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

Anonymous ACL submission

Abstract

We examine the 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 generally outperform earlier models in overall performance, with some variations due to factors such as model architecture and training objectives, their developmental trajectory does not strictly follow the path of human language acquisition. Models show robust improvement in basic and intermediate tasks during pretraining, yet advanced tasks yield minimal gains, highlighting persistent challenges in higher-order linguistic processing. Notably, in generation tasks, experiments show that linguistic features in the training data shape model performance through context-dependent dimensions analogous to those observed in human language.

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 language models (LMs) to emulate human language. Over the past decades, we have witnessed the rise of LMs, which have achieved unprecedented success in language understanding and language generation (e.g., Gemini, Anil et al., 2023; GPT-4, Achiam et al., 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 (Wei et al., 2022a).

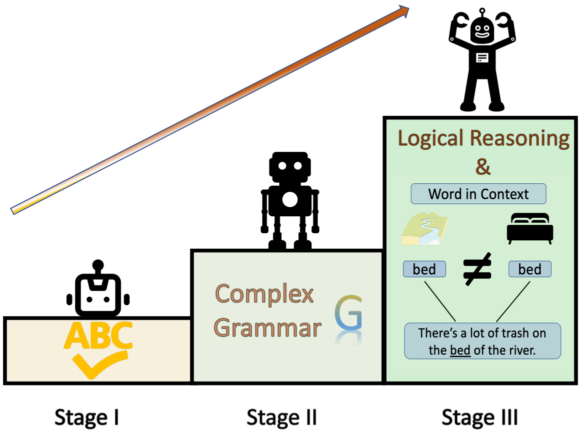


Figure 1: Three-Stage Anatomy of Language Acquisition.

As the capabilities of LMs continue to grow, so does the need for comprehensive evaluations of their performance. 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 understanding of model capabilities in terms of the developmental stages of human language acquisition (Goldberg, 2005)—the focus of this paper. 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. Therefore, 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 pro-

[1] Code and dataset are available at `https://anonymous.4open.science/r/Language-Acquisition-C8F7/README.md`

cess 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 apply our understanding of the human language acquisition process to design and evaluate theory-driven tests of the capabilities of LMs, this 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 and contexts, referred to as *registers*. In LMs, when conditioned on certain tasks, they will reflect some registers but not others, as the task-specific cues selectively activate subsets of linguistic patterns learned from the training data, leading us to examine how LMs have evolved in their register usage over time.

We evaluated 16 LMs released between 2019 and 2024, excluding instruction fine-tuned or chat versions, with varying parameter sizes (see §4.1). Our findings include: (1) LMs learn from vast corpora like humans, but their development does not exactly mimic human language acquisition stages, and their training objective and architecture could be factors that caused the variations; (2) Analysis of model checkpoints shows a steady improvement in model performance with training steps, though more advanced tasks remain challenging; (3) Models also share the context-dependent nature of linguistic feature distribution to some extent.

2 Related Works

Large pre-trained LMs, such as GPT (Radford et al., 2019) and BERT (Devlin et al., 2019), have revolutionized NLP by leveraging vast amounts of data and computational power to capture intricate nuances in language and enhance generative capabilities. After pre-training, these models are fine-tuned for specific tasks, and systematic benchmarking is important to standardize comparisons (Srivastava et al., 2023), highlight areas for improvement, and guide future advancements as models grow in complexity and diversity.

Classical Evaluations. 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 benchmarks 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.

Cognitive and Linguistic Evolution of LMs. Several studies have been conducted to investigate LMs' capabilities of learning language and their developmental abilities. For example, Kallini et al. (2024) evaluated GPT-2 on synthetic variations of impossible languages through systematic alterations of English, revealing that GPT-2 exhibited significant learning difficulties with these impossible languages compared to natural ones, challenging Chomsky's assertions about LMs' universal learning capabilities. Shah et al. (2024) investigated developmental trajectories in pretrained LMs by assessing cognitive abilities across training using standardized metrics in four domains, finding a consistent developmental window where models maximally align with human cognitive patterns. Besides, Li et al. (2024) demonstrated that LMs exhibit human-like patterns in resolving temporary ambiguities, particularly when structural cues such as commas facilitate disambiguation, suggesting fundamental similarities in linguistic processing mechanisms between artificial and human language systems.

While classical benchmarks and recent investigations into the cognitive evolution of LMs provide valuable measures of performance, they overlook a critical perspective: how these models mirror the gradual, stage-based progression observed in human language acquisition. In contrast to evaluating isolated tasks, 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 explore 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 coupled with a human's growth, 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 of refinement 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. For example, 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.[2] For an overview of the datasets, please refer to Table 3 in Appendix F and see Table 6 in Appendix F 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, which presumably to be obtained in early stage.

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, categorizing it into 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 16 LMs [3] in total over a broad time period (2019 to 2024) and with varying model parameter sizes.

[2] Note that we filter the training dataset and restrict the average C-Unit in datasets from the first two stages. 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.

[3] Note that the count of 16 excludes the fine-tuned or chat versions used in the ReClor and generation tasks, as they are the same type and size as the base models.

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`), Gemma2 (`gemma-2-2b`, `gemma-2-9b`) and the intermediate checkpoints of Pythia (Biderman et al., 2023).

4.2 Testing Methods

We use four different strategies to test the performance of LMs based on the specific tasks and model architectures.

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 E 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 16 topics in four different categories, taken from GRE public issue writing prompts (Educational Testing Service). Sample essays written by human testees with high scores (6 and 5) 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 evaluation metric as most of our testing datasets are balanced. CoLA dataset (Warstadt et al., 2018) also uses the Matthews correlation coefficient (see E.1).

Normalized Accuracy: While 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 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). Then we further explored the developmental trend of Pythia during pretraining (§5.2). Finally, we conducted a comprehensive and in-depth evaluation of the models' generative abilities from a linguistic perspective (§5.3).

5.1 Overall Trends in LMs' Development

Here, we focused on the overall development trends of LMs, and whether these models mimic the developmental 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.

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 displays our overall results. In Stage I, we first tackled fundamental tasks of human language acquisition, such as understanding individual

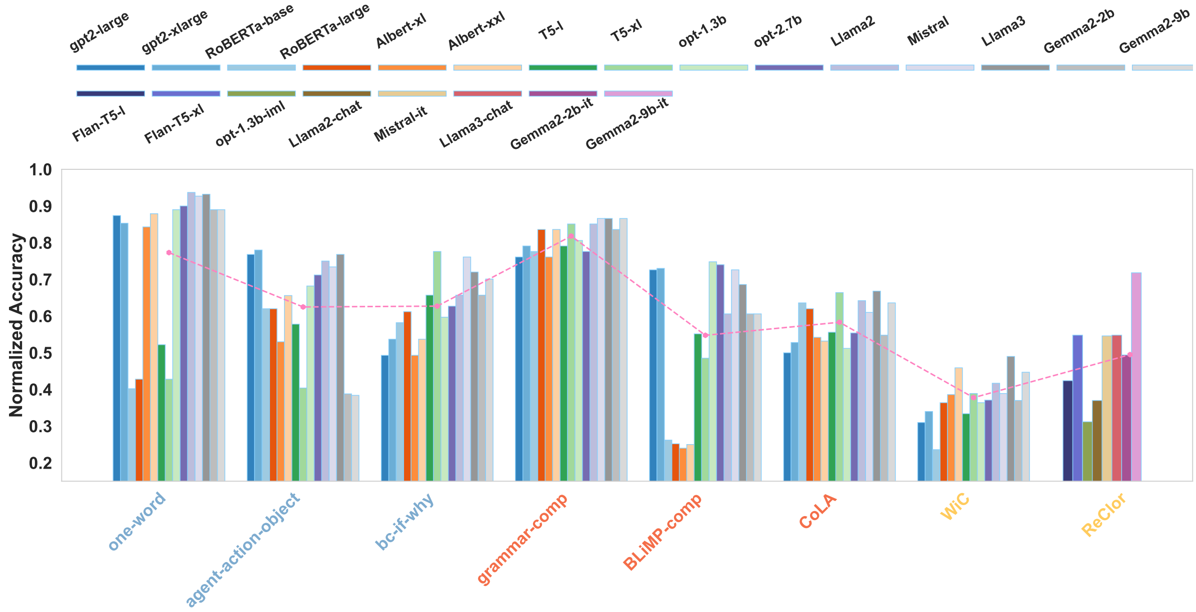


Figure 2: Performance of LMs across three stages. Colors represent stages arranged from left to right: Stage I –> Stage II –> Stage III. The upper legends correspond to models tested in tasks. For each task, models are ordered by their time released, and the tie is broken by their parameter sizes. Results from CoLA also use a different metric; please refer to Figure 11 in Appendix. F. Performance differences appear larger due to normalization.

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.1) on these models, we discovered that T5 and RoBERTa did not perform well on questions that require contextual information. However, the fine-tuned versions of T5 excelled in one-word understanding and the AAO task.

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 5 in Appendix F 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 4 in Appendix F).

Observing the developmental trend also reveals several key architectural and scaling insights (see Appendix B for detailed descriptions) in the model's language acquisition. While increasing model parameters did not consistently improve performance across most language development stages (with ReClor in Stage III being a notable exception), we found that encoder models frequently matched or surpassed larger decoder models in classification tasks, likely due to their bidirectional attention capabilities. Interestingly, for sentence-pair comparison tasks (AAO and one-word), decoder-only models generally outperformed their encoder-only or encoder-decoder models, potentially due to differences in training objectives (e.g., masked language modeling vs. next-token prediction) and the absence of sentence order prediction objective in some models (RoBERTa and T5). These findings suggest that architectural choices and training objectives may be more crucial than model size for specific linguistic tasks to empower the model to learn from the training corpus more effectively. This also indicates that insights from linguistic research can contribute to future improvements, alongside scaling up model parameters and data sizes.

5.2 Language Development in Pretraining

In addition to the investigation of LMs' development as a whole, we also examined the LMs' development during pretraining. Here we selected `Pythia-1B` (Biderman et al., 2023) and tested checkpoints at {5000, 28000, 56000, 84000, 112000, 143000} steps respectively. The experimental results (shown in Figure 3) yield two primary insights: (1) As the training steps increase, the model performance tends to increase. (2) Generally, the model performs better in early-stage tasks than in later-stage tasks—except for the "bc-if-why" task—and Pythia exhibits greater initial gains (or learnability) in earlier-stage tasks.

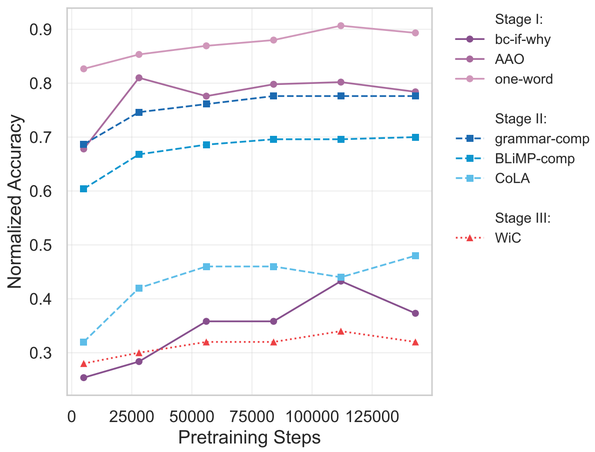


Figure 3: Performance of Pythia 1B with different checkpoints.

Trends in Training Steps. Model performance demonstrates consistent improvement across training steps, analogous to human language acquisition patterns. However, there is a slight dropback between the last two checkpoints in some tasks, which usually appears during training (Shen et al., 2024; Luo et al., 2024). This mirrors patterns in human skill acquisition where progress stabilizes despite continued practice (Vleugels et al., 2020). Significantly, models achieve near-optimal performance after approximately 50000 training steps, with marginal subsequent improvements.

Trends between Stages. Pythia in Stage I tasks demonstrates robust overall performance, with "one-word" tasks achieving high performance, while "bc-if-why" tasks show more modest but also consistent improvement throughout training. Stage II evaluations exhibit progressive enhancement during pretraining, with CoLA demonstrating a slight inital gain but particularly notable developmental trajectories. In contrast, Stage III task WIC consistently yields the low performance metrics with minimal improvement across training iterations, suggesting a persistent challenge in higher-order linguistic processing.

5.3 Generation Ability and Register Theory

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.

Biber's Tagger. Register theory posits that linguistic features—such as vocabulary, syntax, and formality—vary systematically with context, audience, and communicative purpose (Halliday, 1977; Matthiessen, 1993; Biber and Conrad, 2009). Extensive research in linguistics has explored co-occurrence patterns of these features across different contexts based on register theory. Drawing on the Multi Dimensional Analysis Tagger (MAT) by Nini (2019), which replicates the procedure established by Biber (1988), we compared five representative dimensions.

NN (nouns that are not identified as nominalizations or gerunds): This metric evaluates the model's accurate and flexible use of standard noun forms.

AWL (average word length): This metric measures the mean length of the words in the text in orthographic letters.

Clause (a collection of adjectival and adverbial clauses): This metric quantifies the frequency and diversity of clauses.

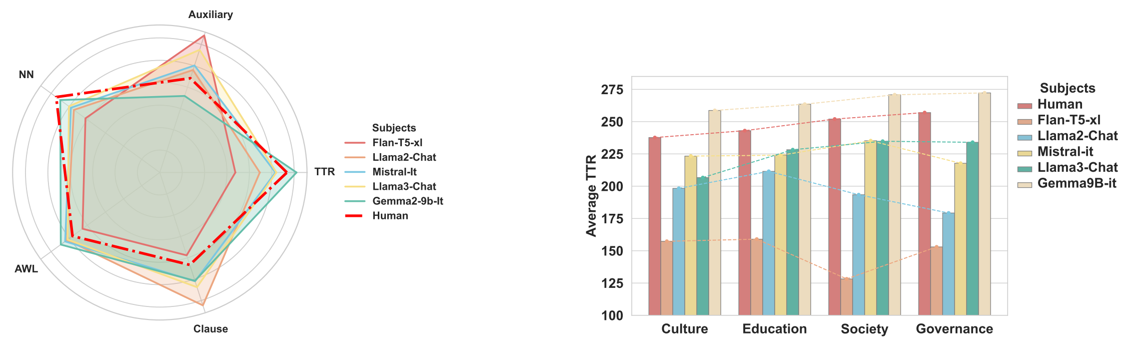
TTR (type-token ratio): This dimension evaluates the richness of the generated text in terms of lexical diversity.

Auxiliary Verbs (e.g., modal verbs expressing possibility, prediction, and necessity): This indicator tracks the usage of auxiliary verbs in the texts.

In all five dimensions, we found that patterns in NN, TTR, and AWL dimensions tend to be more similar to human, while more variations [4] are exhibited in other dimensions (see Figure 4(a)).

Linguistic Features reflect Register. To explore the inter-relationships between linguistic features and registers, we further divided the GRE issue

[4]These variations are discussed in Appedix B.



(a) Generation abilities across Five Dimensions

(b) Average TTR across Four Categories of Topics

Figure 4: Models are ordered by time. (a) shows the comparison of five dimensions of different linguistic features in the essays. (b) shows average TTR across four categories of topics. Average NN also shows certain trend, see Figure 17 in Appendix F.

writing prompts (mentioned in Section 4.2) into four distinct topics (culture, education, society, and governance) and calculated the average TTR for each category separately. As shown in Figure 4(b), the trends in average TTR across these topics converge more closely to human patterns in later models compared to earlier ones—Gemma9b-it even exhibits a higher overall TTR in every category than human data—which suggests that while recent models produce more lexically diverse outputs, they have concurrently evolved to capture the nuanced register-specific variations that characterize natural language.

To investigate how linguistic characteristics influence the variations in the registers, we employed sparse dictionary learning methods (Braun et al., 2024) through two complementary case studies. Our empirical investigation yielded two findings regarding the relationship between register variation and semantic processing:

Semantic Boundaries in AAO. Analysis of the AAO task performance demonstrates that models exhibiting lower accuracy tend to produce subject token representations with less distinct semantic boundaries. This observation aligns with register theory principles, where effective communication relies on maintaining clear semantic distinctions across different contexts (detailed analysis in A.1 of the Appendix).

Lexical Variation under Register Steering. Through targeted activation (see Appendix E.2 for implementation details) steering on context features from "Governance" to "Culture" (Figure 5), we observe systematic decreases in TTR measures. This finding provides empirical support that models also share the the context-dependent nature of linguistic feature distribution, demonstrating how register variations systematically influence lexical diversity patterns.

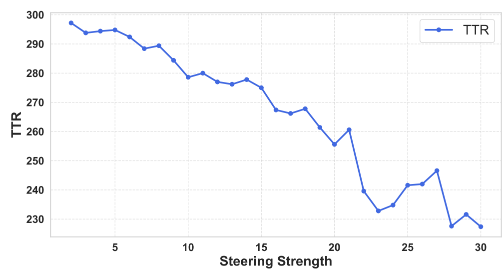


Figure 5: Steering activations of Gemma9b-it with prompt on "Governance" topic to "Culture". Steering strength is normalized by a factor of 100.

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 scale of the corpora they were trained on. Investigations of model checkpoints indicate that models show greater learning abilities in earlier-stage tasks than in later-stage tasks.

The study of register theory further shows that linguistic features of the training data influence the models' performance, demonstrating context-dependent linguistic feature dimensions similar to those observed in human language.

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, because 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.

Douglas Biber and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge University Press.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. *Preprint*, arXiv:2304.01373.

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

Dan Braun, Jordan Taylor, Nicholas Goldowsky-Dill, and Lee Sharkey. 2024. Identifying functionally important features with end-to-end sparse dictionary learning. *Preprint*, arXiv:2405.12241.

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.

J. E. Casal and J. J. Lee. 2019. Syntactic complexity and writing quality in assessed first-year l2 writing. *Journal of Second Language Writing*, 44:51–62.

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.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *Preprint*, arXiv:2309.08600.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, 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.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2024. Scaling and evaluating sparse autoencoders. *Preprint*, arXiv:2406.04093.

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.

Michael AK Halliday. 1977. Text as semantic choice in social contexts. *Grammars and descriptions*, 176225.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, Yu-Gang Jiang, and Xipeng Qiu. 2024. Llama scope: Extracting millions of features from llama-3.1-8b with sparse autoencoders. *Preprint*, arXiv:2410.20526.

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-iml: Scaling language model instruction meta learning through the lens of generalization. *Preprint*, arXiv:2212.12017.

T. Jagaiah, N. G. Olinghouse, and D. M. Kearns. 2020. Syntactic complexity measures: Variation by genre, grade-level, students' writing abilities, and writing quality. *Reading and Writing*, 33(10):2577–2638.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée 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.

Prathyusha Jwalapuram, Shafiq Joty, and Xiang Lin. 2022. Rethinking self-supervision objectives for generalizable coherence modeling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6044–6059, Dublin, Ireland. Association for Computational Linguistics.

Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. 2024. Mission: Impossible language models. *Preprint*, arXiv:2401.06416.

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.

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*.

Andrew Li, Xianle Feng, Siddhant Narang, Austin Peng, Tianle Cai, Raj Sanjay Shah, and Sashank Varma. 2024. Incremental comprehension of garden-path sentences by large language models: Semantic interpretation, syntactic re-analysis, and attention. *Preprint*, arXiv:2405.16042.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant

Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. *Preprint*, arXiv:2408.05147.

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.

Kairong Luo, Haodong Wen, Shengding Hu, Zhenbo Sun, Zhiyuan Liu, Maosong Sun, Kaifeng Lyu, and Wenguang Chen. 2024. A multi-power law for loss curve prediction across learning rate schedules. In *NeurIPS 2024 Workshop on Mathematics of Modern Machine Learning*.

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.

Raphaël Millière. 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.

Raj Sanjay Shah, Khushi Bhardwaj, and Sashank Varma. 2024. Development of cognitive intelligence in pre-trained language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9632–9657, Miami, Florida, USA. Association for Computational Linguistics.

Yikang Shen, Matthew Stallone, Mayank Mishra, Gaoyuan Zhang, Shawn Tan, Aditya Prasad, Adriana Meza Soria, David D. Cox, and Rameswar Panda. 2024. Power scheduler: A batch size and token number agnostic learning rate scheduler. *Preprint*, arXiv:2408.13359.

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.

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.

F. J. Tweedie and R. H. Baayen. 1998. How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities*, 32(5-6):323–352.

L. W. E. Vleugels, S. P. Swinnen, and R. M. Hardwick. 2020. Skill acquisition is enhanced by reducing trial-to-trial repetition. *Journal of Neurophysiology*, 123(4):1460–1471.

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, Zilin Xiao, Hanjie Chen, and Frederick L Oswald. 2024. Will the real linda please stand up... to large language models? examining the representativeness heuristic in llms. *arXiv preprint arXiv:2404.01461*.

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.

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

A.1 Case Study: Under-performance in one-word Understanding

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

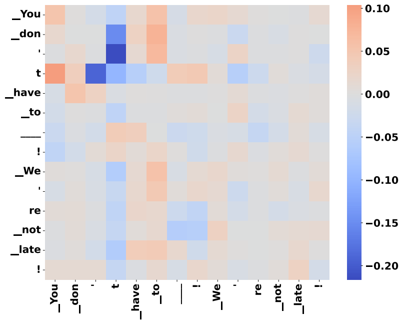
Model choice: wait
Correct choice: rush

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

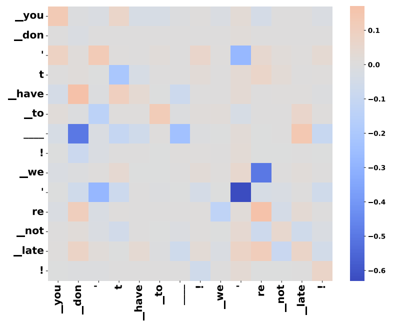
A) dream
B) laugh
C) rush
D) wait

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. (2024).

We carefully designed 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.1, 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. Not surprisingly, in Figure 6, ALBERT-xl which aimed this question paid more attention to the token "late" in the correct sentence compared to T5 which missed this question.



(a) T5-large



(b) ALBERT-xl

Figure 6: Max attention weight differences between sentence 1: "You don't have to rush..." and sentence 2: "You don't have to wait..." for T5 and ALBERT. The "____" token is "rush" in the first sentence and "wait" in the second sentence. We could find the "____" token in ALBERT is more related to the "late" token compared to T5.

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.

Because T5 performed relatively well compared to other models, we speculate that the way it handles multiple-choice questions contributes to its lower performance (see §B). 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).

Dictionary Learning Reflects Register Usage Sparse Autoencoder (SAE) is a powerful unsupervised dictionary learning method that learns a sparse decomposition of models' representa-

Sentences	Gemma2 9b	Gemma2 9b-it	Llama3.1-8b	✔
Good : Melissa will clean a gate. Bad : This mouth will clean a gate.	15675: terms related to oral health and hygiene	6714: anatomical terms related to body parts	6991: references to mouth	Llama 3.1 and Gemma2 9b-it
Good : Tanya admires Melanie. Bad : Music admires Melanie.	10440: mathematical symbols or notions	8262: instances of phrases "I don't" and its variations	16839: references to music and related media	Llama 3.1
Good : A senator drops by every lake. Bad : The muffin drops by every lake.	12754: descriptions of food and beverages , with emphasis on coffee and sweet treats	15081: the prefix "mu" in various forms to identify related biological or chemical substances	10179: mentions of muffins in various contexts	Gemma2 9b
Good : The committee disliked Lissa. Bad : The company disliked Lissa.	13164: references to companies and their details	11832: references to companies and their details	25617: mentions of companies or company-related concepts	None

Table 1: Sentences selected from the AAO tasks. Each entry in the middle three columns represents the feature of the subject token that has the highest activation in the bad sentence. The last column indicates which model(s) choose(s) the correct answer. Highlights in orange show the models activate on correct feature(s) when "making decisions".

tions into interpretable features (Cunningham et al., 2023). Register theory suggests that language varies systematically based on context and corpus. As a result, SAE offers a plausible way to investigate the models could activate on what type of context or which part of corpus (features) given the texts.

Here we use SAE to investigate the AAO tasks in which Gemma2 did not perform well. Because training and scaling of SAEs are computationally intensive and difficult (Gao et al., 2024), we used pretrained SAEs Gemma Scope (Lieberum et al., 2024) and Llama Scope (He et al., 2024). As mentioned in the paper, the data that used to train SAEs are sampled to be representative of the distribution of pretraining data, we could get a fairly well approximation to the pretraining corpus and connect to register theory.

We compared Gemma2-9b, Gemma2-9-it, and Llama3.1-8b on the last layer's residual stream by selecting several examples from the AAO tasks they did correctly or incorrectly. We find an interesting pattern: When the models did the problem correctly, the feature that subject tokens activate retains more precise semantic meanings compared to when they did it incorrectly. For example, at the first row of Table 1, we see that the subject token ("mouth") in the bad example activates on features that are more related (body parts and reference to mouth) for Gemma2-9b-it and Llama3.1-8b. For Gemma2 9b that missed this question, the feature "oral health and hygenie" encompassed more meaning of the later token such as "clean" within the token "mouth". Maintaining a more independent meaning from the context of bad examples is key to aiming this question.

Nonetheless, the last example in Table 1 presents an interesting exception — features from all three models are precise and did not interleave with later tokens. One plausible explanation is that the training corpus may include grammatically incorrect sentences. It does not impede our understanding of the sentence if we say "The company disliked Liss" even if it has mistakes in grammar.

Additionally, having learned that models encode blended semantic meanings in the subject token when they chose the bad sentence, we verified this observation by activation steering (refer to Appendix E.2 for the formulation of steering). By steering toward activations with more precise semantic meaning, the model is more likely to favor the correct answer if it had previously answered incorrectly (see Figure 7). This observation aligns naturally with register theory principles, where formal registers are characterized by clear semantic boundaries and controlled meaning structures. The models' successful performance appears to mirror this principle—maintaining distinct, register-appropriate semantic representations leads to correct responses, while semantic boundary violations, manifesting as blended activations, typically result in errors. Similarly, as shown in Figure 8, by steering Llama3.1-8b's activation of the second example towards the opposite direction of "Music", the model becomes less likely to favor the correct examples. This also confirms the robustness of the features across models and different pretrained SAEs.

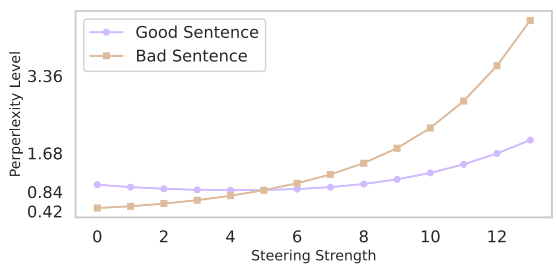


Figure 7: Good Sentence: Tanya admires Melanie. Bad Sentence: Music admires Melanie. Steering towards the direction of "Music". Steering Gemma2-9b towards the direction of "Music".

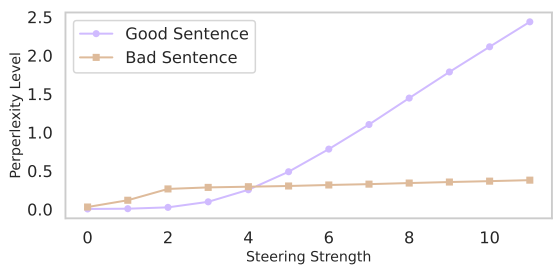


Figure 8: Good Sentence: Tanya admires Melanie. Bad Sentence: Music admires Melanie. Steering Llama3.1-8b towards the **opposite** direction of "Music". We could find that, before steering Llama3.1-8b could choose the correct sentence. However, after steering by a certain strength (around 4), the good sentence has larger perplexity.

B Appendix B: Factors in Models' Language Acquisitions

In this section, we discussed factors that could affect the performance of the models in our language acquisition task.

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 (larger than 20% accuracy) in performance between small models and their larger counterparts. However, this observation does not negate that on certain tasks, larger models could outperform by a certain amount as compared to their smaller counterparts. In fact for the complex task ReClor (in Stage III), 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, Sentence Order Prediction (SOP) might play an important role in one-word understanding tasks (see Appendix A.1 for a complementary case study). Even with larger batch sizes, models such as RoBERTa and T5, which are not trained on SOP, may lack the ability to model sentence-to-sentence transitions, which is essential for that task.

Do Data Size 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 13, 14, 15 in Appendix F). Please refer to the formula in E.3 of Appendix for how the data sizes are estimated.

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.

Variations in Auxiliary and Clause Dimensions in Generation Task. We observed that for some dimensions (Auxiliary and Clause), LMs have larger variations compared to humans. However, a good composition or essay does not necessarily contain more usage of complex features such as auxiliary features and clauses (Jagaiah et al., 2020; Casal and Lee, 2019). While this observation is interesting, it falls outside the scope of our paper, and further research on these topics is encouraged.

C Appendix C: Coherence of Generation

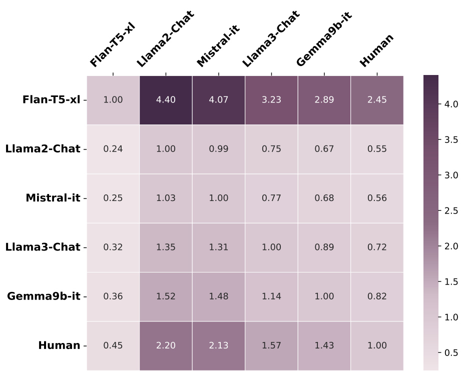


Figure 9: Coherence Comparison across different models.

We observed that, some generated texts from some models are somewhat repetitive in nature. This repetitive nature could cause larger measurements in some dimensions (e.g., clause) in the multidimensional co-occurrence analysis. As a result, we also measured the relative coherence of those essays generated by the models. The relative coherence score is calculated by dividing the target model's coherence score by the reference model's coherence score, where each score is obtained by the pretrained coherence model from Jwalapuram et al. (2022). We made some small modifications to the algorithm to handle negative scores. Please check Appendix E.3 for the details of the algorithm.

In Figure 9, the score in entry `(i, j)` is obtained by using the i-th model as the reference model and the j-th model as the subject model. We found that all models except Flan-T5-xl exhibit higher coherence when compared to humans. This is attributed to the repetitive nature of the texts generated by T5, as previously mentioned. Despite this supremacy, we observed a decreasing trend where the relative coherence of later models tends to decrease. This trend indicates an evolution toward more human-like behavior in the models, suggesting an increased capacity to learn and replicate text with enhanced accuracy and precision.

D Appendix D: 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 2).

E Appendix E: Implementation Details and Metrics

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. Because an MLP is concatenated to each model, fine-tuning is necessary for the models to perform classification.[5] 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 and choose the checkpoint with the lowest validation loss. 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 with rank 64 and lora_alpha 32 (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.

[5]We also compared this method by concatenating question prompts and allowing the model to predict the next token (answer). This approach resulted in at least 20% decrease in performance.

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.

Biber's Tagger and MAT To ensure methodological rigor in our analysis of Type-Token Ratio (TTR), we incorporated a fixed-sample approach with a standardized 600-token threshold. For texts exceeding 600 tokens, TTR is calculated based on the first 600 tokens. For texts with fewer than 600 tokens, the TTR is computed using the full text. This standardization effectively neutralizes the analytical distortions that typically emerge when comparing lexical diversity across texts of varying

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

lengths.

The methodological justification for this approach is grounded in well-established research on lexical statistics. As demonstrated by (Tweedie and Baayen, 1998), TTR values exhibit an inverse relationship with text length, primarily due to the inherent frequency patterns of common lexical items. Our implementation of a uniform 600-token analytical window thus addresses this fundamental methodological challenge, enabling more precise cross-corpora comparisons of lexical diversity.

Our approach is supported by the characteristics of our corpus. The texts have a relatively consistent length, with an average of 480 tokens. This natural consistency helps justify our fixed-sample approach, reducing potential bias in TTR calculations. Together, the standardized analysis and corpus properties provide a strong basis for assessing lexical diversity, mitigating the the bias in TTR measurement.

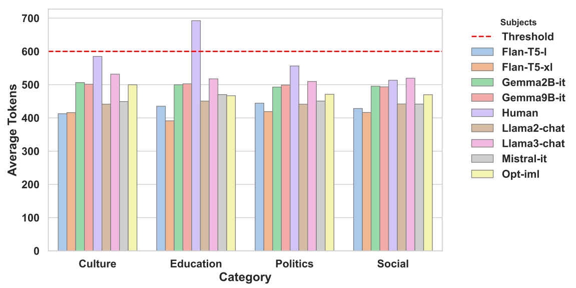


Figure 10: Average Token Length of Each Subject per Category

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

E.1 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

E.2 Activation Steering

Here we describe how we chose and conducted activation steering mentioned in Appendix A.1 and Section 5.3.

Specifically, for the AAO task, we first select the feature's activation that encodes more precise meaning of the subject token and steer the original activation by the following formula:

$$\gamma = \text{strength\_multiple} \times \text{steering\_strength},$$
$$\mathbf{v} = \text{sae.W\_dec}[\text{feature\_index}],$$
$$\mathbf{a} \leftarrow \mathbf{a} + \gamma\,\mathbf{v},$$

where feature_index corresponds to the index of the feature's top activation that encodes more precise meaning of the subject token, $\mathbf{a}$ corresponds to model's activation.

Similarly, for the generation tasks steering, we first selected a feature's activation of the context (e.g., feature that activated on "culture") we want to steer to, and steered the original activation with the above formulation.

E.3 Training Data Size Calculation

We assess the training data size based on either the total token size or the size of its corpus, depending on the information provided in technical reports. For the total token size, we approximate the corpus using the following formula:

$$\texttt{Corpus Size (GB)} = \frac{\mathsf{TT} \times \mathsf{ACT} \times \mathsf{BC}}{10^9} \tag{2}$$

where:

- TT: Total Tokens
- ACT: Average Characters per Token
- BC: Bytes per Characters
- GB: Gigabytes

E.4 Relative Coherence Score Calculation

Algorithm 1 Relative Coherence Score Calculation

Require: ref, sub ▷ Input reference and subject texts

$ref\_tensor \leftarrow$ Preprocessor$([ref])$
$sub\_tensor \leftarrow$ Preprocessor$([sub])$
$ref\_score \leftarrow$ coherenceScore$(ref_tensor$["tokenized_texts"]$)$
$sub\_score \leftarrow$ coherenceScore$(sub_tensor$["tokenized_texts"]$)$
if $sub\_score < 0$ **and** $sub\_score \neq ref\_score$ **then**
    return $\frac{-sub\_score}{ref\_score - sub\_score}$
else if $ref\_score < 0$ **and** $sub\_score \neq ref\_score$ **then**
    return $\frac{-ref\_score}{sub\_score - ref\_score}$
else
    return $\frac{sub\_score}{ref\_score}$
end if

F Appendix F: Tables and Graphs

Stage	Type	Data Split Train	Data Split Test	Aspect
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 3: Tasks from different stages. The Aspect column lists different language aspects tested. AAO = agent-action-object; one-word = one-word understanding dataset.

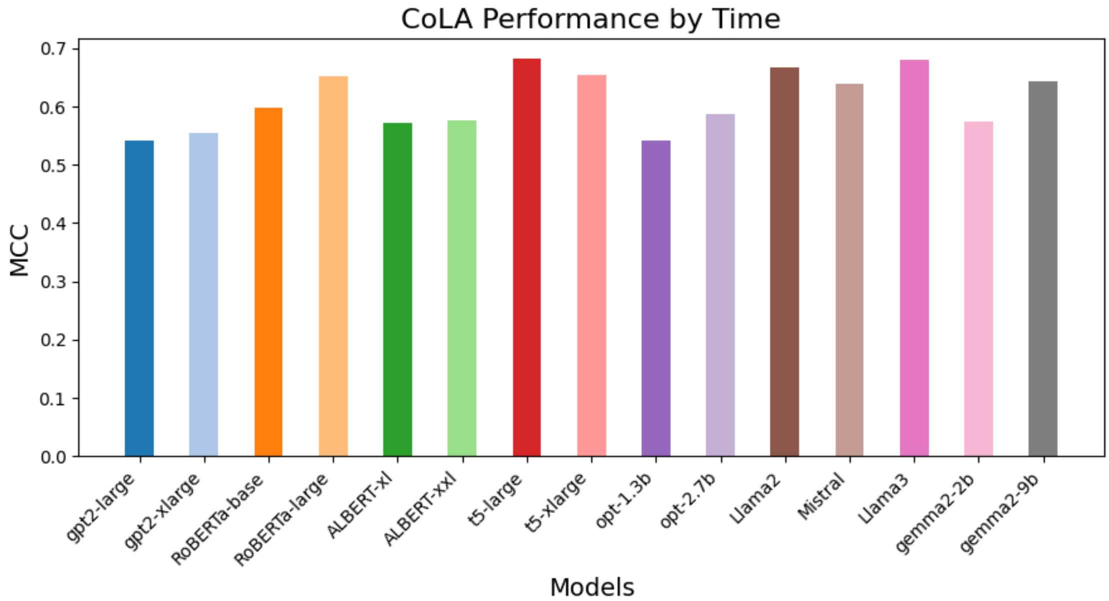


Figure 11: CoLA performance in Stage II measured in Matthews Correlation Coefficient (E.1). The result is obtained by training models at most 20 epochs

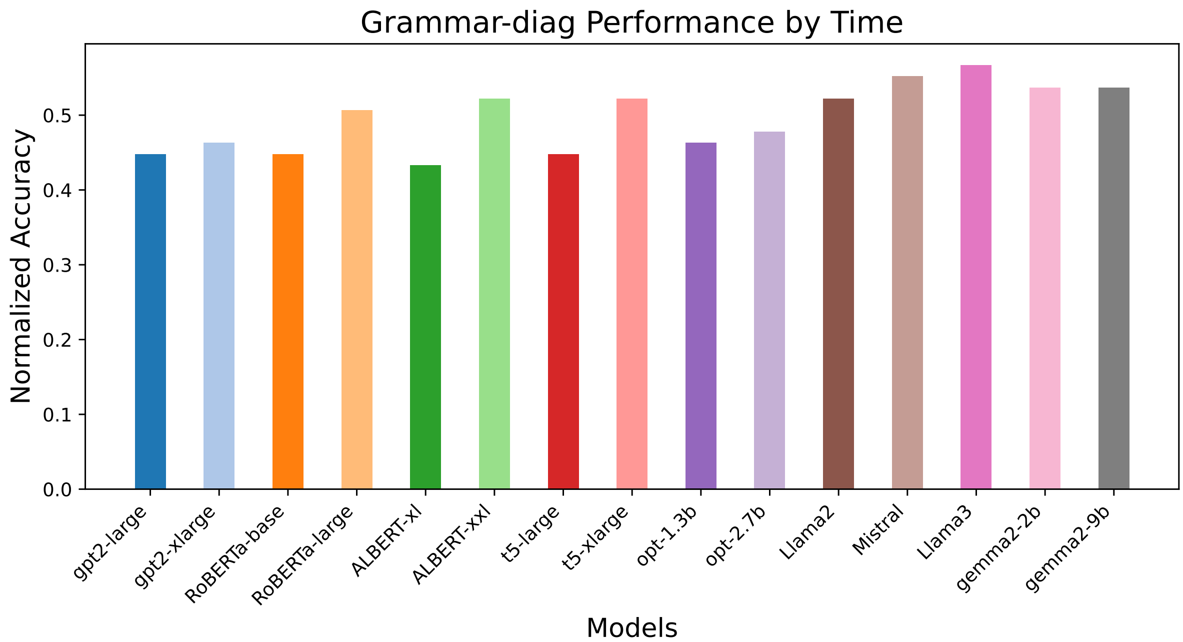


Figure 12: Grammar-diag performance in Stage II. Models are ordered by time. We test on models after fine-tuning on bc-if-why and grammar-comp's training set.

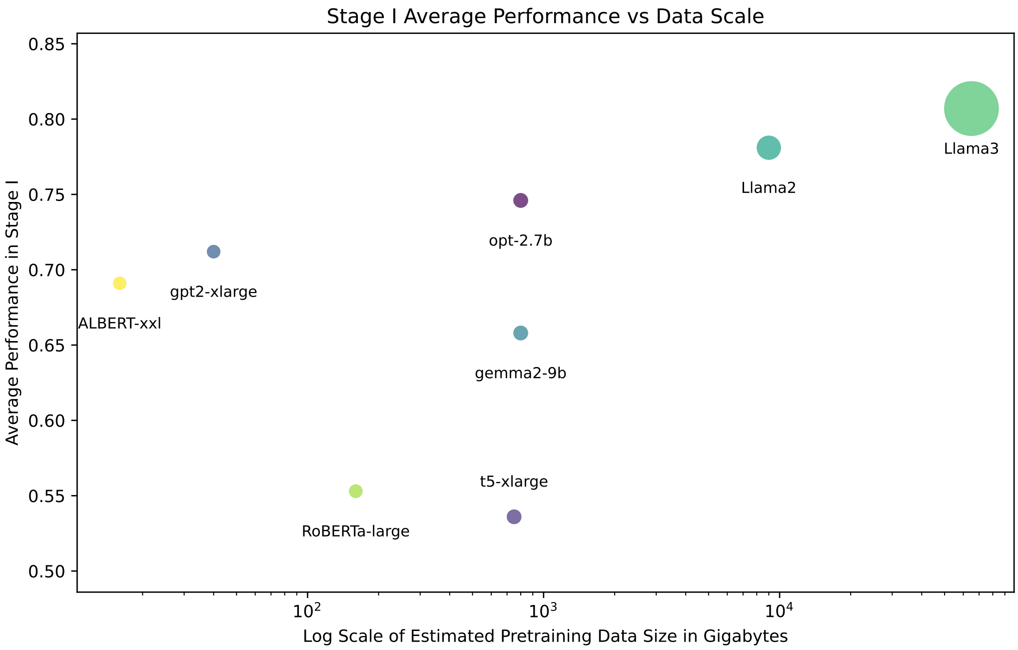


Figure 13: Stage I performance (normalized) vs. their data scale in the logarithm of Gigabyte.

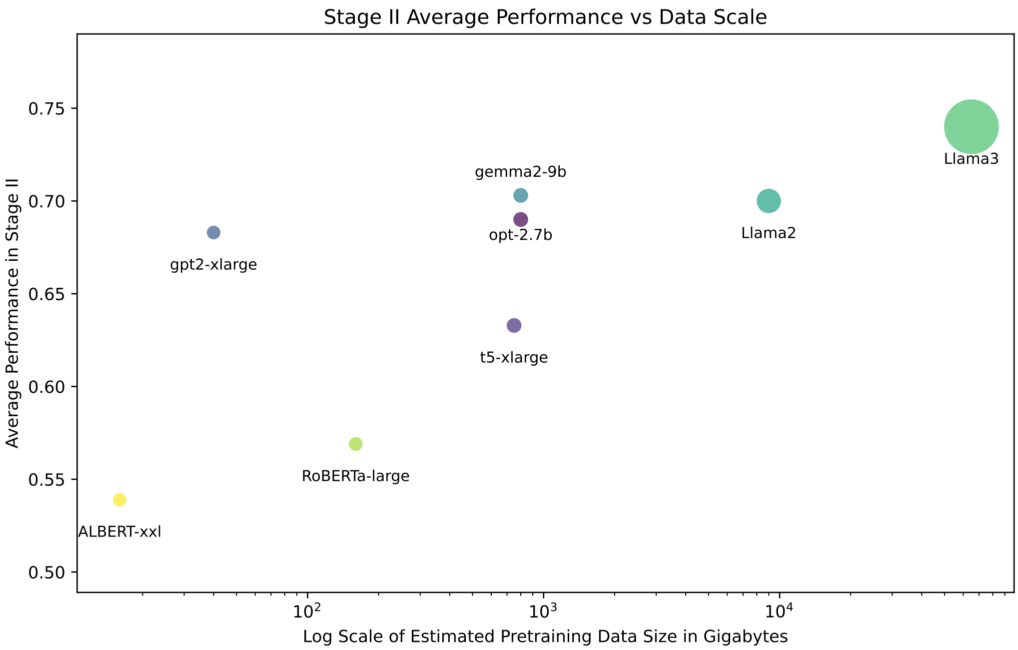


Figure 14: Stage II performance (normalized) vs. their data scale in the logarithm of Gigabyte.

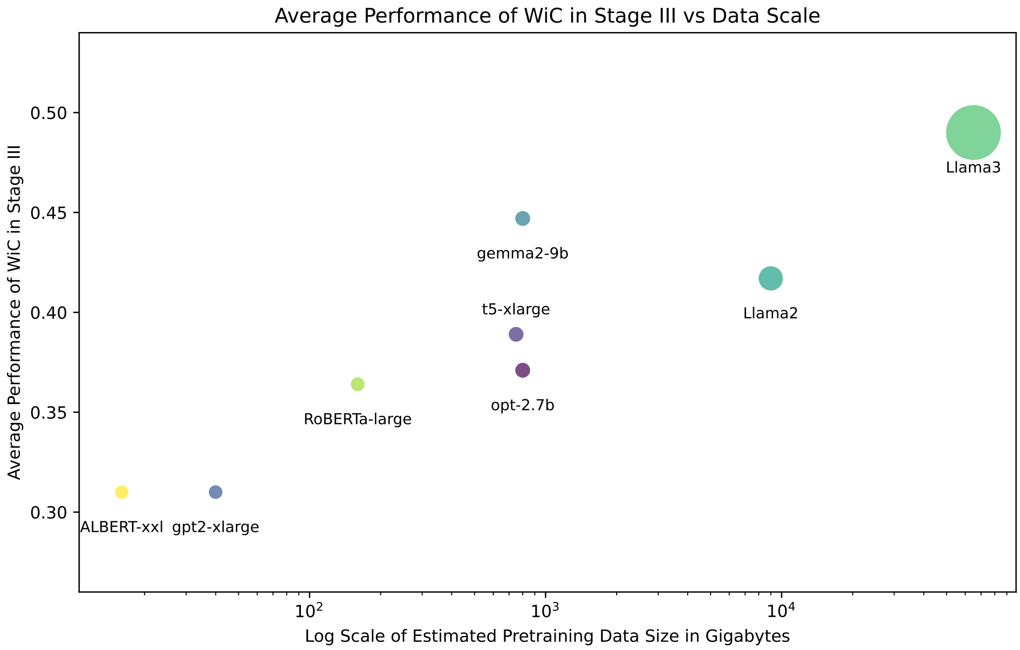


Figure 15: WiC in Stage III performance (normalized) vs. their data scale in the logarithm of Gigabyte.

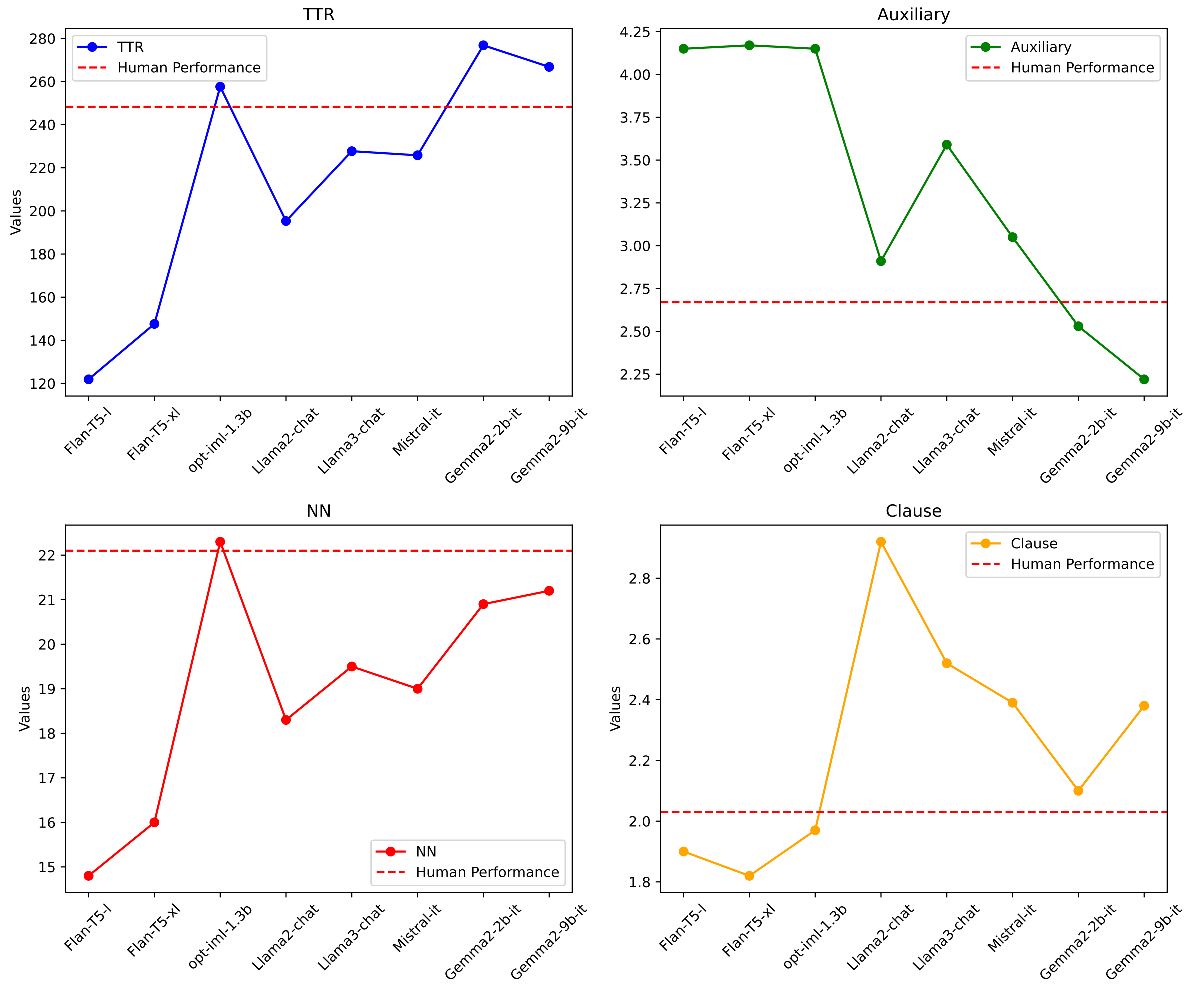


Figure 16: Four different dimensions of linguistic features in generated texts. Models are ordered by time

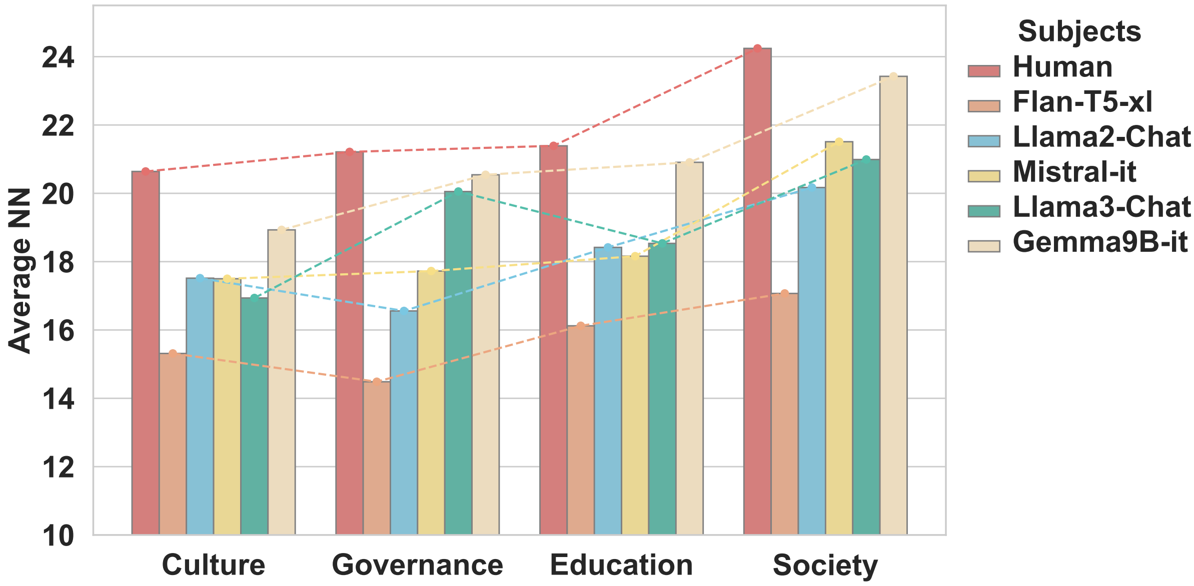


Figure 17: Measurement of average NN across four different categories. Models are ordered by time. We could also find a trend that is similar to human. However, compared to TTR, the NN metric—focused on noun usage—remains less aligned with human patterns.

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 4: 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 5: 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. To access a comprehensive list of results, please refer to our project page which can be found on the first page.

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 6: One example from each dataset.

Example of Essay in Governance Topic

The Art of Simplification: A Necessary Counterpoint to Complexity

The statement, "Anyone can make things bigger and more complex. What requires real effort and courage is to move in the opposite direction—in other words, to make things as simple as possible," resonates deeply with me. While it's undeniably easier to add layers, features, and bureaucracy, true ingenuity lies in stripping away the unnecessary and distilling the essence of a problem or system. This principle applies not only to physical objects but also to social structures, processes, and even our own lives.

The allure of complexity is undeniable. We are constantly bombarded with information, options, and technological advancements, leading to a natural inclination towards accumulating more. Businesses strive for bigger market shares, governments expand their reach, and individuals seek ever-increasing levels of comfort and convenience. This pursuit of "more" often results in convoluted systems, inefficient processes, and a sense of overwhelm.

Consider the healthcare system. While advancements in medical technology have undoubtedly improved patient outcomes, the system itself has become increasingly complex, with multiple stakeholders, intricate insurance networks, and a labyrinthine web of regulations. This complexity leads to higher costs, administrative burdens, and ultimately, a less accessible and equitable system for many.

Similarly, the rise of big government, often cited as an example of excessive complexity, can be detrimental. While government intervention can be necessary to address market failures and provide essential services, an overreaching bureaucracy can stifle innovation, create inefficiencies, and erode individual liberties.

However, the path of simplification is not without its challenges. It requires a willingness to question assumptions, challenge the status quo, and embrace radical transparency. It demands a deep understanding of the underlying principles and a commitment to identifying and eliminating redundancies and inefficiencies.

There are, of course, circumstances where complexity is unavoidable and even beneficial. For instance, advanced scientific research often necessitates intricate models and sophisticated technologies. Similarly, complex legal frameworks are necessary to navigate the intricacies of societal interactions and protect individual rights.

The key lies in finding the right balance. We must strive to simplify wherever possible without sacrificing essential functionality or undermining the need for complexity in specific domains.

This principle of simplification extends beyond the realm of systems and structures. It applies to our personal lives as well. We are constantly bombarded with information and distractions, leading to a sense of overwhelm and a lack of focus. By consciously simplifying our routines, decluttering our environments, and prioritizing meaningful experiences, we can create space for clarity, creativity, and genuine connection.

In conclusion, while the allure of complexity is strong, the true test of ingenuity lies in the ability to simplify. It requires courage, vision, and a commitment to identifying and eliminating the unnecessary. By embracing this principle, we can create more efficient systems, foster greater clarity, and ultimately, live more fulfilling lives.

Example of Essay after steering from Governance to Culture

The adage "simplicity is next to godliness" certainly holds weight when considering the complexities of modern life. While the allure of complexity and the allure of "bigger government" might seem attractive on the surface, the pursuit of simplicity, as the statement suggests, requires a deeper understanding of our values and a willingness to challenge the status quo.
I largely agree with the statement that simplifying complex systems is a greater challenge and requires more courage than making things more complex. Here's why:
1. The Culture of Complexity: We live in a culture that often equates complexity with progress and sophistication. Complex systems, like complex government programs, can seem more sophisticated and complex, and therefore, more impressive. This can lead to a resistance to simplification, as it can be perceived as a sign of weakness or a rejection of progress.
2. The Comfort Zone: We are creatures of habit and comfort. Embracing simplicity often means challenging our ingrained habits and challenging the status quo. It requires a willingness to step outside our comfort zone and embrace a different way of living and interacting with the world.
3. The Power of Simplicity: Simplicity, in its purest form, often leads to greater efficiency, clarity, and cultural impact. Think about the impact of movements like minimalism in design. They're not about deprivation, but rather about focusing on what truly matters and simplifying our lives to free up time and energy for what truly matters.
4. The Challenge of Cultural Shift: Shifting towards a culture that values simplicity over complexity is a monumental task. It requires a cultural shift that values principles like simplicity and mindfulness. It's about shifting the cultural narrative and moving away from the idea that more is always better.
Examples:
* **Technology:** The cultural shift towards minimalism in technology, as seen in the minimalist design of apps and interfaces, reflects a growing appreciation for simplicity. * **Urban Planning:** The move towards walkable cities and neighborhoods reflects a growing appreciation for community and sustainability. **Politics and Governance:** The growing emphasis on cultural humility in governance reflects an awareness of the need for simplicity and thoughtful leadership.
Conclusion:
While the allure of complexity and "bigger government" might seem attractive on the surface, the pursuit of simplicity—as the statement suggests—requires a deeper understanding of our values and a willingness to challenge the status quo. Embracing simplicity can lead to greater well-being, efficiency, and a more sustainable future. It's about recognizing that less can often be more, and that true progress lies in simplifying our lives and focusing on what truly matters.