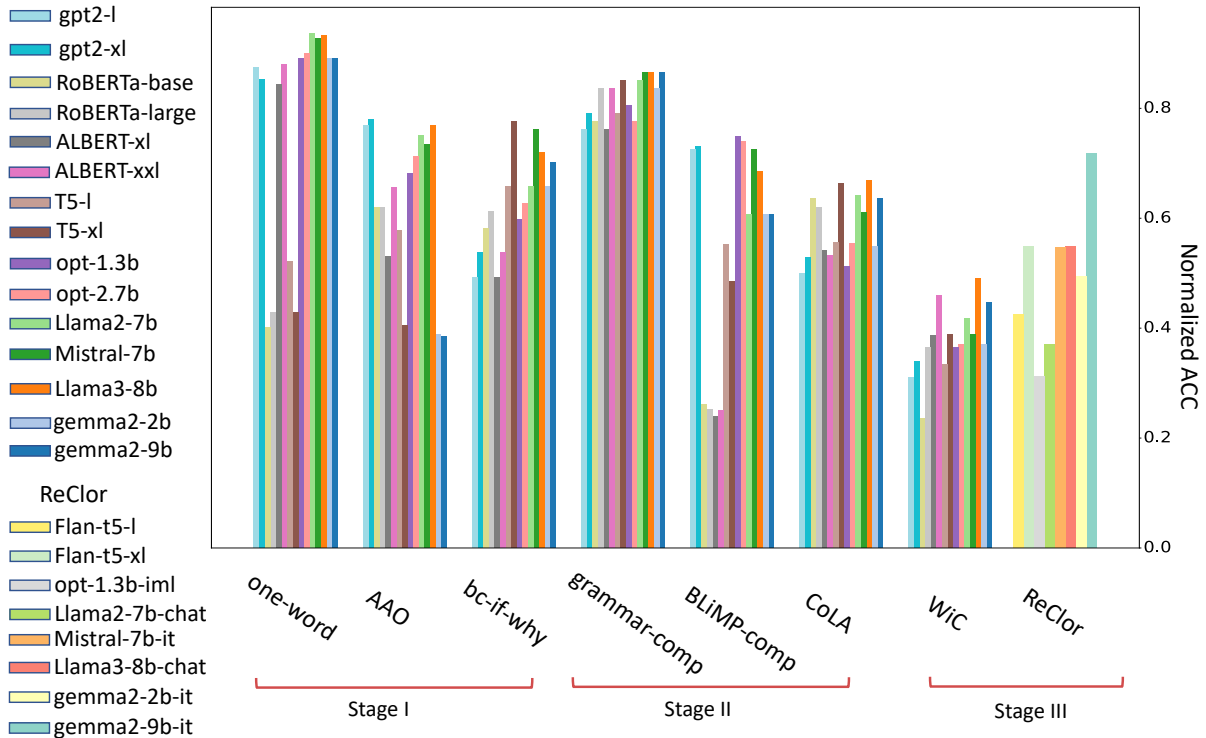


Figure 2: Performance of LMs across three stages. The upper right legend corresponds to models tested in tasks except for ReClor. The lower right legend corresponds to models tested in ReClor. For each task, models are ordered by their time released, and the tie is broken by their parameter sizes. Results from CoLA also use a different metric; please refer to Figure 4 in Appendix E.

Figure 2 displays our overall results. In Stage I, we first tackled fundamental tasks of human language acquisition, such as understanding individual words. Most models performed well at this stage, but a few lagged behind. For example, the accuracy of T5 and RoBERTa was only half that of other models in one-word understanding. We found that Gemma2 performed well in many tasks; however, it fell short compared to other models on the AAO task. After conducting some experiments (see Appendix A) on these models, we discovered that T5 and RoBERTa did not perform well on questions requiring contextual information. However, the fine-tuned versions of T5 and Gemma2 excelled in one-word understanding and the AAO task, respectively.

Stage II involved more complex grammatical knowledge, yet most LMs did not share this difficulty, performing as well as, or even better than, they did in stage I. Notably, despite similar overall performance, there were significant differences in the models’ scores across different grammatical phenomena from BLiMP-comp. Please refer to Table 4 in Appendix E for detailed examples.

In Stage III, performance differences among the

LMs became more pronounced across various tasks. For the WiC task, the LMs failed to demonstrate comparative performance relative to other tasks in Stage I and Stage II. In the ReClor task, the fine-tuned opt-1.3b model and Llama2-chat version performed poorly, while Gemma2-9b-instruct achieved higher accuracy. Moreover, one-shot ICL and CoT learning did not significantly improve model performance in this task (see Table 3 in Appendix E for more details).

Does Scale Matter? Although previous research has shown that the performance of LMs often improves with the expansion of model parameters (Kaplan et al., 2020), in most of the ability tests we conducted across different stages of language development, there was no significant difference in performance between small models and their larger counterparts. The only exception was the complex task ReClor (in Stage III), where larger models significantly outperformed smaller ones.

Just like previous research (e.g., Millière, 2024; Wilcox et al., 2024), our results also support the idea that small models can effectively encode sufficient information for certain tasks, meaning that increasing model parameters is not the only path

to improving performance. Therefore, instead of solely pursuing larger models, drawing insights from linguistic research might be a more effective way to enhance overall model performance (Millière, 2024; Wilcox et al., 2024).

Does Architecture Matter? We noticed that, in classification tasks, encoder models (including T5, which only uses its encoder part for classification), even with smaller numbers of parameters, almost equalize or exceed the performances of decoder models with larger numbers of parameters. The bidirectional property of encoder models could contribute to this.

To master NLI and WiC tasks, it is pivotal to possess the inter-relationship between tokens in two sentences. Consequently, models with encoders could cross-attend to previous and later contextual information in the question and thus manage such tasks well.

For tasks that compare loss between sentence pairs (AAO and one-word), most decoder-only models, such as GPT-2, outperform encoder-only or encoder-decoder models (e.g., T5 and RoBERTa). The differences in architecture determine how they tackle such problems, particularly with prediction loss (e.g., MLM vs. next-token prediction).

We suspect that the randomness introduced by masking tokens (or corruption rates for T5) could contribute to this difference. Additionally, Next Sentence Prediction (NSP) might play an important role in one-word understanding tasks. Even with larger batch sizes, models such as RoBERTa and T5, which are not trained on NSP, may lack the ability to model sentence-to-sentence transitions, which is essential for that task.

Do Data Matter? As the representations in AI models are converging (Huh et al., 2024), the scale and the quality of data that they learn from are the key to their performance. We found that as models' pretraining data scale up, regardless that bigger is not always better, there was a trend to perform better in each stage (see Figure 6, 7, 8 in Appendix E).

Noticeably, Mistral keeps an impressive performance-to-data volume ratio, but it does not bear this advantage in stage III. Although there might be disparities among model sizes, we could anticipate that with a larger amount of training data, LMs could learn richer knowledge and generalize it better.

5.2 Language Models' Generation Ability

We also evaluated the generation abilities of some LMs through the generation task. Here, we regard generation ability as a reflection of LMs' overall capability, as generation requires word-level understanding, flexible use of grammatical knowledge, and strong logical reasoning skills to ensure sentence completeness and fluency.

In the field of linguistics, extensive research has explored co-occurrence patterns of language features. Drawing on the study of the Multi Dimensional Analysis Tagger (MAT) by Nini (2019), which replicates the procedure by Biber (1988), we compared five representative dimensions.

NN (nouns that are not identified as nominalizations or gerunds): The use of nouns is an important component of syntactic structure, helping to assess whether the model handles nouns accurately and flexibly.

AWL (average word length): Word length reflects the complexity of the generated text and the diversity of language style, measuring the model's lexical richness.

Clause (a collection of adjectival and adverbial clauses): The frequency and diversity of clause use reflect the model's ability to generate complex sentences, showcasing its mastery of advanced grammar.

TTR (type-token ratio): This dimension evaluates the richness of the generated text in terms of lexical diversity, indicating the model's flexibility in word choice.

Auxiliary verbs (e.g., modal verbs expressing possibility, prediction, and necessity): The use of auxiliary verbs reveals whether the model can express complex reasoning and logical relationships, serving as an important indicator of reasoning ability in generation tasks.

In all five dimensions, we found that humans tend to exhibit more variation than LMs in the NN (noun usage) and TTR (type-token ratio) dimensions, whereas no significant differences were observed on the other three dimensions (see Figure 3). We believe this variation is due to the following reasons:

(1) NN: Humans, when using language, tend to flexibly choose vocabulary and expressions depending on the context, topic, and purpose of communication. The use of nouns may reflect humans' ability to name objects, concepts, or abstract ideas within a specific context, and this ability becomes

more diverse as topics change. On the other hand, LMs, though trained on large corpora, may rely on more frequent patterns or words during generation rather than adjusting as flexibly as humans do based on context.

(2) TTR: Human language ability is often characterized by broad vocabulary use, especially when dealing with complex or rich topics, where such lexical diversity becomes more apparent. In contrast, LMs might tend to use more common words when generating text, particularly if certain words appear more frequently in the training data, leading to less flexibility in lexical diversity, compared to humans.

Overall, although LMs can simulate a certain level of human linguistic diversity, we believe that due to their reliance on training data, they may not exhibit the same level of variation and flexibility as humans when producing new linguistic expressions. In the other three dimensions (such as word length, use of subordinate clauses, and auxiliary verbs), the differences between LMs and humans were smaller, likely because these dimensions depend more on grammatical structure and syntactic rules, which are more clearly defined in the training of LMs, allowing them to match human performance in these areas.

Language Models’ Development in Generation

We also explored the relationship between these five dimensions and the development trends of LMs. We found that except for Clause and Auxiliary verbs, NN, AWL, and TTR showed significant progress (see Figure 9). This phenomenon may be due to improvements in the training corpora for models. The progress in NN and AWL may reflect an enhancement in the models’ ability to generate complex and precise vocabulary. As LMs developed, their vocabulary size, semantic understanding, and contextual processing capabilities improved through learning from training data, enabling them to generate richer vocabulary and longer, more complex structures. The increase in TTR indicates that the model can use a wider range of vocabulary when generating text, rather than repeatedly using the same words. This could be attributed to the model’s ability to better capture lexical diversity when processing large-scale training corpora and reflect this diversity in its generation tasks.

In contrast, the trends for Clause and Auxiliary verbs showed less noticeable changes, possibly be-

cause these features involve more complex grammatical structures and logical reasoning. Models have made progress in vocabulary generation, yet they still face significant challenges in accurately generating more complex clauses and auxiliary verbs. This may require deeper syntactic understanding and stronger logical reasoning abilities, which are improving at a slower pace.

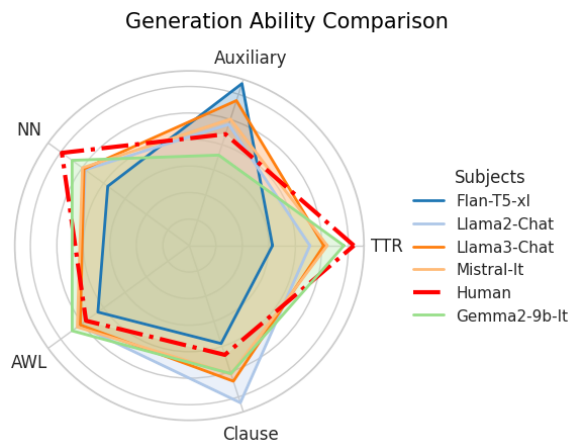


Figure 3: Generation Abilities of six models along five selected dimensions.

6 Conclusion

We evaluated LMs by incorporating theories from human language acquisition. Building on classical language development theories, we proposed a three-stage framework to assess the abilities of LMs. By and large, we observed that LMs do not conform to human language acquisition patterns. Although some LMs performed competitively in the later stages, they struggled with tasks in the earlier stages. This may be due to their specific architectures, parameter sizes, and the language corpora they were trained on.

Models show smaller differences from human performance in areas where information is easier to extract from the corpus, such as average word length, clause structure, and auxiliary verb usage in generation tasks. For dimensions that do not vary significantly across corpora, the models’ performance similarly does not show a significant improvement.

Register theory offers a plausible explanation for these observations, suggesting that the linguistic features of the training data substantially influence the models’ abilities.

Limitations

This evaluation was necessarily limited by the genres of our collected dataset, which consisted entirely of text. Texts represent only part of the information acquired during human language acquisition. For example, Barreto (2019) introduced visual questions in the CELF-5 that assessed children’s understanding of spatial terms, requiring the examinee to identify the position of an object in a picture. Similarly, the TOLD-P:5 (Newcomer and Hammill, 2018) assessed children’s spoken language skills through tasks such as defining spoken words and demonstrating an understanding of their meanings. To explore this topic further, a multimodal dataset incorporating images, videos, and speech would have been necessary.

Moreover, since the aforementioned assessments were commercially available, accessibility issues arose concerning such datasets. In the spirit of open science, future work should focus on creating similar datasets that are open to a wide range of research communities.

Additionally, research by McMurray et al. (2014) showed individual differences in human language abilities. Similarly, LMs could have been developed to model such variations more closely.

Finally, due to the rapid advancements in LMs and their increasing parameter sizes, a continuous and sustainable evaluation of these models might have been required.

Ethics Statement

The datasets we compiled are all publicly available for research purposes (under CC-BY 4.0 license or unspecified). We have manually checked each example from the one-word understanding we collected and modified to ensure it does not contain any harmful information or bias.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Monica Barreto. 2019. *CELF-5*, pages 1–4. Springer New York, New York, NY.

Douglas Biber. 1988. *Variation across Speech and Writing*. Cambridge University Press, Cambridge, UK. Tagger for the multidimensional functional analysis of English texts.

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural language toolkit*.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, and et Noah Goodman. 2022. *On the opportunities and risks of foundation models*. *Preprint*, arXiv:2108.07258.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. *Language models are few-shot learners*. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2024. Benchmarking large language models on answering and explaining challenging medical questions. *arXiv preprint arXiv:2402.18060*.

Noam Chomsky. 2014. *Aspects of the Theory of Syntax*. 11. MIT press.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. *Scaling instruction-finetuned language models*. *Preprint*, arXiv:2210.11416.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Chunyan Deng, Yilun Zhao, Xiangru Tang, Mark Gestein, and Arman Cohan. 2024. *Investigating data*

- contamination in modern benchmarks for large language models. *Preprint*, arXiv:2311.09783.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *Bert: Pre-training of deep bidirectional transformers for language understanding*. *Preprint*, arXiv:1810.04805.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Educational Testing Service. Gre: Graduate record examinations. <https://www.ets.org/gre.html>. Accessed: 2024-10-02.
- EnglishTestsOnline.com. 2024. 262 everyday vocabulary: Collective nouns test. Accessed: 2024-10-11.
- Arnold Gesell, Frances Lillian Ilg, Louise Bates Ames, and Glenna E Bullis. 1946. The child from five to ten.
- Adele E Goldberg. 2005. *Constructions at work: The nature of generalization in language*. Oxford University Press.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. *Don't stop pretraining: Adapt language models to domains and tasks*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.
- Michael AK Halliday. 1977. Text as semantic choice in social contexts. *Grammars and descriptions*, 176225.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. *Measuring massive multitask language understanding*. *Preprint*, arXiv:2009.03300.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Mark Grefenstette. 2020. *spacy: Industrial-strength natural language processing in python*. Version 2.3.5.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. *Lora: Low-rank adaptation of large language models*. *arXiv preprint arXiv:2106.09685*.
- Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. *The platonic representation hypothesis*. *Preprint*, arXiv:2405.07987.
- Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O'Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Ves Stoyanov. 2023. *Opt-impl: Scaling language model instruction meta learning through the lens of generalization*. *Preprint*, arXiv:2212.12017.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. *Mistral 7b*. *Preprint*, arXiv:2310.06825.
- Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. *Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension*. *Preprint*, arXiv:1705.03551.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. *Scaling laws for neural language models*. *Preprint*, arXiv:2001.08361.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. *Albert: A lite bert for self-supervised learning of language representations*. *arXiv preprint arXiv:1909.11942*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. *Roberta: A robustly optimized bert pretraining approach*. *arXiv preprint arXiv:1907.11692*.
- Walter Loban. 1976. Language development: Kindergarten through grade twelve. ncte committee on research report no. 18.
- CMIM Matthiessen. 1993. Register in the round: diversity in a unified theory of register analysis. *Register analysis: Theory and practice*, pages 221–292.
- Bob McMurray, Cheyenne Munson, and J. Bruce Tomblin. 2014. *Individual differences in language ability are related to variation in word recognition, not speech perception: evidence from eye movements*. *Journal of Speech, Language, and Hearing Research*, 57(4):1344–1362.
- Rapha  l Milliere. 2024. Language models as models of language. *arXiv preprint arXiv:2408.07144*.
- Phyllis L. Newcomer and Donald D. Hammill. 2018. *Test of Language Development-Primary: Fifth Edition (TOLD-P:5)*. Pro-Ed, Austin, TX.

- Andrea Nini. 2019. The multi-dimensional analysis tagger. In Tony Berber Sardinha and Marcia Veirano Pinto, editors, *Multi-Dimensional Analysis: Research Methods and Current Issues*, pages 67–94. Bloomsbury Academic, London; New York.
- RC O’Donnel, WJ Griffin, and RC Norris. 1967. Syntax of kindergarten and elementary school children. *National Council of Teachers of English, Champaign*, 111.
- Oxford_Learner’s_Dictionary. 2024. Oxford learner’s dictionaries. <https://www.oxfordlearnersdictionaries.com/>. Accessed: 2024-10-11.
- Rhea Paul. 2007. *Language Disorders from Infancy Through Adolescence: Assessment & Intervention*.
- Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. **Wic: the word-in-context dataset for evaluating context-sensitive meaning representations**. *Preprint*, arXiv:1808.09121.
- Steven Pinker. 2003. *The language instinct: How the mind creates language*. Penguin UK.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. **Language models are unsupervised multitask learners**. *OpenAI Blog*, 1(8).
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. **Exploring the limits of transfer learning with a unified text-to-text transformer**. *Journal of Machine Learning Research*, 21:1–67.
- Abhilasha Ravichander, Aakanksha Naik, Carolyn Rose, and Eduard Hovy. 2019. Equate: A benchmark evaluation framework for quantitative reasoning in natural language inference. *arXiv preprint arXiv:1901.03735*.
- Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. **Detecting pretraining data from large language models**. *Preprint*, arXiv:2310.16789.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. **Recursive deep models for semantic compositionality over a sentiment treebank**. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, and etc Alexander W. Kocurek. 2023. **Beyond the imitation game: Quantifying and extrapolating the capabilities of language models**. *Preprint*, arXiv:2206.04615.
- Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. *arXiv preprint arXiv:2305.08377*.
- Mildred C Templin. 1957. Certain language skills in children; their development and interrelationships.
- Ingilizce Test. 2024. Elementary vocabulary test. https://ingilizcetest.weebly.com/uploads/6/1/3/4/61346255/elementary_vocabulary_tests_and_answer_key.pdf. Accessed from google: 2024-10-11.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2020. **Superglue: A stickier benchmark for general-purpose language understanding systems**. *Preprint*, arXiv:1905.00537.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. **Glue: A multi-task benchmark and analysis platform for natural language understanding**. *Preprint*, arXiv:1804.07461.
- Pengda Wang, Hwayeon Myeong, and Frederick L Oswald. 2024a. On putting the horse (raters and criteria) before the cart (variance components in ratings). *Industrial and Organizational Psychology*, 17(3):309–313.
- Pengda Wang, Zilin Xiao, Hanjie Chen, and Frederick L Oswald. 2024b. Will the real linda please stand up... to large language models? examining the representativeness heuristic in llms. *arXiv preprint arXiv:2404.01461*.
- Pengda Wang, Huiqi Zou, Zihan Yan, Feng Guo, Tianjun Sun, Ziang Xiao, and Bo Zhang. 2024c. Not yet: Large language models cannot replace human respondents for psychometric research.
- Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2023. **Blimp: The benchmark of linguistic minimal pairs for english**. *Preprint*, arXiv:1912.00582.
- Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.
- Albert Frank Watts. 1944. The language and mental development of children.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. **Emergent abilities of large language models**. *Preprint*, arXiv:2206.07682.

- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.
- Ethan Gotlieb Wilcox, Michael Hu, Aaron Mueller, Tal Linzen, Alex Warstadt, Leshem Choshen, Chengxu Zhuang, Ryan Cotterell, and Adina Williams. 2024. Bigger is not always better: The importance of human-scale language modeling for psycholinguistics.
- Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. *Preprint*, arXiv:1704.05426.
- Terry Winograd. 1971. Procedures as a representation for data in a computer program for understanding natural language.
- Haotian Xia, Zhengbang Yang, Junbo Zou, Rhys Tracy, Yuqing Wang, Chi Lu, Christopher Lai, Yanjun He, Xun Shao, Zhuoqing Xie, Yuan-fang Wang, Weining Shen, and Hanjie Chen. 2024. Sportu: A comprehensive sports understanding benchmark for multimodal large language models. *arXiv preprint arXiv:2410.08474*.
- Guotong Yu. 2024. Gre sample writing. <https://github.com/yugt/GRE-Sample-Writing>. Accessed: 2024-10-02.
- Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. *Preprint*, arXiv:2002.04326.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *Preprint*, arXiv:2205.01068.

A Appendix A: Case Study

An example question in one-word-understanding that T5 made a mistake

Model choice: wait
Correct choice: rush

You don't have to _____! We're not late!

- A) dream
- B) laugh
- C) rush
- D) wait

We also investigate questions that RoBERTa and T5 answered incorrectly in the one-word understanding task, which all other models, including decoder-only and encoder-only models, answered correctly. After a thorough inspection of the testing examples that RoBERTa and T5 did not answer correctly, we identified two common points: (1) The models tend to choose answers that form more frequent collocations. For example, the models prefer “think about” over “complain about.” “Think about” can be used in a wider variety of contexts, including contemplation, consideration, and planning, whereas “complain about” has a negative connotation and is more context-specific. (2) Most of these questions require information from the surrounding context, either before or after the blank that needs to be filled in, which is similar to the findings of the case study in Wang et al. (2024b).

We carefully selected 50 examples from our training dataset on one-word understanding and tested RoBERTa-base and T5-large on these examples. All of the selected questions are composed of either those requiring context knowledge or those relying solely on collocation knowledge. To solve example A, the models must attend to the second sentence to understand that “not late” is related to “don’t have to rush,” rather than focusing solely on the first sentence.

RoBERTa RoBERTa-base answered 23 out of 50 examples correctly with an accuracy of 46%. Upon closer investigation, we found that, out of the 27 questions RoBERTa made mistakes on, 60% (16 questions) required context, while 40% (11 questions) were related to collocation.

T5 For the same set of examples, T5-large correctly answered 28 out of 50 examples, achieving

an accuracy of 56%. Of the 22 questions that T5 answered incorrectly, 16 (73%) required some contextual knowledge, while 6 (27%) involved collocations.

Since T5 performed relatively well compared to other models, we speculate that the way it handles multiple-choice questions contributes to its lower performance (see §5.1). As a result, we tested Flan-T5 (both large and 3b) on this task. We found that their performance, measured by normalized accuracy, increased to 0.807 (Flan-T5-l) and 0.898 (Flan-T5-xl).

Gemma2 Similarly, we tested instruction-fine-tuned versions of Gemma2 on the AAO task, where it underperforms. Their normalized accuracy rises to 0.87 and 0.85 for the 2b and 9b models, respectively, approaching the performance of other models. By fine-tuning on a wider variety of datasets, it enables generalization across a range of tasks.

B Appendix B: Data Contamination

There has been an increasing concern in data contamination nowadays (Deng et al., 2024). In this section, we investigate whether the pretraining data contain any datasets used in our evaluation. We apply the MIN-K% Prob method (Shi et al., 2024). This method selects the top k% of tokens with the highest negative log-likelihood and then computes the average log-likelihood. It is based on the hypothesis that an unseen example is likely to contain a few outlier words with low probabilities under the LMs, whereas a seen example is less likely to have words with such low probabilities. We follow the same settings as in that research and choose $k = 20$. If the number of tokens is between zero and one after multiplying the token length by 20%, we round it up to one.

In the following paragraph, we list the selection methodology:

one-word-understanding: We selected all instances of our test datasets and included sentences containing the correct answers.

AAO: We selected all examples from the test set, including both sentence_good and sentence_bad.

bc-if-why: We included all instances in the test datasets, incorporating both the premise and the hypothesis.

grammar-comp: In the test data, we randomly selected 1,000 examples and kept all other settings the same as in bc-if-why.

BLiMP-comp: For each grammatical phenomenon, we selected 50 examples, resulting in 2,800 instances. All other settings were the same as in AAO.

CoLA: All of the test examples were selected.

grammar-diag: We included all of the examples in the test datasets. The settings were the same as in bc-if-why.

WiC: Both sentences, one and two, were included.

ReClor: We tested the “context” part in each question. For this question, we tested the instructional fine-tuned and the chat version of the models.

Across each task, we presented the average MIN-K% probability for all individual sentences. For encoder-only models, we adapted this method by calculating the logits after masking each token in every sentence. To measure the relative MIN-K% probability, we randomly generated a sequence of all alphabets with a length of 10.

Overall, all models demonstrated comparatively low probabilities. We found that, in most datasets, the models are within 5% of the probabilities from random letters. However, gemma2-2b slightly exceeds 5% in the AAO dataset, which we consider acceptable (see Table 1).

C Appendix C: Implementation Details and Metrics

C.1 Implementation Details

Classification For BERT-style encoder models (Devlin et al., 2019), a special token, [CLS], is used as input to an MLP for prediction. In decoder models such as GPT-2 (Radford et al., 2019), the hidden state of the last token is connected to a classification head. For T5 (Raffel et al., 2020), with an encoder-decoder architecture, we use only the encoder to make predictions. Since an MLP is concatenated to each model, fine-tuning is necessary for the models to perform classification. Otherwise, the results will be random guesses. We fine-tune the models on grammar-comp for 1 epoch due to the large amount of data, and other classification tasks for 20 epochs maximum using four NVIDIA A-6000 GPUs. The learning rates we used range from 1e-6 to 1e-4, depending on model sizes and data sizes. Training batch sizes range from 1 to 16, given different parameter sizes. We also use LoRA (Hu et al., 2021) for models with large parameter sizes (Llama2-7b, Llama3-8b, Mistral-7b,

Gemma2-9b) due to the limitations of computational resources.

Minimal Pair and Vocabulary For decoder models, the average loss of the sequence is computed to determine which sentence is better. For BERT-style models, Masked Language Modeling is used to make predictions. For minimal pair questions (AAO and BLiMP-comp), special masks (e.g., <MASK>) are placed at the positions where the two sentences differ. Of the masked words, we select the one with a larger probability among the prediction of the masked positions. Similarly, for one-word understanding, we masked the blanks in the sentence. Then we choose one of the four words/phrases with the largest probability. T5, which is very similar to BERT-style models, uses Span Predictions. We compare the probability of the words it predicts between the span: <extra_id_0> word(s) predicted <extra_id_1>.

Generation Configuration The number of tokens generated by the LMs is set between a minimum of 500 and a maximum of 600 to ensure meaningful and comparable results across all chosen models. We keep the default generation parameters for all models, with two exceptions: Flan-T5 (Chung et al., 2022) and OPT-IML (Iyer et al., 2023) tend to generate repetitive sentences, so we relax their sampling criteria and apply top-k sampling with a probability of 0.9.

Other For filtering examples from datasets, we use the nltk (Bird et al., 2009) and spaCy(Honnibal et al., 2020) packages in Python.

C.2 Matthews Correlation Coefficient Formulation:

MCC =

$$\frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (1)$$

where:

- FP: False Positive
- FN: False Negative
- TP: True Positive
- TN: True Negative

Models	AAO	one-word	bc-if-why	grammar-comp	BLiMP-comp	CoLA	grammar-diag	WiC	ReClor	Random letters
opt-1.3b	12.75	10.18	9.13	9.42	12.51	10.37	9.11	10.41	10.30	10.29
opt-2.7b	12.8	10.17	9.16	9.43	12.54	10.38	9.05	10.39	/	10.22
T5-large	12.75	10.18	9.13	9.42	12.78	10.37	9.11	10.41	0.73	4.88
T5-3b	12.75	10.18	9.13	9.42	13.27	10.37	9.11	10.41	0.62	5.00
gpt2-large	12.66	10.55	8.91	9.15	12.54	10.04	9.02	9.73	/	9.87
gpt2-xl	12.67	10.47	8.86	9.13	12.38	10.04	8.99	9.70	/	9.84
Llama2-7b	11.58	9.24	8.96	8.87	11.29	9.63	8.36	9.89	8.31	9.87
Llama3-8b	13.13	10.35	9.80	9.70	12.69	10.58	9.03	10.85	11.00	11.00
Mistral-7b	12.16	9.80	9.64	9.42	12.14	10.18	8.72	11.27	7.08	10.12
gemma-2-2b	20.22	14.06	13.25	13.52	19.60	15.22	12.62	16.26	8.62	15.54
gemma-2-9b	22.14	14.82	13.85	14.03	21.50	16.12	12.94	16.63	9.11	17.11
ALBERT-xlarge	11.62	8.72	7.46	7.65	11.03	8.13	7.08	8.27	/	11.19
ALBERT-xxlarge	12.65	8.72	7.46	7.65	12.07	8.13	7.08	8.27	/	11.17
RoBERTa-base	12.87	9.29	7.27	7.10	11.92	7.91	5.82	7.99	/	9.89
RoBERTa-large	12.50	8.81	6.83	6.62	11.50	7.61	5.36	7.45	/	9.29

Table 1: MIN-K% Prob measured in %. Models measured in the ReClor task are the fine-tuned or chat version of that model.

D Appendix D: Interdisciplinary Collaboration

We would like to emphasize the importance of interdisciplinary collaboration. As LMs continue to evolve and mature, their potential applications across various fields are becoming increasingly evident. For example, they can be used in sports assessment (Xia et al., 2024), assist in questionnaire design in the social sciences (Wang et al., 2024c), answer clinical case questions (Chen et al., 2024), and even help with candidate screening (Wang et al., 2024a).

Interdisciplinary collaboration not only provides innovative technological solutions for various fields, but also brings unique insights from different disciplines into computer science, facilitating a better understanding of the underlying problems. For instance, collaboration between computer science, linguistics, and psycholinguistics offers new perspectives and methods, aiding in understanding the natural language processing capabilities of models from the viewpoint of language formation and development.

Such interdisciplinary collaboration transcends the limitations of individual disciplines, fostering the integration and innovation of knowledge, and enabling more complex and intelligent technological solutions. This trend presents new opportunities for future research and practice, driving societal progress.

E Appendix E: Tables and Graphs

Stage	Type	Data Split		Aspect
		Train	Test	
I	one-word	598	255	word-level
	AAO	-	1k	preliminary common sense
	bc-if-why	1.4k	348	causality conditionality
II	grammar-comp	170k	19k	grammar
	CoLA	6.8k	1.7k	
	grammar-diag	-	645	
	BLiMP-comp	-	56k	
III	WiC	5.4k	1.4k	word meaning under context
	ReClor	4.6k	1k	logical reasoning
	generation	-	10	logical composition

Table 2: Tasks from different stages. The Aspect column lists different language aspects tested. AAO = agent-action-object; one-word = one-word understanding dataset.

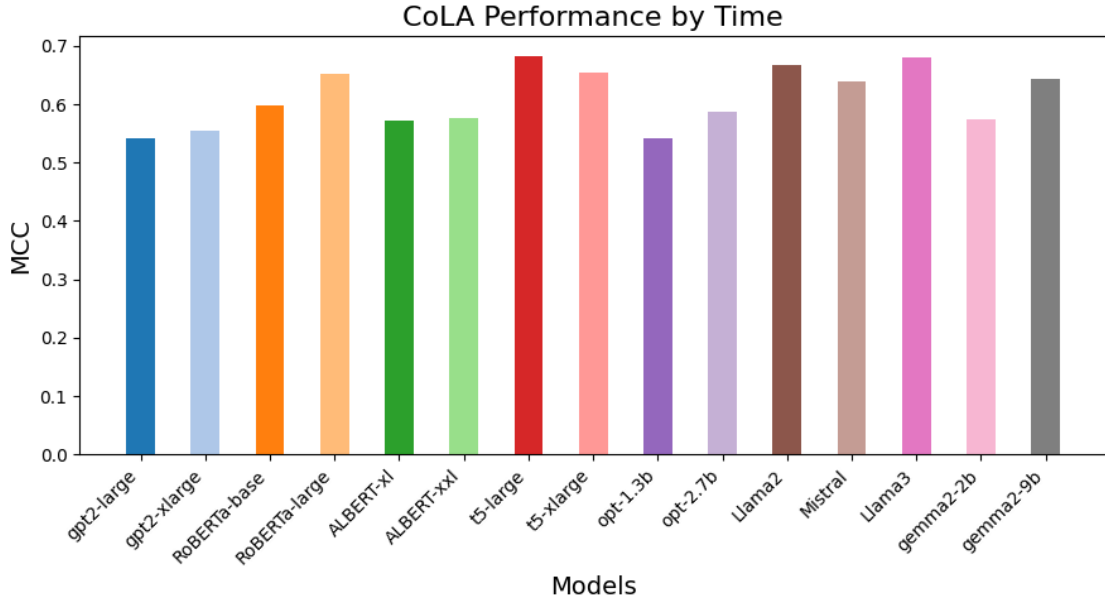


Figure 4: CoLA performance in Stage II measured in Matthews Correlation Coefficient (C.2). The result is obtained by training models at most 20 epochs

Models	Raw Accuracy	1-shot ICL	0-shot CoT
opt-iml-1.3b	0.31	0.32 +0.06	0.32 +0.06
Flan-t5-l	0.42	0.38 -0.05	0.42 +0.00
Flan-t5-xl	0.55	0.55 +0.00	0.54 -0.00
Gemma2-2b-it	0.49	0.46 -0.03	0.49 +0.00
Gemma2-9b-it	0.72	0.76 +0.04	0.71 -0.01
Llama2-7b-chat	0.37	0.36 -0.01	0.36 -0.01
Llama3-8b-chat	0.58	0.56 -0.03	0.43 -0.15
Mistral-7b-it	0.55	0.55 +0.00	0.53 -0.02

Table 3: Model Performance with raw accuracy on ReClor Dataset with 1-shot ICL and 0-shot CoT.

Grammar Phenomena	RoBERTa-base	T5-l	Gemma2-9b	Human
passive_2	0.60	0.87	0.75	0.86
determiner_noun_agreement_with_adj_irregular_1	0.50	0.83	0.89	0.94
superlative_quantifiers_2	0.89	0.76	0.71	0.85
wh_questions_subject_gap_long_distance	0.72	0.90	0.80	0.85
superlative_quantifiers_1	0.42	1.00	0.71	0.94
causative	0.72	0.78	0.65	0.98

Table 4: Selected results from BLiMP-comp of detailed grammar phenomena. We could notice the discrepancy in performance among the three models in these tasks, while humans could maintain high performance relatively.

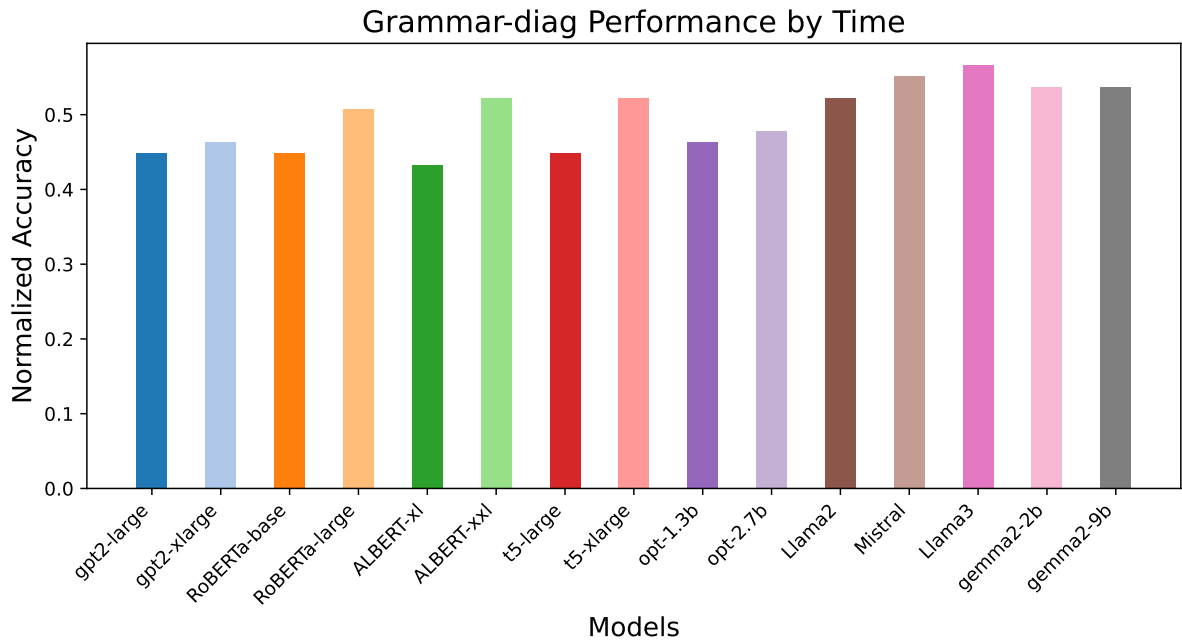


Figure 5: Grammar-diag performance in Stage II. Models are ordered by time. We test on models after fine-tuning on bc-if-why and grammar-comp’s training set.

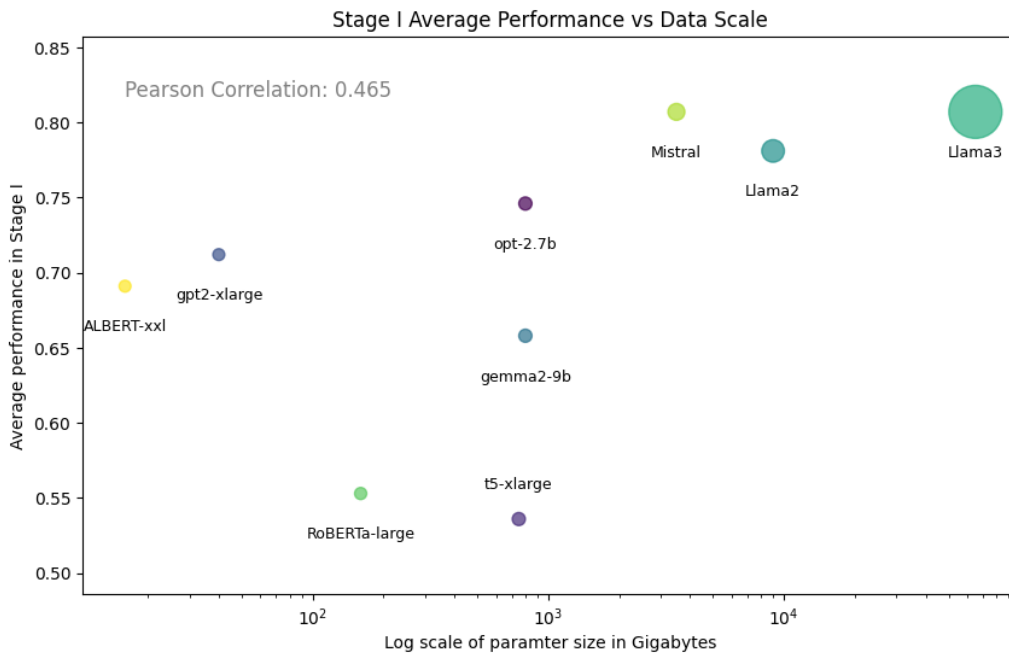


Figure 6: Stage I performance (normalized) vs. their data scale in the logarithm of Gigabyte.

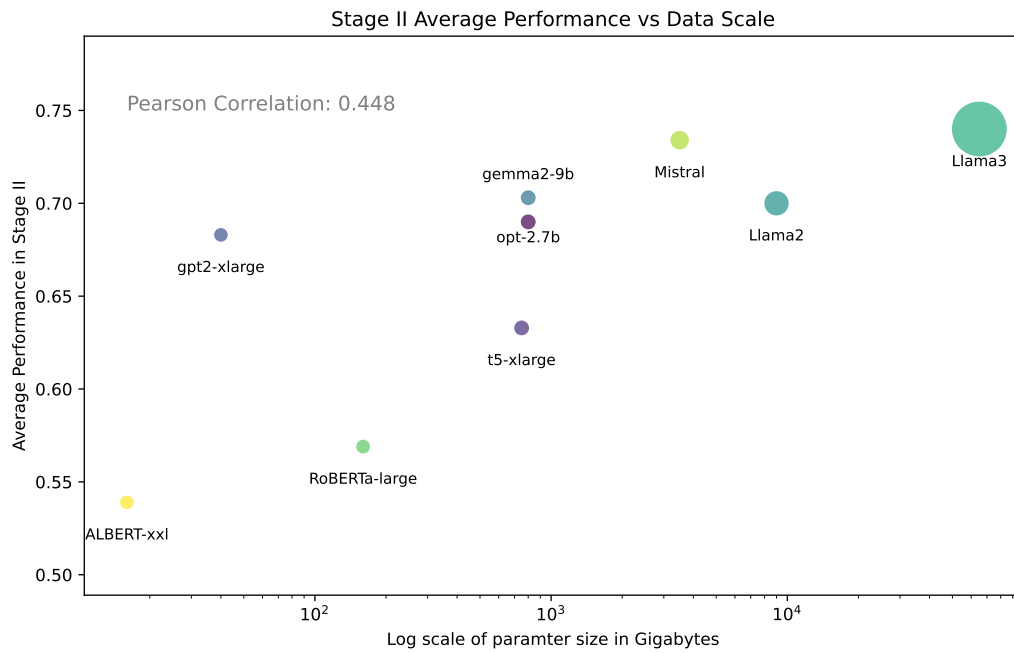


Figure 7: Stage II performance (normalized) vs. their data scale in the logarithm of Gigabyte.

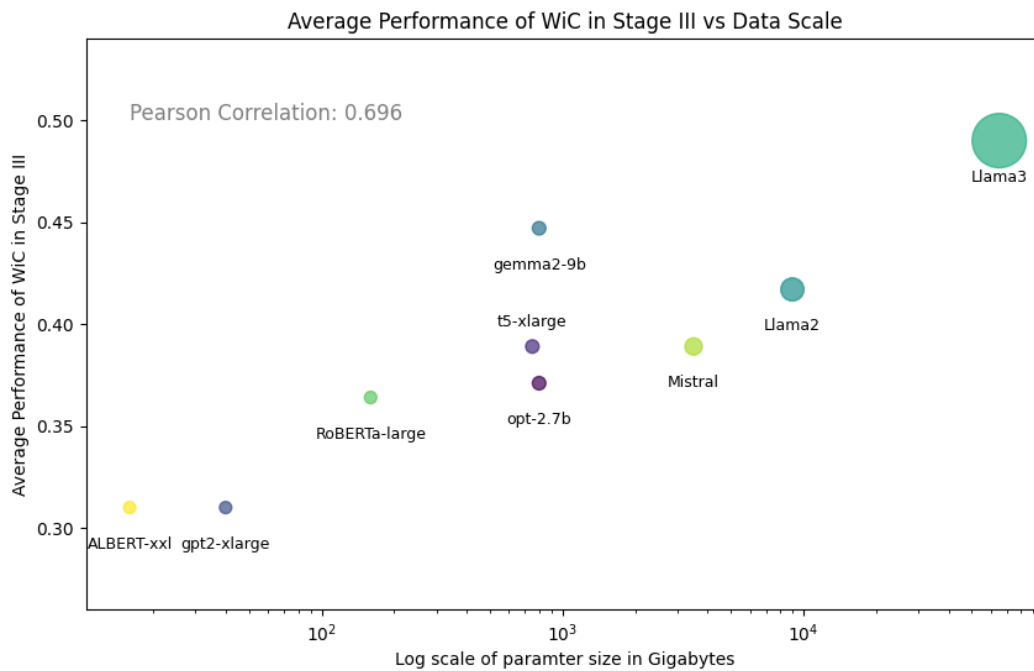


Figure 8: WiC in Stage III performance (normalized) vs. their data scale in the logarithm of Gigabyte.

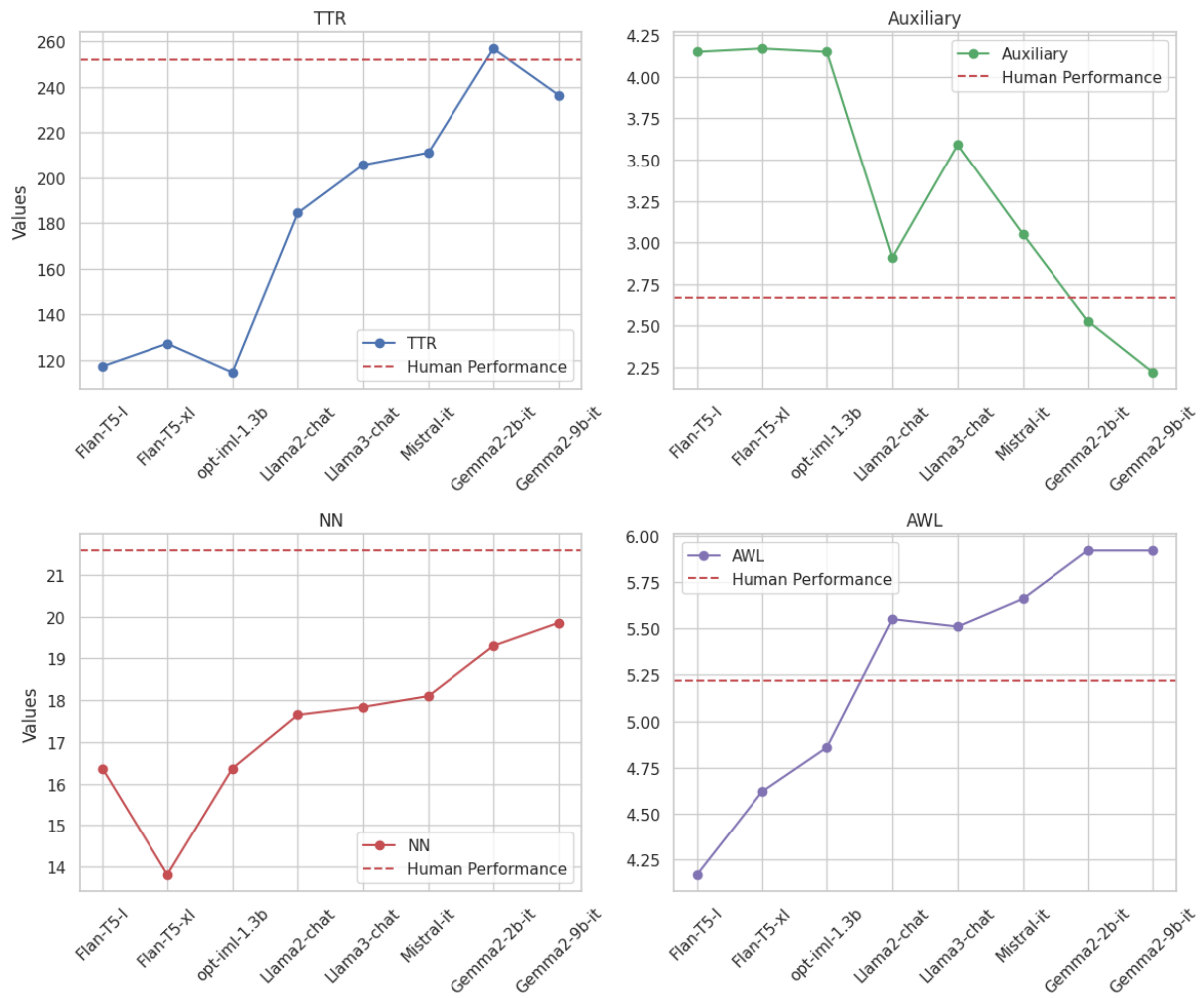


Figure 9: Four types of grammar metrics. Models are ordered by time

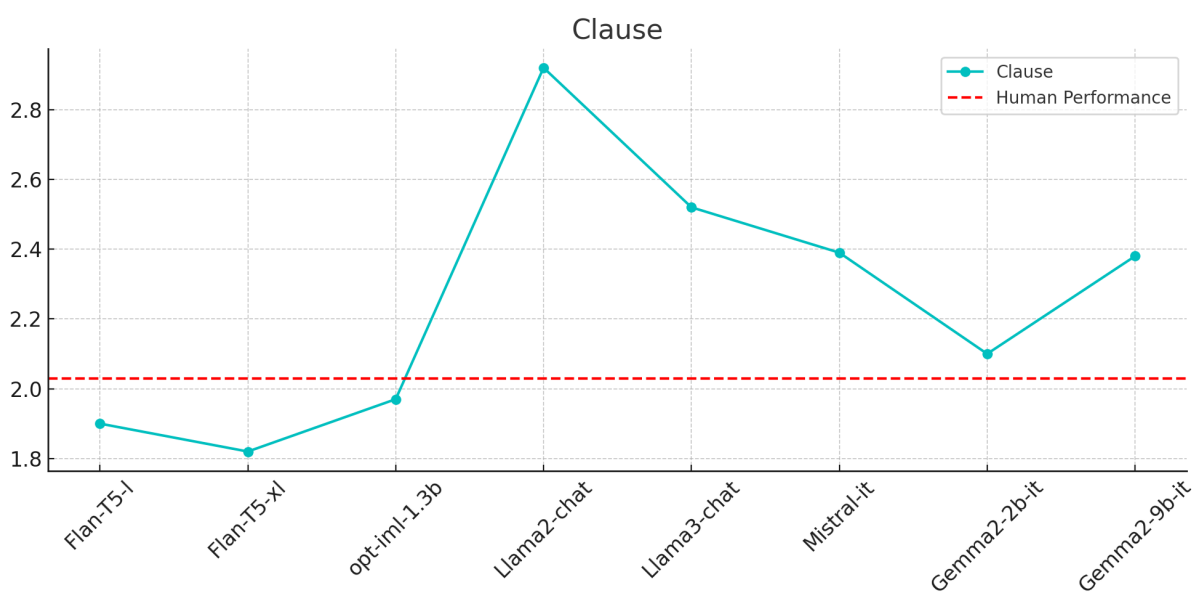


Figure 10: Clause grammar metrics. Models are ordered by time

Examples of each task

one-word understanding

Question: When you say something to someone's ear quietly and secretly, you _____.

- A) repeat
- B) whisper
- C) discuss
- D) cry

Correct Answer: B

Agent-Action-Object (AAO)

sentence_good: Tanya conceals Adam.

sentence_bad: This ice cream conceals Adam.

bc-if-why

Premise: If we keep up, they'll route.

Hypothesis: They'll route if we keep up.

Label: Entailment

grammar-comp

Premise: For Master P, neither is an appealing prospect.

Hypothesis: Master P found both projects to be appealing.

Label: Contradiction

CoLA

sentence: The in loved peanut butter cookies.

Label: 0 (False)

BLiMP-comp: determiner_noun_agreement_adj_2

sentence_good: Cynthia scans these hard books.

sentence_bad: Cynthia scans this hard books.

WiC

word: carry

sentence1: You must carry your camping gear.

sentence2: Sound carries well over water.

Label: F (False)

ReClor

Context: In a business whose owners and employees all belong to one family, the employees can be paid exceptionally low wages. Hence, general operating expenses are much lower than they would be for other business ventures, making profits higher. So a family business is a family's surest road to financial prosperity.

Question: The reasoning in the argument is flawed because the argument

- A) ignores the fact that in a family business, paying family members low wages may itself reduce the family's prosperity
- B) presumes, without providing justification, that family members are willing to work for low wages in a family business because they believe that doing so promotes the family's prosperity
- C) ignores the fact that businesses that achieve high levels of customer satisfaction are often profitable even if they pay high wages
- D) presumes, without providing justification, that only businesses with low general operating expenses can succeed

Answer: A

Table 5: One example from each dataset.