

Miss Universal Grammar: Pretraining BabyLMs with Linguistic and Non-Linguistic Sources

Erika Varis Doggett*

Disney Research Imagineering
Walt Disney Imagineering
erika.varis.doggett@disney.com

Mary Katie Kennedy*

University of Southern California
mkkenned@usc.edu

Abstract

Inspired by linguistic theory and recent work exploring linguistic biasing of neural networks, we propose a “Universal Grammar” (UG) pre-training paradigm for transformer-based neural networks, preparing the network to learn language more effectively, especially in low data volume scenarios. We accomplish this “universal” pretraining with a combination of synthetic language-like sequences created with a context-free grammar, biological sequence sources like protein amino acids, and material sources like crystal structures. Experiments with the BabyLM Challenge datasets show improved language acquisition in English, Dutch, and Chinese with our MissUniverse models, compared to not using “UG” pretraining. This pretraining paradigm has interesting implications for both the future of language models, and potentially how we think about human Universal Grammar.

1 Introduction

Children demonstrate a reasonable mastery over their native language around age five (Gleitman and Newport, 1995), a skill that continues to expand into a proficiency over complex grammatical constructions, a wider vocabulary, and even an awareness of what language to use in which social contexts for specific conversational goals (Clark, 2024). All of this is achieved despite children being exposed to no more than 50M words by age five (Warstadt and Bowman, 2024; Frank, 2023), and less than 100M words by age 13 (Gilkerson et al., 2017). In contrast, large language models (LLMs) often need to train on orders of magnitude more data to even broach human proficiency. To further highlight this divide between the language-acquisition efficiency of humans and LLMs, bilingual children—those who acquire two languages from an early age as their two first languages—have

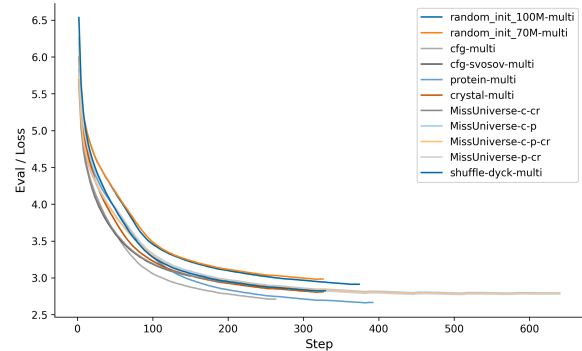


Figure 1: Eval loss curves comparing different data sources to induce a “Universal Grammar” in stage 1 pretraining, for multilingual models.¹

an overall reduced input of both languages due to their inputs being split between the two languages; despite this, research has found that bilingual children utilize vocabulary diversity and syntactic complexity as rich as their monolingual counterparts (Nicoladis and Jiang, 2018; Korade et al., 2024).

Observation of children’s remarkable language-learning ability led to the notion of the Poverty of the Stimulus (PotS) (Chomsky, 1965, 1981b; Lightfoot, 1989; Crain, 1991). In brief, PotS raises the paradoxical problem that the inputs children receive are not sufficient enough to specify the correct grammar that children ultimately converge upon, rather than the numerous other possible grammatical explanations for the language inputs. A solution to this problem has come to be known as “Universal Grammar” (UG), which has largely been defined variously over the years by Noam Chomsky as the biological grammar possessed by all humans (Roberts, 2017). While this theory has received its criticisms (Evans and Levinson, 2009; Tomasello, 2009; Ambridge and Lieven, 2014), we take its core notions as inspiration for our BabyLM.

¹Only the competition MissUniverse models were run for the full 10 epochs; all others underwent early stopping when loss plateaued.

*Both authors contributed equally to this work.

Building on the tradition of inducing syntactic biases in LLMs (Gessler and Schneider, 2023; Mueller and Linzen, 2023; Hu et al., 2025), we aim to induce an analogous “UG” state for an artificial neural network, inspired by hierarchical fundamentals and Merge operations proposed in human theoretical UG (Chomsky, 1995). To achieve this, we extend the “pre-pretraining” methodology of Hu et al. (2025) and pretrain on data that demonstrates hierarchical structures and geometric folding. Our work diverges from the human-centric limitation of traditional conceptualizations of UG in that our data takes inspiration not only from human language, but also from biological and physical “languages” and their associated “syntax.” Recent work has found LLMs develop non-trivial, deep representations of human language through training on unlabeled text (Hewitt and Manning, 2019; Kulmizev et al., 2020; Li et al., 2021a; Liu et al., 2026); the same mechanisms that have enabled its success in human language have also been shown to be applicable in biological processes such as proteins (Rao et al., 2021; Jumper et al., 2021; Lin et al., 2023b) as well as physical materials such as crystal structures (Antunes et al., 2024; Zimmermann et al., 2025; Zhang et al., 2026).

By pretraining on data sources that demonstrate hierarchical syntax, we aim to induce in our model its own “UG”, one that is not limited only to human language, but which nevertheless demonstrates structural properties that are analogous to the basic principles of syntax in human language, such as Internal and External Merge (Chomsky, 1995). By focusing on data that demonstrates hierarchical structures wherein long-distance elements are folded into a close proximity or in which certain elements combine only with other elements, we seek to induce the basic fundamentals of syntax as the models’ own UG that might enable it to better acquire a variety of human languages.

While our artificial “UG” proposal here is inspired by the theory of human UG, it should by no means be taken as an endorsement of that theory for either humans or ANNs. We merely draw the analogy between a theoretical framework that has been proposed for humans, and a training paradigm for transformer-based ANNs: if human languages all share the same basic linguistic principles, can we prime a neural network to be quicker at acquiring such principles before exposure to natural language ever occurs?

2 Background

2.1 Linguistic Principles

In the field of linguistics, a well-known framework of syntax is generative Minimalist Syntax, which has developed out of the foundational theories of Noam Chomsky (1957; 1981a; 1986; 1965). This constituency-based family of grammatical theories is built on the binary-branching X-bar theory and its core operations of *External Merge* and *Internal Merge* (also known as “Move”). These Merge operations serve to connect elements; External Merge involves the formation of phrases when heads take their arguments and specifiers (e.g., the verb “eat” taking “apples” as its argument to form the verb phrase “eat apples”), while Internal Merge involves the movement of formed phrases to new landing spots in the syntax (e.g., “What did you see them eat ~~what~~?”). After all operations are applied in the course of derivation, the sentence is then linearized when read from left-to-right along the leaf nodes.

2.2 LLMs are capable of discovering syntax

The interpretability field of LLMs has amassed a number of converging studies that indicate models’ ability to discover syntax, such as the work of Hewitt and Manning (2019) who found that dependency structures are recoverable from the contextualized representations of sentences while other work, such as Liu et al. (2026), find that LLMs display behaviors reflecting tree-like representations of constituency.

Within the space of multilingual models, several studies have converged upon evidence that multilingual models share an abstract, conceptual representation space; Starace et al. (2023) found linguistic information, such as part-of-speech, to have shared representations across languages, which Brinkmann et al. (2025) further expanded by finding morphosyntactic concepts such as grammatical number, gender, and tense to be shared across languages. Models furthermore have been discovered to display similar behavior in their representations across various layers. Several researchers have found early layers constitute an input space that aligns to either the input language or the semantics; middle layers focus on task-solving and convert the representations into a shared abstract concept space, which often is more aligned to English representations; and late layers finally convert the concepts back into the output language (Zhao et al., 2024; Wendler et al., 2024; Tamo et al., 2026).

Models’ ability to converge upon shared, universal grammatical features and to furthermore share abstract representations are indicative of models’ abilities to not only extract syntax, but to create their own iteration of a Universal Grammar-like functionality if trained upon enough data.

2.3 Syntactic Induction in LLMs

Previous research in LLMs has sought to induce syntactic biases within models during pretraining (Zhou et al., 2020; Xu et al., 2021; Bai et al., 2021; Li et al., 2021b; Zhang et al., 2022; Gessler and Schneider, 2023; Yang et al., 2026). Given that models are able to extract core grammatical concepts across languages if provided enough unlabeled training data, we seek to discover whether inducing the basic fundamental operations of syntax within models can provide the Universal Grammar tools, and reduce the necessary monolingual and multilingual training data while achieving improved performance. Our work in particular builds off the research of Hu et al. (2025), which found that a stage of pretraining on the context-sensitive, hierarchical formal language of k -Shuffle Dyck conferred an inductive bias that reduced the necessary token budget by 33%.

Using the methodology of Hu et al. (2025), we seek to pretrain our models on a variety of different training data sources that demonstrate hierarchical structures, geometric folding akin to Internal Merge, and mildly context-sensitive grammars in an attempt to induce in our models a UG similar to what is postulated to exist in humans.

3 Data

The inspiration for our data comes from linguistic and non-linguistic grammars, similar to Papadimitriou and Jurafsky (2023) who pretrained on music. Hu et al. (2025) found that among their set of synthetic formal languages, the best data for pre-pretraining was the context-sensitive k -Shuffle Dyck language. This parallels human language, which generally falls within the mildly context-sensitive tier of grammar (Joshi et al., 1990).

However, syntax is found not only in human language, but in nature as well. In this section, we describe the three datasets we use for pretraining to induce a syntactic bias within our models.

3.1 Proteins

Proteins are three-dimensional structures that are the result of a linearized amino acid chain, which

will always² fold in a *specific* configuration depending upon the chemical bonds that form between the atoms (Anfinsen, 1973). However, searching the space of all possible folding configurations would take an exorbitant time; despite this, proteins are able to fold into their specific shape in mere milliseconds or less (Dill et al., 2007), and this near instantaneous folding of a seemingly innumerable space of possibilities has come to be known as the Levinthal Paradox (Levinthal, 1968).

Early research identified computational linguistics as a means to solve this problem. Dill et al. (2007) developed their Zipping and Assembly Mechanism by Dynamic Programming (ZAMDP), inspired by the computational linguistic methods of CKY algorithm and chart parsing, to solve local problems in order to achieve a global structure for protein and RNA secondary structures. They further observe that basic biopolymer structures, such as β hairpins in protein structures or the stem loops of secondary RNA structures, demonstrate simplistic context-free grammars with recursively nested links. More complex structures, such as protein configurations like helix bundles and β sheets along with RNA secondary pseudoknots, demonstrate crossing dependencies like that observed in Dutch. In this regard, proteins are mildly context-sensitive, similar to human language.

Recently, the advent of LLMs has allowed the field to advance in leaps and bounds, with AlphaFold v2 winning the 2024 Nobel Prize for Chemistry for solving the Levinthal Paradox. Countless scores of LLMs have demonstrated remarkable potential for models to assist with protein structures (Jumper et al., 2021; Baek et al., 2021; Rao et al., 2021; Vig et al., 2021; Wang et al., 2022; Lin et al., 2023b; Fang et al., 2023). Even more notably, the protein language model ESM2 (Lin et al., 2023b) masks out 15% of amino acids in a protein’s amino acid chain, and trains the model to predict the missing acid, which requires the model to create complex, contextualized representations of the protein, similar to how LLMs develop rich representations of sentences (Brinkmann et al., 2025; Hewitt and Manning, 2019; Nikolaev and Padó, 2023; Kamath et al., 2024). These well-contextualized representations of the linearized chain of amino acids provide the necessary information for a secondary model, ESMFold, to efficiently predict three dimen-

²Prions are an exception to this as they are misfolded proteins that induce misfolding in adjacent proteins. For discussion, see Ezzat and Espay (2023).

sional protein structures.

Dataset Because ESM2 and ESMFold (Lin et al., 2023b) demonstrate that linearized amino acid chains of proteins encodes the information necessary to predict complex protein folding, we use the Huggingface protein dataset released by the ESM2 team (Lin et al., 2023a). This protein dataset, built off of the UniRef dataset (Suzek et al., 2015), is composed of clustered sequences according to length and similarity, increasing variation while limiting redundancy. For training budgets, we subsampled the data to 1M sequences or less and filtered to sequences that could fit within a 1024 token context window, resulting in a total of 966,428 sequences.

3.2 Crystal Structure

The biomedical field is not the only area to leverage LLMs’ remarkable abilities for syntax discovery. The materials science field has also demonstrated LLMs’ potential within the realm of crystals and chemistry (Jablonka et al., 2023; Zimmermann et al., 2025; Khastagir et al., 2025; Zhang et al., 2026). Among this research includes the exploration of LLMs as a means to identify new, three dimensional crystalline materials—structures which are composed of a repeating minimal unit of atoms in a three dimensional space—with various chemical properties for use in batteries, semiconductors, solar cells, and more (Khastagir et al., 2025). Unlike many protein models, many crystal generation models utilize an autoregressive (AR) task to generate plausible novel crystalline structures (Antunes et al., 2024; Khastagir et al., 2025).

Because crystals represent three dimensional structures that are formed according to a chemical syntax with possibly infinitely repeating forms, reminiscent of human language’s ability for recursion, we select crystals as another dataset for pretraining to inject syntax.

Dataset Antunes et al. (2024) trained their crystal generative model CrystaLLM model on Crystallographic Information Files (CIFs), which contains structural information about the material, including a tabular list of atom’s element and their xyz coordinates. We use the dataset from CrystaLLM,³ which is composed of data from the Materials Project (Jain et al., 2013), OQMD (James E. Saal, 2013),

and NOMAD (Scheidgen et al., 2023). We used the linearized list of atoms, and a standard BPE tokenizer trained on a subsample of the data.⁴

3.3 Synthetic Language

While Hu et al. (2025) used formal languages for pretraining, we were inspired by simple context-free grammars (CFGs) for creating language-like sequences, and added high-level distributional patterns to mimic the mild context-sensitive patterns found in natural language, like vocabulary frequency distributions, morphological patterns, and semantic clustering. We outline the pertinent features in the following sub-sections. Using the full system, we generate 1,000,000 sequences with a mean length of 20.5 words (ranging from 7 to 26 words each).

Grammar rules The context-free grammar rules are SVO (Subject-Verb-Object) with head-initial tree branching, and includes clause, inflection, and tense syntactic phrase options; preposition phrases; adverbs; adjectives; determiner+noun phrases; and sentential-level conjunctions. The tree depth is limited to 12 levels, for compute purposes. We also did some exploration into including an alternate SOV word order, outlined in section 5.2.

Zipfian distribution The base vocabulary size is set to 1,000 nonce words with English phonotactics. About 60% of the words were monosyllabic, and 40% bisyllabic.

The words were assigned to the different syntactic categories, and Zipfian frequencies assigned to them, with boosts depending on how often the CFG rules use the category (see Table 5 in the Appendix). Thus, we create a plausible distribution of higher and lower frequency words within the sequences.

Morphological connections Within our base lexical vocabulary (adjectives, adverbs, nouns, and verbs) we took 30% of each category to transform with morphological affixes (again, English plausible phonotactically). The morphological items received a frequency assignment of 10% of the target

³<https://github.com/lantunes/CrystaLLM#obtaining-the-training-data>

⁴Early tests with the xyz coordinates showed poor performance, so we omitted the coordinates. However, we recognize that CrystaLLM and other previous work used customized tokenizers with the xyz coordinate information. Future work can explore whether coordinates and other spatial information can be beneficially worked into our UG pretraining through additional preprocessing, tokenization, or other such measures.

frequency for the lexical category, so the morphological rule is present but does not overwhelm the vocabulary. This expansion added about 500 words to the base vocabulary.

Agglutination While languages like English and Chinese largely rely on strict word order and minimal morphological affixation, other languages are highly agglutinative. To mimic agglutination, we introduced a post-lexical fusion effect.

We create a small set of templates for items that can be fused together, like verb-object and determiner-noun, with a corresponding rate of application (e.g., 15%). For example, words like “the” “cat” might be concatenated with an optional linking vowel such as “o”, resulting in the newly fused word “thecat” or “thecat”.

Semantic clustering To artificially create semantic clustering effects, we assign 4–5 arbitrary semantic fields to our lexical categories. An example of semantic fields for adjectives would be physical (e.g., size/color), evaluative (e.g., good/bad), emotional (e.g., sad/happy), temporal, and quantitative (e.g., amount).

We assign a compatibility matrix to the different semantic fields to identify which ones “go together” (e.g., object nouns with physical adjectives), and assign probability weights of high (0.9), medium (0.7–0.8), or default (0.3) for the different semantic field combinations.

At sequence generation, when a word is chosen its semantic field becomes an active context, and the frequency for the next word is multiplied by the semantic weights of its matrix field assignments, as illustrated in 1.

$$W = f_{zipf} + (f_{zipf} * w_{sem}) \quad (1)$$

Thus, if a word would otherwise receive a frequency score of 0.5 but it falls within a high compatibility semantic field, it receives a final frequency weight of 0.5 (base Zipfian) + (0.5 base Zipfian * 0.9 compatibility weight) = 0.95, whereas one that falls within a low compatibility semantic field receives a final weight of 0.5 + (0.5 * 0.3) = 0.65.

CFG feature ablation Ablation experiments with the different CFG features—morphological affixation, agglutination, semantic clustering—showed that a CFG using all of our linguistically-inspired features had better loss curves when training on subsequent BabyLM data than CFG using only one or a subset of features (see Figure 2).

We generated 1,000,000 sequences of 20.5 average word length using this system.

4 Method

4.1 Stages

Following [Hu et al. \(2025\)](#), we approached training in two stages: stage 1, in which we train on the non-human language data sources in order to induce a “UG” state, and stage 2, which trains on the BabyLM datasets for natural language learning.

We performed experiments on the multilingual BabyLM challenge sets (English, Dutch, and Chinese), with results reported in section 5, along with some testing on the BabyLM strict track (100M token) dataset and strict-small (10M), which are reported in the Appendix.

Our challenge track models take the MissUniverse moniker, while all others represent experiment ablations and explorations.

4.2 Tokenizers

For the sake of simplicity in these experiments, we used a custom BPE tokenizer that is trained up front before beginning stage 1 or stage 2 training. The tokenizer is trained from the stage 1 data (e.g., CFG, proteins, crystals, or a mix thereof) and a subset of stage 2 data (i.e., BabyLM data) so that the same tokenization and vocabulary embeddings can be used for both stages.⁵

For English and Dutch, we use whitespace separation and lowercase normalization, and for Chinese we use NFC unicode normalization and byte fallback for the BPE tokenizer. Our vocabulary size is approximately 8,000 tokens.

We also include a test with a custom shared tokenizer, trained on all data sources (CFG, proteins, crystals, English, Dutch, and Chinese)⁶, and increased the vocabulary size to 32,000. The results for this shared tokenizer test are available in the Appendix.

4.3 Models

Our models implement GPT-2 ([Radford et al., 2019](#)) as a base architecture as it fits well on a single

⁵Future research may consider experimenting with tokenizer adaptation between stages so that the network could be trained with a stage 1 tokenizer for non-language data before being adapted to a stage 2 tokenizer for natural language data.

⁶We used ratios according to our most multilingual test case for this tokenizer: approximately 30% allocated to CFG, proteins, and crystals (in a 1:1:0.5 ratio themselves, by sequence count) and 70% allocated to a 1:1:1 mix of English, Dutch, and Chinese.

GPU, allowing for a high degree of experimentation with low computational costs.⁷ We performed tests on both large (approximately 0.7B param) and small (124M param) GPT-2 model sizes,⁸ the latter of which is reported in the Appendix.

Following Hu et al. (2025), we used a max input length of 1024 to take advantage of potential long distance dependencies in protein sequences. Since the individual CFG and crystal sequences were often much shorter, we packed multiple “sentences” into the input. We found this improved results over training on too many padding tokens.⁹

4.4 Task

We trained our models on a standard autoregressive task, focusing on the effects of the stage 1 data pretraining. We selected the autoregressive architecture as our underlying base because its unidirectional nature mimics how humans process language inputs incrementally, building expectations of the sentence’s syntactic structure as we process the linearized outputs of a speaker or text (Marslen-Wilson, 1975; Konieczny, 2000; van Gompel and Pickering, 2007). Furthermore, GPT2 specifically has been found to be a better predictor of online human language processing (Lee et al., 2024).¹⁰

We leave experimentation with masked language modeling tasks to future research.

Evaluation We primarily evaluate our UG pretraining method on a multilingual mixture of English, Dutch, and Chinese datasets with the provided BabyLM evaluation pipelines. When relevant, we compare our models to vanilla GPT2 base-

lines that trained from random initialization solely on natural language, without interrupted staged training.

5 Results

We report the by-language results of our MissUniverse models in Table 1,¹¹ and compare our models to two baselines: a vanilla GPT2 model trained on the full 100M BabyLM token budget (random init 100M) and a vanilla GPT2 model trained on an amount of natural language data equivalent to the roughly 70M tokens the MissUniverse models (random init 70M).

From Table 1, we can observe that few instances of our MissUniverse models outperform the 100M baseline overall, with the exceptions of our CFG-protein and protein-crystal mix for Chinese evals. That being said, several of our MissUniverse models outperform the 70M baseline, which trained on an equivalent amount of natural language data. When examining the overall multilingual zero-shot average (see Table 2), only MissUniverse-c-p outperforms the equivalent 70M baseline while still falling short of 100M tokens of natural language training.

We might conclude from this that only the CFG-protein mix confers a linguistic bias that exceeds what can be learned from 70M tokens alone, and that this linguistic boost does not surpass what could have been learned from utilizing the 30M tokens on natural language. However, two key observations stand out. First, when the overall score is calculated using zero-shot and fine-tuning results (Table 2), all MissUniverse models exceed both baselines with the exception of MissUniverse-c-p, which still outperforms the 70M baseline. Second, we observe that in Table 1, many of our models outperform the baselines when the benchmark is *not* evaluating grammatical knowledge. In Table 2, we split evaluations into grammatical (BLiMP and MultiBLiMP, when relevant) and non-grammatical (HellaSwag, Winogrande, GlobalPIQA, XStoryCloze, and XCOMPS, when relevant). Upon doing so, we find that stage 1 pretraining by and large confers a notable benefit when the evaluation is *not* testing the model’s grammatical ability, with our models exceeding even the 100M baseline for Chinese (p-cr) and English (c-p-cr and c-p).

⁷Initial experiments suggest that our approach of pretraining on non-human language data—including synthetic CFGs, protein sequences, and crystal CIF sequences—should extend to other small language models like the Phi family of models (Jawaheripi et al., 2023); however, due to resource constraints, we focus our experimentation on the GPT-2 architecture.

⁸While the aim was for 1B param and 124M param sizes, due to implementing a more limited vocabulary size in order to train better on our pretraining data mixes, our true “large” model size was 706.7M parameters. Likewise, the 124M parameter model is truly 98M parameters. For reference purposes in this paper, we use “0.7B param” for our large GPT-2 model size, and “124M param” for our small GPT-2 model size.

⁹When stage 1 data sources were mixed in a single model, packing was heterogeneous.

¹⁰While humans receive and produce language incrementally, research suggests that they build hierarchical expectations during processing and utilize hierarchical planning during production (Lee et al., 2013). Future work could consider exploring how unidirectional and bidirectional training objectives or a combination of the two affect syntactic bias induction.

¹¹Note that random init 70M-multi is averaged from two runs.

Multilingual Model (NL Evals)	Avg (NL)	BLiMP	MultiBLiMP	HellaSwag	Winogrande	XStoryCloze(NL)	XCOMPS	GlobalPIQA	FineTune
random init 100M-multi	<u>53.76</u> †	<u>73.75</u> †	<u>87.77</u> †	<u>26.75</u> †	49.88†	47.60	51.41†	<u>39.14</u> †	37.79†
random init 70M-multi	53.14	72.34	86.25	26.53	49.59	47.99*	51.19	38.10	36.63
MissUniverse-c-p-cr	52.94	71.61	85.16	26.48	50.04†*	48.37†*	50.77	38.12†	37.18†
MissUniverse-c-p	53.05	72.44†	85.20	26.65†	49.63†	<u>50.39</u> †*	50.95	36.11	36.85†
MissUniverse-p-cr	52.92	72.09	84.94	26.29	<u>50.21</u> †*	49.15†*	50.12	37.61	<u>38.27</u> †*
MissUniverse-c-cr	52.94	71.37	85.33	26.65†	50.04†*	48.99†*	<u>51.50</u> †*	36.68	38.09†*
Multilingual Model (ZH Evals)	Avg (ZH)	BLiMP		HellaSwag	Winogrande	XStoryCloze	XCOMPS(Zho)	GlobalPIQA	FineTune
random init 100M-multi	<u>47.47</u> †	<u>74.27</u> †		26.76†	49.38	46.67	52.56	35.15†	39.99
random init 70M-multi	47.45	72.32		26.74	<u>51.07</u> *	47.33*	52.58*	34.64	40.61*
MissUniverse-c-p-cr	47.19	72.69†		26.64	48.06	<u>48.84</u> †*	<u>52.70</u> †*	34.18	<u>40.61</u> †*
MissUniverse-c-p	<u>47.67</u> †*	73.83†		26.85†*	49.63*	47.91†*	51.18	<u>36.65</u> †*	39.86
MissUniverse-p-cr	47.47 †*	71.77		<u>26.92</u> †*	49.79*	<u>48.84</u> †*	52.33	35.15†*	40.51*
MissUniverse-c-cr	46.85	72.76†		26.89†*	49.96*	<u>47.52</u> †*	51.25	32.74	40.06*
Multilingual Model (EN Evals)	Avg (EN)	BLiMP	MultiBLiMP	HellaSwag	Winogrande	XStoryCloze		GlobalPIQA	FineTune
random init 100M-multi	52.48 †	65.30†	83.64†	26.72	<u>51.36</u> †	49.77†		38.08†	41.04†
random init 70M-multi	51.44	64.46	81.82	26.84*	<u>51.36</u> *	49.54		34.62	40.60
MissUniverse-c-p-cr	51.66 †	64.33	79.61	26.88†*	49.22	<u>49.85</u> †*		<u>40.08</u> †*	42.64†*
MissUniverse-c-p	52.09 †	64.21	81.17	26.77†*	51.20	49.69†		39.52†*	41.94†*
MissUniverse-p-cr	51.33	64.74†	80.00	26.57	51.28	49.23		36.15†	42.98†*
MissUniverse-c-cr	52.01 †	64.17	82.73†	26.79*	49.79	49.46		39.11†*	<u>43.05</u> †*

Table 1: Multilingual zero-shot evaluation results for Dutch (NL, top), Chinese (ZH, middle), and English (EN, bottom). Average reports the zero-shot average per language. Best performers are underlined, and those that beat or tie with random init (70M) are marked with a † while those that beat or tie with random init (100M) are marked with *.

We postulate that this may be due to stage 1 acquiring some of the basic operations of syntax, thus freeing up some of the model’s resources in stage 2 for world-knowledge acquisition, though this seems to come at the cost of more language-specific syntactic acquisition. It is possible that the lagging grammatical knowledge is due to our models needing to learn not just three grammars, but five or even six grammars: the grammar of our CFG, of proteins, and of crystal structures, in addition to the three natural languages. To test this, we conduct two follow-up experiments: we reduce the number of grammars the model is exposed to in stage 1 training (Section 5.1), and we reduce the number of languages in stage 2 down to a single, monolingual condition (Section 5.2).

5.1 Single-Source Stage 1

To reduce the number of grammars the models are exposed to, we limit the number of stage 1 inputs to a single non-language source. We additionally include a new CFG that includes both SVO and SOV word order in order to see if that improves the models’ performance on Dutch, which has varied word-order.

As Table 3 demonstrates, this method did not drastically impact scores. English scores remain relatively unchanged, with several models beating the 70M baseline but not the 100M. Chinese and Dutch show slight changes: previously no model beat any baseline for Dutch, while protein and shuffle-dyck are able to exceed the 70M baseline. For Chinese, previously two models (CFG-protein

and protein-crystal) outperformed both baselines, whereas for single-source only shuffle-dyck is able to tie with the 70M. Overall, we do not find reducing the number of input sources for stage 1 to simplify the system enough for our multilingual models to excel over the baselines for zero-shot results.¹²

5.2 Monolingual Stage 2

As a second follow-up exploration, we reduce the number of languages the model trains on. Rather than focusing on multilingual models as our MissUniverse models did, we instead turn our attention to the effects of stage 1 training effects on *monolingual* models. For baseline comparisons, we trained a vanilla GPT2 baseline on 100M tokens of a single language: English, Chinese, Dutch. We then initiated stage 1 training on a data mix of CFG, proteins, and crystals, as well as single-source inputs, including shuffle-dyck. Because Dutch demonstrates SOV and has some measure of flexible word-ordering, we additionally included single-source CFGs that were SVO, SOV, and a mix of the two.

When we compare the results of Table 4 to those in Table 1 and Table 3, we find a large increase in the number of models that exceed a model that was trained on 100M of natural language data. Previously, only the CFG-protein and CFG-crystal

¹²In this follow-up experiment, we did not fine-tune the new single-source models due to compute restraints. Future research may consider exploring the differences between multi-source and single-source stage 1 training on fine-tune evaluations.

Model	Multilingual Avgs				NL Avgs		ZH Avgs		EN Avgs	
	Overall	Zero-shot	Gram	Other	Gram	Other	Gram	Other	Gram	Other
random init 100M-multi	44.72 [†]	51.37 [†]	76.95 [†]	42.23 [†]	80.76 [†]	42.96 [†]	74.27 [†]	42.10	74.47 [†]	41.48 [†]
random init 70M-multi	44.31	50.81	75.44	42.01	79.30	42.68	72.32	42.47 [*]	73.14	40.59
MissUniverse-c-p-cr	44.76 [†] [*]	50.72	74.68	42.16 [†]	78.39	42.76 [†]	72.69 [†]	42.08	71.97	41.51 [†] [*]
MissUniverse-c-p	44.56 [†]	51.05 [†]	75.37	42.37 [†] [*]	78.82	42.75 [†]	73.83 [†]	42.44 [†]	72.69	41.80 [†] [*]
MissUniverse-p-cr	44.99 [†] [*]	50.70	74.71	42.12 [†]	78.52	42.68 [†]	71.77	42.61 [†] [*]	72.37	40.81 [†]
MissUniverse-c-cr	44.90 [†] [*]	50.72	75.25	41.95	78.35	42.77 [†]	72.67 [†]	41.67	73.45 [†]	41.29 [†]

Table 2: Multilingual averages, with the Overall including fine-tune evals in addition to zero-shots. Zero-shot evals are further sub-divided into grammatical benchmarks (BLiMP and MultiBLiMP when available) and non-grammatical “other” benchmarks (HellaSwag, Winogrande, XstoryCloze, XCOMPS when available, and GlobalPIQA). Best-performing models are underlined; models that outperform or tie with the 70M baseline are denoted with [†] while those that outperform the 100M baseline are marked with ^{*}.

Multiling Model	Multiling Avg	NL Avg	ZH Avg	EN Avg
rand init 100M	51.37 [†]	53.76 [†]	47.47 [†]	52.48 [†]
rand init 70M	50.81	53.14	47.45	51.44
CFG (SVO)	50.14	52.40	46.60	51.05 [†]
CFG (SVO-SOV)	50.70	52.21	47.43	52.20 [†]
protein	50.91 [†]	53.19 [†]	47.42	51.75 [†]
crystal	50.54	52.55	46.38	52.35 [†]
shuffle-dyck	50.83 [†]	53.52 [†]	47.45 [†]	51.08

Table 3: Single source multilingual zero-shot evaluation results for Dutch (NL, top), Chinese (ZH, middle), and English (EN, bottom). Best models are underlined. Those that beat or tie with the random init baselines are marked with [†] for random init (70M) and ^{*} for random init (100M).

MissUniverses outpaced their 100M baseline and only for Chinese. However, if the model’s required training is reduced from learning three languages to learning merely one, the benefits of the syntactic induction begin to outpace the benefits of training on an extra 30M tokens of natural language data.

By reducing the models’ natural language acquisition demands, the benefits of multi-source stage 1 training become more prevalent, particularly for Dutch, which exceeds the baseline for all but the SVO-protein mix. Interestingly, we find the SVO-cr mix to benefit all three languages evaluated in BabyLM.

We postulate the strength of SVO-cr across Chinese, English, and Dutch is due in part to the linguistically-informed nature of the CFG, especially for these languages that demonstrate SVO order. It is possible that the CFG is closer in nature to the syntax demonstrated in the natural language data, thus conferring a greater benefit. Evidence of this hypothesis comes from the SVO, SOV, and mixed SVO-SOV comparisons in Dutch, a language with flexible word-order (Dryer, 2013). By training on a CFG that has both SVO and SOV word-order, the Dutch zero-shot average increases by roughly 3% compared to either word-orders

Monolingual Model	ZH Avg	EN Avg	NL Avg
rand init 100M	48.89	52.12	55.32
SVO-p-cr	48.85	51.80	56.14 [*]
SVO-p	48.72	52.00	54.75
p-cr	47.55	52.03	56.14 [*]
SVO-cr	<u>49.66</u> [*]	52.50 [*]	55.71 [*]
CFG (SVO)	49.29 [*]	52.30 [*]	54.08
CFG (SOV)			54.42
CFG (SVO-SOV)			55.77 [*]
protein	48.16	52.62 [*]	54.42
crystal	49.11 [*]	51.98	55.80 [*]
shuffle-dyck	49.29 [*]	51.95	<u>56.20</u> [*]

Table 4: Monolingual zero-shot evaluation results, trained on Chinese (ZH), English (EN), or Dutch (NL) in stage 2. Best models are underlined, and those that beat or tie with the random init baseline are marked with a ^{*}.

alone.

In sum, we find that when the language requirements of stage 2 are reduced, the benefits of our UG-induction method through stage 1 training emerge more clearly. Future research may consider exploring whether the benefits of stage 1 UG-induction emerges in bilingual models, or if the induction is best limited to monolingual models when the token budget is capped at 100M tokens.

6 Discussion

For multilingual models, our “UG” pretraining confers a benefit on non-grammar-focused zero-shot evals but not grammar ones, perhaps still suffering from the 30M loss of natural language data. When the model is given additional training time via fine-tuning (a “stage 3” of training), our MissUniverse models catch up and surpass the baselines, even 100M. When “stage 3” isn’t available (i.e., zero-shots), reducing the number of input sources in stage 1 is not enough to surpass what would have

been gained by just giving the model an extra 30M tokens of natural language. However, if the model is required to only acquire one language, then again we see the benefits of stage 1 training.

6.1 Universal Grammar

While we refer to a Universal Grammar style biasing throughout this work, we do not imply that our CFG, crystals, and protein mixes are actually representative of the UG in humans, if one exists. Instead, this work explores what a UG-like bias could look like in artificial neural networks using synthetic language-like sequences and biological data sources that demonstrate properties similar to Merge and Move (e.g., protein folding). With this biasing, we achieve high performance in non-grammatical tasks, monolingual settings, and fine-tuning on downstream tasks.

However, the success of non-language sources for network biasing raises some interesting questions regarding the human brain bias for language. Perhaps the theorized human Universal Grammar may also be better described as a primed state for generalized downstream learning, akin to our stage 1 pre-trained models, rather than an explicit framework of language patterns.¹³

6.2 Future Steps

There are several avenues for future looking research. One, exploring what about each stage 1 source confers which benefits. We already postulate that the CFG may confer more benefits when its word order is aligned to the target language, but further investigation is needed. Likewise, which grammatical operations did the stage 1 sources help with, and did it hurt? Why does protein perform well for Dutch but poorly for Chinese?

In continued regards to the biological sources, it is unclear whether the performance boost often seen by adding crystals to proteins is due to additional structural complexity, or whether the crystals merely had a regularization effect.

Other aspects of language modeling and acquisition, like tokenizer adaptation from stage 1 to stage 2, incorporating a masked language modeling task, or dynamically changing task objectives to more closely mimic human cognitive development (Thelen and Smith, 1994), we similarly believe are promising future avenues for investigation.

¹³After all, if these patterns occur in other biological sequences then perhaps UG is less about language than it is about truly universal patterns.

7 Conclusions

It has already been demonstrated in the literature that LLMs share grammatical representations across multiple languages (Starace et al., 2023; Wendler et al., 2024; Zhao et al., 2024; Brinkmann et al., 2025). And Dhaliwal et al. (2026) find that adding additional languages to training improves model performance, particularly for low-resource languages. We hypothesize that by training on our data mixes, which are not derived from any human language but which never the less demonstrate structured data with basic syntactic foundations that are shared with human languages, we prime our models to a non-language-specific “UG”-like state that enables the models to better perform on a variety of languages, despite the reduced natural language input, and learn downstream (fine-tuning) tasks better.

Our results align with previous literature finding success in biasing neural networks towards linguistic patterns with careful pretraining of the network (McCoy et al., 2020; Mueller and Linzen, 2023; Papadimitriou and Jurafsky, 2023; Lindemann et al., 2024; Hu et al., 2025; Petty et al., 2025; Aryabumi et al., 2025; Kim et al., 2024).

Despite seeing less natural language data, our results pretraining on different sources of syntax demonstrate just how “universal” a UG really can be, and suggest interesting future research at the intersection of language, networks, and the natural world.

Limitations

As much as we promote here a “Universal Grammar” pretraining, the synthetically generated CFG we created is still somewhat English-centric, with an SVO head-right grammatical structure and using English phonotactic nonce words. While we begin early explorations into word order alternations, future research in this area should still address a wider variety of grammatical structures to more closely match existing human linguistic typology.

Our tests were limited to the languages presented, and it is possible that upon further testing, our results may not hold cross-linguistically with a wider set. Initial tests with Farsi (see Appendix section E) merit further investigation regarding word order influence.

We also were not able to test every possible biological source sequence fully. Virus genome sequences for example, may be an area of potential

study, or may not prove beneficial despite similar syntactic properties in genome sequences.

The results we present here are limited to autoregressive language modeling, and may or may not hold with a masked language modeling task.

Likewise, compute resources and timing constraints limited our ability to test full ablations. Some variations may change our existing comparative rankings, and as discussed briefly in the Appendix, our results here may not hold equally well with smaller (124M parameter) models. We have also not had the chance to fully test models larger than 0.7B parameters, similarly limiting the extent of our findings. We hope to test this method on larger models in the future.

References

- Ben Ambridge and Elena V. M. Lieven. 2014. Child language acquisition: Why universal grammar doesn’t help. *Linguistic Society of America*, 90(3):53–90.
- Christian B. Anfinsen. 1973. [Principles that govern the folding of protein chains](#). *Science*, 181(4096):223–230.
- Luis M. Antunes, Keith T. Butler, and Ricardo Grau-Crespo. 2024. Crystal structure generation with autoregressive large language modeling. *Nature Communications*, 15.
- Viraat Aryabumi, Yixuan Su, Raymond Ma, Adrien Morisot, Ivan Zhang, Acyr Locatelli, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2025. [To code or not to code? exploring impact of code in pre-training](#). In *The Thirteenth International Conference on Learning Representations*.
- Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb R. Glassman, Andy DeGiovanni, Jose H. Pereira, Andria V. Rodrigues, Alberdina A. van Dijk, Ana C. Ebrecht, and 13 others. 2021. [Accurate prediction of protein structures and interactions using a three-track neural network](#). *Science*, 373(6557):871–876.
- Fan Bai, Alan Ritter, and Wei Xu. 2021. [Pre-train or annotate? domain adaptation with a constrained budget](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5002–5015, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. 2025. [Large language models share representations of latent grammatical concepts across typologically diverse languages](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6131–6150, Albuquerque, New Mexico. Association for Computational Linguistics.
- Noam Chomsky. 1957. *Syntactic Structures*. Mouton.
- Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press.
- Noam Chomsky. 1981a. *Lectures on government and binding*. Foris Publications.
- Noam Chomsky. 1981b. *Rules and Representations*. Basil Blackwell.
- Noam Chomsky. 1986. *Knowledge of Language: Its Nature, Origin, and Use*. Praeger.
- Noam Chomsky. 1995. *The Minimalist Program*. MIT Press, Cambridge, MA.
- Eve V. Clark. 2024. *First Language Acquisition*, 4th edition. Cambridge University Press.
- Stephen Crain. 1991. Language acquisition in the absence of experience. *Behavioral and Brain Sciences*, pages 597–612.
- Mehak Dhaliwal, Shashwat Chaurasia, Yao Qin, Dezhi Hong, and Thomas Butler. 2026. [English is not all you need: Systematically exploring the role of multilinguality in llm post-training](#). *Preprint*, arXiv:2604.13286.
- Ken A. Dill, Adam Lucas, Julia Hockenmaier, Liang Huang, David Chiang, and Aravind K. Joshi. 2007. [Computational linguistics: A new tool for exploring biopolymer structures and statistical mechanics](#). *Polymer*, 48(15):4289–4300.
- Matthew S. Dryer. 2013. [Order of subject, object and verb \(v2020.4\)](#). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.
- Nick Evans and Stephen C. Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioural and Brain Sciences*, 32:429–492.
- Kariem Ezzat and Alberto J. Espay. 2023. [Chapter 2 - the allure and pitfalls of the prion-like aggregation in neurodegeneration](#). In Alberto J. Espay, editor, *Precision Medicine in Neurodegenerative Disorders, Part II*, volume 193 of *Handbook of Clinical Neurology*, pages 17–22. Elsevier.
- Xiaomin Fang, Fan Wang, Lihang Liu, Jingzhou He, Dayong Lin, Yingfei Xiang, Kunrui Zhu, Xiaonan Zhang, Hua Wu, Hui Li, and Le Song. 2023. A method for multiple-sequence-alignment-free protein structure prediction using a protein language model. *Nature Machine Intelligence*, 5:1087–1096.

- Michael C. Frank. 2023. [Bridging the data gap between children and large language models](#). *Trends in Cognitive Sciences*, 27(11):990–992.
- Luke Gessler and Nathan Schneider. 2023. [Syntactic inductive bias in transformer language models: Especially helpful for low-resource languages?](#) In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 238–253, Singapore. Association for Computational Linguistics.
- Jill Gilkerson, Jeffrey A. Richards, Judith K. Montgomery, Steven F. Warren, Charles R. Greenwood, D. Kimbrough Oller, John HL Hansen, and Terrence D. Paul. 2017. Mapping the early language environment using all-day recordings and automated analysis. *American Journal of Speech-Language Pathology*, 26(2):248–265.
- Lila R. Gleitman and Elissa L. Newport. 1995. *An Invitation to Cognitive Science*, 2nd edition, volume 1, chapter The Invention of Language by Children: Environmental and Biological Influences on the Acquisition of Language. MIT Press.
- Alain Goriely. 2024. Eighty-six billion and counting: do we know the number of neurons in the human brain? *Brain*, (3):689–691.
- Daniil Gurgurov, Tanja Baeumel, Josef van Genabith, and Simon Ostermann. 2026. [Sparse subnetwork enhancement for underrepresented languages in large language models](#). *Preprint*, arXiv:2510.13580.
- John Hewitt and Christopher D. Manning. 2019. [A structural probe for finding syntax in word representations](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.
- Michael Y. Hu, Jackson Petty, Chuan Shi, William Merrill, and Tal Linzen. 2025. [Between circuits and Chomsky: Pre-pretraining on formal languages imparts linguistic biases](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9691–9709, Vienna, Austria. Association for Computational Linguistics.
- Kevin Maik Jablonka, Qianxiang Ai, Alexander Al-Feghali, Shruti Badhwar, Joshua D. Bocarsly, Andres M. Bran, Stefan Bringuier, L. Catherine Brinson, Kamal Choudhary, Defne Circi, Sam Cox, Wibe A. de Jong, Matthew L. Evans, Nicolas Gastellu, Jerome Genzling, Maria Victoria Gil, Ankur K. Gupta, Zhi Hong, Alishba Imran, and 34 others. 2023. [14 examples of how llms can transform materials science and chemistry: a reflection on a large language model hackathon](#). *Digital Discovery*, 2(5):1233–1250.
- Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. 2013. The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1).
- Muratahan Aykol Bryce Meredig C. Wolverton James E. Saal, Scott Kirklin. 2013. [Materials design and discovery with high-throughput density functional theory: The open quantum materials database \(oqmd\)](#). *JOM*, 65:1501–1509.
- Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, and 1 others. 2023. Phi-2: The surprising power of small language models. *Microsoft Research Blog*.
- Aravind K Joshi, K. Vijay-Shanker, and David Weir. 1990. *Foundational Issues in Natural Language Processing*, chapter The Convergence of Midly Context-Sensitive Gramamr Formalisms. MIT Press, Cambridge, MA, USA.
- Jaap Jumelet, Abdellah Fourtassi, Akari Haga, Bastian Bunzeck, Bhargav Shandilya, Diana Galvan-Sosa, Faiz Ghifari Haznitrana, Francesca Padovani, Francois Meyer, Hai Hu, Julien Etxaniz, Laurent Prévot, Linyang He, María Grandury, Mila Marcheva, Negar Foroutan, Nikitas Theodoropoulos, Pouya Sadeghi, Siyuan Song, and 7 others. 2025. [Babybabbellm: A multilingual benchmark of developmentally plausible training data](#). *Preprint*, arXiv:2510.10159.
- John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Židek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, and 15 others. 2021. Applying and improving alphafold at casp14. *Proteins*, 89(12):1711–1721.
- Gaurav Kamath, Sebastian Schuster, Sowmya Vajjala, and Siva Reddy. 2024. [Scope ambiguities in large language models](#). *Transactions of the Association for Computational Linguistics*, 12:738–754.
- Subhojyoti Khastagir, Kishalay Das, Pawan Goyal, Seung-Cheol Lee, Satadeep Bhattacharjee, and Niloy Ganguly. 2025. [Llm meets diffusion: A hybrid framework for crystal material generation](#). In *Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS)*.
- Najoung Kim, Sebastian Schuster, and Shubham Toshniwal. 2024. [Code pretraining improves entity tracking abilities of language models](#). *Preprint*, arXiv:2405.21068.
- Lars Konieczny. 2000. Locality and parsing complexity. *Journal of Psycholinguistic Research*, 29:627–645.

- Chloe Korade, Elena Nicoladis, and Monique Charest. 2024. Monolingual and bilingual children’s production of complex syntactic structures. *Journal of Child Language*, 51(2):454–469.
- Mark Kretschmann. [link].
- Artur Kulmizev, Vinit Ravishankar, Mostafa Abdou, and Joakim Nivre. 2020. Do neural language models show preferences for syntactic formalisms? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4077–4091, Online. Association for Computational Linguistics.
- Eun-Kyoung Rosa Lee, Sathvik Nair, and Naomi H. Feldman. 2024. A psycholinguistic evaluation of language models’ sensitivity to argument roles. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3262–3274, Miami, Florida, USA. Association for Computational Linguistics.
- Eun-Kyung Lee, Sarah Brown-Schmidt, and Duane G. Watson. 2013. Ways of looking ahead: Hierarchical planning in language production. *Cognition*, 129(3):544–562.
- Cyrus Levinthal. 1968. Are there pathways for protein folding? *J. Chim. Phys.*, 65:44–45.
- Bingzhi Li, Guillaume Wisniewski, and Benoit Crabbé. 2021a. Are Transformers a modern version of ELIZA? Observations on French object verb agreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4599–4610, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Zhongli Li, Qingyu Zhou, Chao Li, Ke Xu, and Yunbo Cao. 2021b. Improving BERT with syntax-aware local attention. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 645–653, Online. Association for Computational Linguistics.
- David Lightfoot. 1989. The child’s trigger experience: Degree-0 learnability. *Behavioral and Brain Sciences*, 12:321–334.
- Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. 2023a. Esm-2 uniref pretraining data. https://huggingface.co/datasets/nvidia/esm2_uniref_pretraining_data.
- Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. 2023b. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130.
- Matthias Lindemann, Alexander Koller, and Ivan Titov. 2024. Strengthening structural inductive biases by pre-training to perform syntactic transformations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11558–11573, Miami, Florida, USA. Association for Computational Linguistics.
- Wei Liu, Ming Xiang, and Nai Ding. 2026. Active use of latent tree-structured sentence representation in humans and large language models. *Nature Human Behavior*, 10:303–316.
- William D. Marslen-Wilson. 1975. Sentence perception as an interactive parallel process. *Science*, 189(4198):226–228.
- Andrea Matarazzo and Riccardo Torlone. 2025. A survey on large language models with some insights on their capabilities and limitations. *Preprint*, arXiv:2501.04040.
- R. Thomas McCoy, Erin Grant, Paul Smolensky, Thomas L. Griffiths, and Tal Linzen. 2020. Universal linguistic inductive biases via meta-learning. *Preprint*, arXiv:2006.16324.
- Aaron Mueller and Tal Linzen. 2023. How to plant trees in language models: Data and architectural effects on the emergence of syntactic inductive biases. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11237–11252, Toronto, Canada. Association for Computational Linguistics.
- Elena Nicoladis and Zixia Jiang. 2018. Language and cognitive predictors of lexical selection in storytelling for monolingual and sequential bilingual children. *Journal of Cognition and Development*, 19(4).
- Dmitry Nikolaev and Sebastian Padó. 2023. Investigating semantic subspaces of transformer sentence embeddings through linear structural probing. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 142–154, Singapore. Association for Computational Linguistics.
- Isabel Papadimitriou and Dan Jurafsky. 2023. Injecting structural hints: Using language models to study inductive biases in language learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8402–8413, Singapore. Association for Computational Linguistics.
- Jackson Petty, Sjoerd van Steenkiste, and Tal Linzen. 2025. How does code pretraining affect language model task performance? *Transactions on Machine Learning Research*.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

- Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. 2021. [Transformer protein language models are unsupervised structure learners](#). In *International Conference on Learning Representations*.
- Ian Roberts. 2017. *The Oxford Handbook of Universal Grammar*, chapter Introduction. Oxford University Press.
- Markus Scheidgen, Alvin Noe Ladines Lauri Himanen 1*, David Sikter, Mohammad Nakhaee, Ádám Fekete, Theodore Chang, Amir Golparvar, José A. Márquez, Sandor Brockhauser, Sebastian Brückner, Luca M. Ghiringhelli, Felix Dietrich, Daniel Lehmberg, Thea Denell, Andrea Albino, Hampus Näsström, Sherjeel Shabih, Florian Dobener, Markus Kühbach, and 9 others. 2023. Nomad: A distributed web-based platform for managing materials science research data. *Journal of Open Source Software*, 8(90).
- Giulio Starace, Konstantinos Papakostas, Rochelle Choenni, Apostolos Panagiotopoulos, Matteo Rosati, Alina Leidinger, and Ekaterina Shutova. 2023. [Probing LLMs for joint encoding of linguistic categories](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7158–7179, Singapore. Association for Computational Linguistics.
- Baris E. Suzek, Yuqi Wang, Hongzhan Huang, Peter B. McGarvey, Cathy H. Wu, and the UniProt Consortium. 2015. [Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches](#). *Bioinformatics*, 31(6):926–932.
- J. Ben Tamo, Daniel Carlander-Reuterfelt, Jonathan Rubin, Oleg Poliannikov, Dezhi Hong, and Mingxian Wang. 2026. [Linguamap: Which layers of LLMs speak your language and how to tune them?](#) In *The Fourteenth International Conference on Learning Representations*.
- Esther Thelen and Linda B Smith. 1994. *A dynamic systems approach to the development of cognition and action*. MIT Press.
- Michael Tomasello. 2009. Ug is dead. *Behavioral and Brain Sciences*, 32:470–471.
- Roger P. G. van Gompel and Martin J. Pickering. 2007. *The Oxford Handbook of Psycholinguistics*, chapter Syntactic Parsing. Oxford University Press.
- Jesse Vig, Ali Madani, Lav R. Varshney, Caiming Xiong, richard socher, and Nazneen Rajani. 2021. [{BERT}ology meets biology: Interpreting attention in protein language models](#). In *International Conference on Learning Representations*.
- Teng Wang, Zhenqi He, Wing-Yin Yu, Xiaojin Fu, and Xiongwei Han. 2025. [Large language models are good multi-lingual learners : When LLMs meet cross-lingual prompts](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4442–4456, Abu Dhabi, UAE. Association for Computational Linguistics.
- Wenkai Wang, Zhenling Peng, and Jianyi Yang. 2022. Single-sequence protein structure prediction using supervised transformer protein language models. *Nature Computational Science*, 2:804–814.
- Alex Warstadt and Samuel R. Bowman. 2024. [What artificial neural networks can tell us about human language acquisition](#). *Preprint*, arXiv:2208.07998.
- Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. [Do llamas work in English? on the latent language of multilingual transformers](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.
- Agata Wolna, Aaron Wright, Colton Casto, Samuel Hutchinson, Benjamin Lipkin, and Evelina Fedorenko. 2026. [The extended language network: Language-responsive brain areas whose contributions to language remain to be discovered](#). *Journal of Neuroscience*.
- Zenan Xu, Daya Guo, Duyu Tang, Qinliang Su, Linjun Shou, Ming Gong, Wanjuan Zhong, Xiaojun Quan, Daxin Jiang, and Nan Duan. 2021. [Syntax-enhanced pre-trained model](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5412–5422, Online. Association for Computational Linguistics.
- Xiulin Yang, Arianna Bisazza, Nathan Schneider, and Ethan Gotlieb Wilcox. 2026. [A unified assessment of the poverty of the stimulus argument for neural language models](#). *Preprint*, arXiv:2602.09992.
- Lei Zhang, Zhaorui Liu, Ben Ni, and Quan Wang. 2026. [Large language models \(llms\) for materials design](#). *Advanced Functional Materials*, 36(30):e25897.
- Shuai Zhang, Wang Lijie, Xinyan Xiao, and Hua Wu. 2022. [Syntax-guided contrastive learning for pre-trained language model](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2430–2440, Dublin, Ireland. Association for Computational Linguistics.
- Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. [How do large language models handle multilingualism?](#) In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Junru Zhou, Zhuosheng Zhang, Hai Zhao, and Shuailiang Zhang. 2020. [LIMIT-BERT : Linguistics informed multi-task BERT](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4450–4461, Online. Association for Computational Linguistics.
- Yoel Zimmermann, Adib Bazgir, Zartashia Afzal, Fariha Agbere, Qianxiang Ai, Nawaf Alampara, Alexander Al-Feghali, Mehrad Ansari, Dmytro Antypov, Amro

Aswad, Jiaru Bai, Viktoriia Baibakova, Devi Dutta Biswajeet, Erik Bitzek, Joshua D. Bocarsly, Anna Borisova, Andres M Bran, L. Catherine Brinson, Marcel Moran Calderon, and 125 others. 2025. [Reflections from the 2024 large language model \(llm\) hackathon for applications in materials science and chemistry](#). *Preprint*, arXiv:2411.15221.

A Context-Free Grammar Details

Tables 5, 6, 7, and 8 provide additional details for the creation of the context-free grammar.

Note, the compatibility matrices and semantic fields outlined do not represent real word meanings or assignments, but imagined groupings. All words are assigned randomly to different syntactic categories and imaginary semantic groups that may be loosely represented by the categories presented in these tables.

Category	Frequency boost
Determiners	10.0x
Nouns	8.0x
Tense/Inflection	6.0x
Verbs	5.0x
Conjunctions	3.0x
Complementizers	2.0x
Prepositions	1.5x
Adjectives/Adverbs	1.5x

Table 5: CFG syntactic categories and their frequency multipliers.

B Ablations

B.1 CFG feature ablations

Eval loss curves on CFG feature ablations showed most efficient training using all developed CFG features (Figure 2).

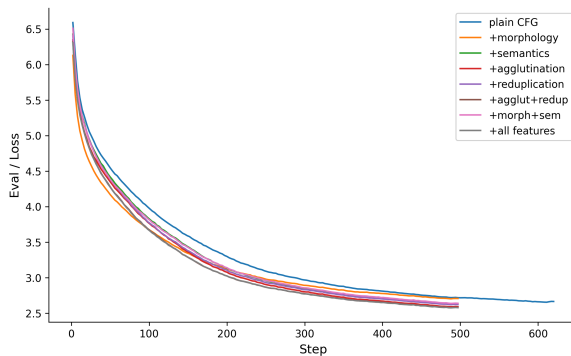


Figure 2: Eval loss curve ablations for CFG features, for stage 2 training with English BabyLM data.

B.2 Data mixing ablations

We also experimented with different proportions of CFG, protein, and crystal mixes. We present a subset of those results for comparison in Appendix F, Table 10.

C BabyLM English: Strict

Like Hu et al. (2025), our pretraining paradigm achieved lower loss faster than natural language training for our 0.7B strict (100M) models (see Figure 1). In particular, our stage 1 UG pretraining paradigm achieved a noticeable improvement in loss curves compared to three baselines: a random initialization baseline (i.e., skipping stage 1 pretraining and only training directly on BabyLM data for the full 1B total tokens seen), training on BabyLM natural language in both stage 1 and stage 2, and the Shuffle-Dyck pretraining language used in Hu et al. (2025), demonstrating the benefits of these rich sources of syntax for UG.

Mixing proteins and crystals in a 1:1 ratio resulted in good performance on the English BabyLM eval pipeline, as did CFG and proteins (Table 10). Crystal data did not appear to differ greatly from Shuffle-Dyck as a solo source for stage 1 training. When mixed in a 2:2:1 ratio with CFG and proteins, it performed well with our 0.7B model, but still not as high as either proteins+crystals or CFG+proteins.

D Small Models (124M params)

D.1 Small model size

Our early experiments were on the largest models we could run with our compute resources—the nearly 0.7B param size of GPT2—and showed high separation between the stage 1 pretraining data conditions. However, we also ran tests on the same 100M dataset size in a smaller 124M parameter GPT2 model. In an attempt to match posted BabyLM GPT2 2025 baselines, we also dropped the sequence packing stage 2 training.¹⁴

In the smaller model, the CFG and protein data sources did not appear to hold the same performance gain from our specialized data mixes (see Figure 3) and Table 10 in Appendix F). Two-stage training on natural language (BabyLM data) in both stages performed the best for the Strict dataset

¹⁴The pulled config for the BabyLM GPT2 baseline indicated a max sequence length of 1024. As such, we kept the long max sequence in place for these 124M tests.

	motion	cognition	creation	interaction	change	perception
living	0.9	0.8	-	0.9	-	0.9
objects	0.6	-	0.8	-	-	-
places	0.7	-	-	-	-	-
abstract	-	0.9	-	-	0.7	-
substances	-	-	-	-	0.8	-

Table 6: CFG noun-verb compatibility matrix for semantic clustering

	physical	evaluative	emotional
living	0.8	-	0.7
objects	0.9	0.8	-
places	0.8	-	-
abstract	-	0.9	0.8

Table 7: CFG noun-adjective compatibility matrix for semantic clustering

Pattern	Application Rate
V D N	0.15
P D N	0.12
D N	0.3
T V	0.21

Table 8: CFG agglutination template. Linking morphemes are [‘’, ‘’, ‘’, ‘a’, ‘o’, ‘i’], with blanks representing bare concatenation.

size,¹⁵ followed by Shuffle-Dyck stage 1 training, then our CFG+protein+crystal mix. Interestingly, in the smaller size, CFG+protein+crystal did better than only CFG+protein, which performed better in the larger model size.

Likewise, the 124M model with the Strict-Small dataset for stage 2 training showed slightly higher performance with CFG+protein+crystal compared to CFG+protein.¹⁶

D.2 Strict-Small

For the Strict-Small test with a model size of 124M parameters, we also pulled the stage 1 training down to 500 steps (approximately 16.5M tokens)¹⁷.

We compared stage 1 training mixes of

¹⁵Our 124M model two-stage training with BabyLM data did still adhere to the 100M token budget.

¹⁶We do not include comparisons with BabyLM in both stages for the Strict-Small exploration because we did not limit stage 1 training to only 3M tokens, thus BabyLM in both stages would have just been additional training beyond the 10M English token budget.

¹⁷Since our stage 1 training is already over the Strict-Small challenge token budget, we did not submit this version to the formal challenge

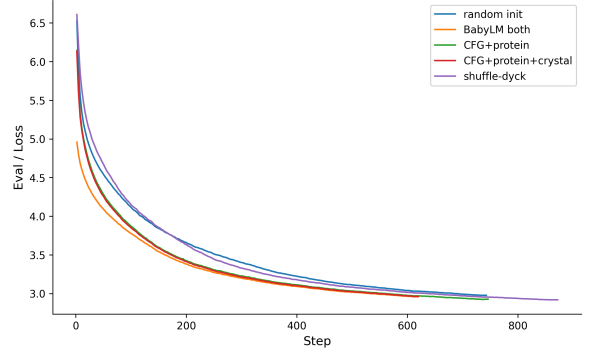


Figure 3: Eval loss curves comparing stage 1 data mixes for the 124M param model size for strict track (100M).

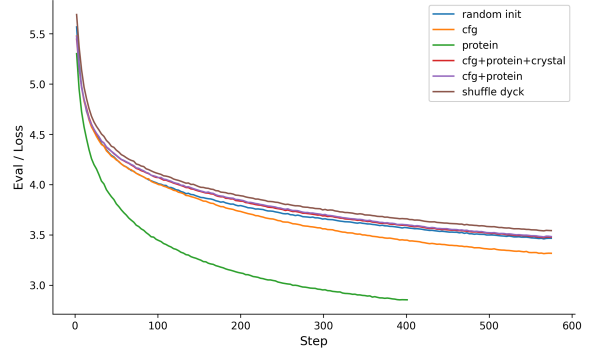


Figure 4: Eval loss curves comparing stage 1 data mixes for the 124M model size on the strict-small 10M dataset.

CFG, proteins, shuffle dyck, CFG+protein, CFG+protein+crystals, and random init (i.e., no stage 1 training).

For this small model and dataset size, stage 1 training only on proteins performed dramatically better in the eval loss curve during training (see Figure 4), but not noticeably better in eval results (Table 10 in Appendix F). We saw somewhat similar behavior in ablations with our full 0.7B model on the larger Strict dataset, where the loss curves looked good for protein only but were not borne out in eval results (though the difference in loss curves was not so stark in the larger model ablation tests). From this we emphasize the need to look

closely at downstream eval in addition to rapid loss improvement.

As a further note on these Strict-Small tests, in keeping with our goal of creating a theoretical “UG” network state before language training begins, we did not limit stage 1 training to 30% of a total 10M token budget. Instead, stage 2 only was considered the language acquisition test for the small network, and used the entire 10M budget. This setup more closely mimics stage 1 as analogous to a “hardcoded” or “biological” bias towards syntactic structures.

D.3 Discussion

Our 124M models performed several points lower than our 0.7B parameter models for the strict track. One explanation for improved performance of 0.7B param model compared to 124M param model when trained on our UG-inducing datamixes is the complexity of the data and the models’ ability to extract said linguistic data. [Matarazzo and Torlone \(2025\)](#) report the Scaling Law of LLMs, noting that as models increase in parameter size, so do their linguistic capabilities as the greater number of parameters enables the model to better capture and model the linguistic relations within the data. As such, larger models often achieve greater linguistic performance than their smaller counterpart models ([Matarazzo and Torlone, 2025](#)), a finding backed by smaller models performing worse than their larger counterparts on multilingual tasks ([Wang et al., 2025](#); [Gurgurov et al., 2026](#)).

In light of this, we postulate that the larger number of parameters in the 0.7B sized GPT2 model better enables the model to take advantage of the shared UG principles in the data mixes in the first stage, then focus on vocabulary and knowledge acquisition in stage 2.

The higher success with larger models vs. smaller ones show we still have progress to be made with regards to network size and shape, considering that current LLMs have 1 trillion or more parameters ([Kretschmann](#)). Compare this to the human brain, which has between 61B and 99B neurons ([Goriely, 2024](#)), of which, only 5% of gray matter is used in language processing ([Wolna et al., 2026](#)).

E Additional Languages

We began investigating additional languages from the BabyBabel datasets ([Jumelet et al., 2025](#)),

choosing Farsi¹⁸ as it has a canonically SOV word order and fits nicely with the word order exploration. However, as seen in Table 9, it does not follow expected patterns with an SOV CFG.

Monolingual Model	ZH Avg	EN Avg	NL Avg	FA Avg
rand init 100M	48.89	52.12	55.32	50.11
SV0-p-cr	48.85	51.80	56.14*	49.10
SV0-p	48.72	52.00	54.75	49.56
p-cr	47.55	52.03	56.14*	49.82
SV0-cr	<u>49.66*</u>	52.50*	55.71*	49.76
CFG (SV0)	49.29*	52.30*	54.08	49.47
CFG (SOV)			54.42	48.49
CFG (SV0-SOV)			55.77*	48.77
protein	48.16	<u>52.62*</u>	54.42	48.93
crystal	49.11*	51.98	55.80*	<u>50.34*</u>
shuffle-dyck	49.29*	51.95	<u>56.20*</u>	

Table 9: Zero-shot evaluation results for monolingual models trained on Chinese (ZH), English (EN), Dutch (NL), or Persian/Farsi (FA) in stage 2. Best models are *underlined*, and those that beat or tie with the random init baseline are marked with a *.

While Dutch shows a performance boost from a CFG word order well aligned with its own distribution, Farsi poses a challenge to this explanation. It is the SVO word-order that outperforms the other two CFGs, despite Farsi being canonically SOV. Perhaps the child-directed language included in the dataset includes instances of SVO order, though this requires a thorough investigation of the relative prevalence of each word-order in the Farsi BabyBabel data to answer.

F Full Tables

¹⁸We would like to acknowledge and thank Zahra Mirrazi for advising us on the syntax of Farsi in various domains.

Model (0.7B params Strict)	Text Avg	BLiMP	BLiMP Supp.	EWoK	Entity Tracking	Comps
CFG only	50.172	67.07	61.64★	51.02	18.48★	52.65
protein only	49.806	65.85	61.84★	50.18	18.94★	52.22
crystal only	50.88★	66.92	64.12★	52.72★	17.48★	53.16
CFG+protein	50.794	66.63	63.45★	52.43★	18.26★	53.2
protein+crystal	51.376	67.91	<u>64.85</u>	52.5★	18.12★	53.5
CFG+crystal	51.024★	68.02	62.95★	52.07★	18.39★	53.69
CFG2+protein1+crystal1	49.174	64.04	60.83★	50.4	17.33★	53.27
CFG2+protein2+crystal1	50.814	65.97	62.35★	<u>52.88</u>	<u>19.39</u>	53.48
BabyLM both	48.188	63.19	57.93	50.4	18.97	50.45
random init	50.844	<u>71.06</u>	60.3	52.04	16.29	<u>54.53</u>
Model (124M params Strict)	Text Avg	BLiMP	BLiMP Supp.	EWoK	Entity Tracking	Comps
CFG+protein	49.234★	63.9	57.34★	52.55	17.96★	54.42
CFG+protein+crystal	49.27★	63.09	58.53★	52.11	18.35★	54.27
shuffle dyck	48.56	64.26	56.75★	51.11	16.25	54.43
BabyLM both	50.444	64.66	<u>59.96</u>	<u>53.41</u>	<u>19.45</u>	54.74
random init	49.038	<u>65.38</u>	54.16	52.88	17.85	<u>54.92</u>
Model (124M params Strict-small)	Text Avg	BLiMP	BLiMP Supp.	EWoK	Entity Tracking	Comps
CFG+protein	47.1	61.3	57.97★	49.4	17.11	49.72
CFG+protein+crystal	47.84	62.28	57.97★	51.34★	17.59	50.02★
shuffle dyck	46.794	62.85	53.14	<u>51.37</u>	16.74	49.87
random init	47.808	<u>63.94</u>	57.08	50.2	<u>17.89</u>	49.93
CFG only	47.232	60.94	<u>58.62</u>	49.73	16.8	50.07★
crystal only	47.23	62.6	56.91	50.09	16.73	49.82
protein only	46.512	60.45	54.06	49.72	17.81	<u>50.52</u>

Table 10: Data mix eval results (0.7B param model + 100M strict track top; 124M param model + 100M strict track middle; 124M param model + 10M strict-small track bottom). The random init comparison ran over the full English data (100M or 10M, depending). For the 0.7B param strict models, two versions of the cfg+protein+crystal mix are included: one with a ratio of 2:1:1, and the other with 2:2:1. All other instances of cfg+protein+crystal are with the 2:2:1 ratio. Best models are underlined, and those that outperformed random init are marked with a ★.

Multilingual Model (NL Evals)	Avg	POS(NL)	MNLI(NL)	SIB200(NL)	Belebele(NL)	ARC(NL)	TruthfulQA(NL)	BMLama(NL)	INCLUDE(NL)	
random init 100M-multi	37.79†	92.04†	43.86	56.00	22.73†	27.71†	16.07	18.05	25.89†	
random init 70M-multi	36.63	91.82	43.92†	56.50★	21.59	23.75	18.75★	19.78★	16.96	
MissUniverse-c-p-cr	37.18†	91.26	45.95†★	52.50	21.02	25.83†	16.96★	12.67	31.25†★	
MissUniverse-c-p	36.85†	91.84†	43.92†★	48.00	20.45	25.83†	17.86★	17.47	29.46†★	
MissUniverse-p-cr	38.27†★	92.50†★	40.48	64.50†★	19.89	26.88†	22.32†★	13.74	25.89†★	
MissUniverse-c-cr	38.09†★	92.25†★	42.85	62.00†★	20.45	24.79†	20.54†★	16.80	25.00†	
Multilingual Model (ZH Evals)	Avg	POS(ZH)	MNLI(ZH)	XNLI(ZH)	SIB200(ZH)	Belebele(ZH)	ARC(ZH)	TruthfulQA(ZH)	BMLama(ZH)	INCLUDE(ZH)
random init 100M-multi	39.99	88.65†	45.78†	42.75	76.50†	17.61	28.96†	22.32	13.25†	24.11
random init 70M-multi	40.61★	88.20	44.26	43.90★	74.00	23.86★	26.46	27.68★	11.26	25.89★
MissUniverse-c-p-cr	40.61★	87.88	45.38†	44.70†★	75.00†	23.30★	27.08†	20.54	13.08†	28.57†★
MissUniverse-c-p	39.86	88.71†★	46.06†★	43.40★	66.00	25.57†★	24.17	25.00★	13.99†★	25.89†★
MissUniverse-p-cr	40.51★	89.02†★	46.96†★	47.75†★	72.00	26.14†★	22.71	27.68†★	17.14†★	15.18
MissUniverse-c-cr	40.06★	88.94†★	48.48†★	42.95★	70.00	21.02★	23.94	28.57†★	13.41†★	23.21
Multilingual Model (EN Evals)	Avg	POS(EN)	MNLI(EN)	XNLI(EN)	SIB200(EN)	Belebele(EN)	ARC(EN)	TruthfulQA(EN)	BMLama(EN)	INCLUDE(EN)
random init 100M-multi	41.04†	90.17	46.57†	45.40	59.50†	21.02	26.25†	21.43	17.96	
random init 70M-multi	40.60	90.34★	44.31	45.45★	55.50	21.59★	25.42	18.75	23.43★	
MissUniverse-c-p-cr	42.64†★	89.31	44.09	45.80†★	62.50†★	21.02★	26.88†★	38.39†★	13.16	
MissUniverse-c-p	41.94†★	89.95	46.57†★	44.45	62.50†★	24.43†★	24.58	24.11†★	18.96†★	
MissUniverse-p-cr	42.98†★	90.32★	44.99†★	52.50†★	66.00†★	21.02★	21.46	27.68†★	19.87†★	
MissUniverse-c-cr	43.05†★	90.14	44.65†	46.15†★	72.00†★	21.02★	26.25†★	26.79†★	17.38	

Table 11: Multilingual fine-tune evaluation results for Dutch (NL, top), Chinese (ZH, middle), and English (EN, bottom). Averages are calculated for the MissUniverse models, as well as two single-source stage 1 models that outperformed three of the MissUniverse models). Best performers are underlined.

Model	Multilingual Avgs			NL Avgs		ZH Avgs		EN Avgs	
	Fine-tune	Synt-Sem	Other	Synt-Sem	Other	Synt-Sem	Other	Synt-Sem	Other
random init 100M-multi	39.62 †	61.90†	29.14†	67.95†	27.74†	59.06†	30.46	60.71†	29.23†
random init 70M-multi	39.33	61.53	28.89	67.87	26.22	58.79	31.52*	60.03	28.94
MissUniverse-c-p-cr-multi	40.16 †*	61.70†	29.98†*	<u>68.61</u> †*	26.70†	59.32†*	31.26*	59.73	32.39†*
MissUniverse-c-p-multi	39.56 †	61.86†	29.07†	67.88†	26.51†	59.39†*	30.10	60.32†	30.92†*
MissUniverse-p-cr-multi	44.99 †*	<u>63.06</u> †*	30.01†*	66.49	28.87†*	61.24†*	30.14	<u>62.60</u> †*	31.21†*
MissUniverse-c-cr-multi	40.39 †*	62.05†*	<u>30.19</u> †*	67.55	28.26†*	60.12†*	30.03	60.31†	<u>32.69</u> †*

Table 12: Fine-tune evals sub-divided into syntax-semantics benchmarks (POS, MNLI, and XNLI, when relevant) and world-knowledge “other” benchmarks (SIB200, Belebele, ARC, TruthfulQA, BMLama, and INCLUDE, when relevant). Best-performing models are underlined; models that outperform or tie with the 70M baseline are denoted with † while those that outperform the 100M baseline are marked with *.

Multilingual Model (NL Evals)	Avg	BLiMP(NL)	MultiBLiMP(NL)	HellaSwag(NL)	Winogrande(NL)	XStoryCloze(NL)	XCOMPS(NL)	GlobalPIQA(NL)
random init 100M-multi	<u>53.76</u> †	<u>73.75</u> †	<u>87.77</u> †	<u>26.75</u> †	49.88†	47.60	51.41†	39.14†
random init 70M-multi	53.14	72.34	86.25	26.53	49.59	47.99*	51.19	38.10
protein-multi	53.19 †	72.06	84.81	26.46	49.22	48.22†*	51.48†*	40.09†*
shuffle-dyck-multi	53.52 †	70.97	85.63	26.21	49.05	50.15†*	<u>51.51</u> †*	<u>41.14</u> †*
CFG (SOV & SVO)-multi	52.21	68.98	85.59	26.23	<u>50.29</u> †*	49.54†*	50.14	34.70
crystal-multi	52.55	71.94	84.86	26.48	49.30	48.68†*	50.90	35.70
CFG (SVO)-multi	52.40	69.47	83.35	26.30	48.60	48.68†*	50.81	39.61†*
Multilingual Model (ZH Evals)	Avg	BLiMP(ZH)		HellaSwag(ZH)	Winogrande(ZH)	XStoryCloze(ZH)	XCOMPS(ZH)	GlobalPIQA(ZH)
random init 100M-multi	<u>47.47</u> †	<u>74.27</u> †		26.76†	49.38	46.67	52.56	35.15†
random init 70M-multi	47.45	72.32		26.74	51.07*	47.33*	52.58*	34.64
protein-multi	47.42	71.86		26.58	<u>51.28</u> †*	46.59	52.57*	35.65†*
shuffle-dyck-multi	47.45 †	70.73		26.83†*	50.95*	<u>48.30</u> †*	51.71	<u>36.17</u> †*
CFG (SOV & SVO)-multi	47.43	72.24		<u>26.87</u> †*	49.71*	47.45†*	<u>52.61</u> †*	35.70†*
crystal-multi	46.38	72.48†		26.80†*	50.62*	47.21*	51.95	29.21
CFG (SVO)-multi	46.60	69.32		26.52	49.71*	46.98*	51.88	35.17†*
Multilingual Model (EN Evals)	Avg	BLiMP(EN)	MultiBLiMP(EN)	HellaSwag(EN)	Winogrande(EN)	XStoryCloze(EN)		GlobalPIQA(EN)
random init 100M-multi	<u>52.48</u> †	65.30†	<u>83.64</u> †	26.72	<u>51.36</u> †	49.77†		38.08†
random init 70M-multi	51.44	64.46	81.82	26.84*	<u>51.36</u> *	49.54		34.62
protein-multi	51.75 †	<u>66.27</u> †*	81.04	26.90†*	50.29	49.85†*		36.14†
shuffle-dyck-multi	51.08	64.96†	80.13	26.88†*	48.06	49.38		37.07†
CFG (SOV & SVO)-multi	52.20 †	64.50†	80.00	26.98†*	51.28	49.38		<u>41.05</u> †*
crystal-multi	52.35 †	64.36	82.08†	<u>27.16</u> †*	50.70	49.23		40.55†*
CFG (SVO)-multi	51.05	63.68	79.35	26.51	<u>51.61</u> †*	<u>50.00</u> †*		35.12†

Table 13: Multilingual zero-shot evaluation results for Dutch (NL, top), Chinese (ZH, middle), and English (EN, bottom) for single-source stage 1 training. Models appear in order of highest multilingual zeroshot average (top within each language) to lowest. Single-source stage 1 models that outperform all other models in this table and those in Table 1 are underlined. Those that beat or tie with the random init baselines (see Table 1 for random init values) are marked with † for random init (70M) and * for random init (100M).

Multilingual Model (NL Evals)	Avg	BLiMP(NL)	MultiBLiMP(NL)	HellaSwag(NL)	Winogrande(NL)	XStoryCloze(NL)	XCOMPS(NL)	GlobalPIQA(NL)	Piqa-par(NL)	Piqa-nonpar(NL)
sharedtok CFG-multi	0.5305 †	0.7272	0.8683	0.2600	0.4979*	0.4729	0.5155†	0.3718†	0.2136†	0.53†*
sharedtok protein-multi	0.5371 †	0.7262	0.8726†	0.2601	0.4864	0.4853†*	0.5179†	<u>0.4112</u> †*	0.2524†	<u>0.57</u> †*
sharedtok crystal-multi	0.5334 †	<u>0.7525</u>	0.8764†	0.2605	0.4938	0.4807*	<u>0.5227</u> †*	0.3474†	0.1748	0.52†
sharedtok shuffle-dyck-multi	0.5275	0.7315	0.8687	0.2573	<u>0.5095</u> †*	0.4752	0.5182†	0.3323†	0.1845	0.48
sharedtok random init 70M-multi	0.5277	0.7351	0.8713	0.2644	0.5012	0.4814	0.5134	0.3273	0.1845	0.47
sharedtok random init 100M-multi	0.5386 †	0.7362	<u>0.8820</u>	0.2618	0.4971	0.4783	0.5189	0.3959	<u>0.2718</u>	0.52
sharedtok MissUniverse-c-p-cr	0.5299	0.7293	0.8593	0.2609	0.4889	0.4745	0.5243†*	0.3721†	0.1942†	0.55†*
Multilingual Model (ZH Evals)	Avg	BLiMP(ZH)		HellaSwag(ZH)	Winogrande(ZH)	XStoryCloze(ZH)	XCOMPS(ZH)	GlobalPIQA(ZH)	Piqa-par(ZH)	Piqa-nonpar(ZH)
sharedtok CFG-multi	0.4678	0.7092		0.2654†	0.4897*	0.4721	0.5284†	0.3418†*	0.2136*	0.47†*
sharedtok protein-multi	0.4786 †*	<u>0.7516</u> †*		0.2671†*	0.5021*	0.4652	0.5244†	0.3612†*	<u>0.2524</u> †*	0.47†*
sharedtok crystal-multi	0.4750 †*	0.7225		<u>0.2698</u> †*	0.4955*	0.4628	0.5331†	0.3662†*	<u>0.2524</u> †*	0.48†*
sharedtok shuffle-dyck-multi	0.4615	0.6936		0.2666†*	0.4913*	0.4667	0.5188	0.3318†*	0.2136*	0.45†*
sharedtok random init 70M-multi	0.4713	0.7231		0.2659	<u>0.5070</u>	0.4776	0.5226	0.3314	0.2427	0.42
sharedtok random init 100M-multi	0.4721	0.7445		0.2651	0.4889	0.4807	<u>0.5415</u>	0.3120	0.2039	0.42
sharedtok MissUniverse-c-p-cr	0.4703	0.7297†		0.2651	0.4823	0.4737	0.5189	0.3521†*	0.1942	0.51†*
Multilingual Model (EN Evals)	Avg	BLiMP(EN)	MultiBLiMP(EN)	HellaSwag(EN)	Winogrande(EN)	XStoryCloze(EN)		GlobalPIQA(EN)	Piqa-par(EN)	Piqa-nonpar(EN)
sharedtok CFG-multi	0.5185	0.6573	0.8208	0.2664†	0.5004†	0.4892		0.3771	0.1942	<u>0.56</u> †*
sharedtok protein-multi	0.5217	0.6692†	0.8221	0.2629	0.4889	0.4961*		0.3911*	0.2621	0.52*
sharedtok crystal-multi	0.5207	0.6474	0.8247	0.2677†*	0.5054†*	0.4977*		0.3814	0.2427	0.52*
sharedtok shuffle-dyck-multi	0.5167	0.6387	0.8065	0.2647†*	0.4979†	0.4961*		0.3961*	0.2621	0.53*
sharedtok random init 70M-multi	0.5260	0.6633	0.8299	0.2646	0.4930	<u>0.5039</u>		<u>0.4011</u>	0.2621	0.54
sharedtok random init 100M-multi	0.5270	<u>0.6711</u>	<u>0.8390</u>	0.2644	0.5037	0.4930		0.3908	0.2816	0.50
sharedtok MissUniverse-c-p-cr	0.5159	0.6486	0.8272	0.2651†*	0.4938†	0.4992*		0.3614	0.2427	0.48

Table 14: Shared tokenizer evaluation results for multilingual zero-shot evals in Dutch (NL, top), Chinese (ZH, middle), and English (EN, bottom). Averages are calculated from the listed evals. Best performers are underlined, and those that beat random init are marked with a *. Note: these results do not include the finetuning eval tasks.

Multilingual Model (ZH Evals)	Avg	BLiMP(ZH)	HellaSwag(ZH)	Winogrande(ZH)	XStoryCloze(ZH)	XCOMPS(ZH)	GlobalPIQA(ZH)	Piqa-par(ZH)	Piqa-nonpar(ZH)	
random init 70M-ZH	0.4678	0.7172	0.2630	0.4897*	0.4768*	0.5228	0.3373	0.1845	0.49*	
random init 100M-ZH	0.4703 †	0.7241†	0.2671†	0.4798	0.4737	<u>0.5348</u> †	0.3420†	0.2039†	0.48	
CFG (SOV & SVO) 70M-ZH	0.4725 †*	0.6982	0.2690†*	0.4946†*	0.4930†*	0.5193	0.3611†*	0.2621†*	0.46	
CFG (SOV & SVO) 100M-ZH	0.4758 †*	0.7284†*	0.2723†*	0.5078†*	0.4760†	0.5186	0.3517†*	0.2233†*	0.48*	
Multilingual Model (EN Evals)	Avg	BLiMP(EN)	MultiBLiMP(EN)	HellaSwag(EN)	Winogrande(EN)	XStoryCloze(EN)	GlobalPIQA(EN)	Piqa-par(EN)	Piqa-nonpar(EN)	
rand init 70M-EN	0.5074 *	0.6142	0.8013*	0.2702*	0.5062*	0.4814	0.3711*	0.2621*	0.48*	
rand init 100M-EN	0.5068	0.6342†	0.7987	0.2684	0.4979	0.4853†	0.3562	0.2524	0.46	
CFG (SOV & SVO) 70M-EN	0.5063	0.6345†*	0.8013†*	0.2655	0.4889	0.4915†*	0.3562*	0.2524*	0.46*	
CFG (SOV & SVO) 100M-EN	0.5112 †*	0.6429†*	0.8130†*	0.2696*	0.4979*	<u>0.4985</u> †*	0.3455	0.3010†*	0.39	
Multilingual Model (NL Evals)	Avg	BLiMP(NL)	MultiBLiMP(NL)	HellaSwag(NL)	Winogrande(NL)	XStoryCloze(NL)	XCOMPS(NL)	GlobalPIQA(NL)	Piqa-par(NL)	Piqa-nonpar(NL)
random init 70M-NL	0.5297	<u>0.7129</u> *	0.8366	0.2644	0.4938	0.4899	0.5040	0.4061	0.2621*	0.55
random init 100M-NL	0.5376 †	0.7078	<u>0.8494</u> †	0.2661†	<u>0.5161</u> †	0.4923†	<u>0.5155</u> †	<u>0.4161</u> †	0.2621†	<u>0.57</u> †
CFG (SOV & SVO) 70M-NL	0.5264	0.7004	0.8250	0.2640	0.5103†	0.4876	0.5015	0.3961	0.2621†*	0.53
CFG (SOV & SVO) 100M-NL	0.5282	0.7116*	0.8460†	0.2679†*	0.4782	<u>0.4992</u> †*	0.4989	0.3954	0.3107†*	0.48
Multilingual Model (FA Evals)	Avg	BLiMP(FA)	MultiBLiMP(FA)	HellaSwag(FA)	Winogrande(FA)	XStoryCloze(FA)	XCOMPS(FA)	GlobalPIQA(FA)	Piqa-par(FA)	Piqa-nonpar(FA)
random init 70M-FA	0.4755 *	0.6753	0.2642	0.4901	0.5170*	0.5207*	<u>0.3859</u> *	0.2718*	0.50*	
random init 100M-FA	0.4734	0.6792†	<u>0.2669</u> †	0.5033†	0.5139	0.5207†	0.3567	0.2233	0.49	
CFG (SOV & SVO) 70M-FA	0.4705	0.6784†	0.2637	0.4942	0.5178	<u>0.5232</u> †*	0.3459	0.2718†*	0.42	
CFG (SOV & SVO) 100M-FA	0.4806 †*	<u>0.6976</u> †*	0.2663†	<u>0.5058</u> †*	0.5209*	0.5070	0.3858*	<u>0.2816</u> †*	0.49†*	

Table 15: Multilingual zero-shot evaluation results for Dutch (NL, top), Chinese (ZH), English (EN), and additional evals for Persian (FA, bottom) using SOV-SVO stage 1 mix and 70M tokens of natural language as well as 100M to accommodate for the fourth language. Best performers are underlined, and those that beat random init (70M) are marked with a † while random init (100M) winners are noted with *. Note: these results do not include the finetuning eval tasks.

Model (Monolingual ZH)	Avg	BLiMP(ZH)		HellaSwag(ZH)	Winogrande(ZH)	XStoryCloze(ZH)	XCOMPS(ZH)	GlobalPIQA(ZH)	Piqa-par(ZH)	Piqa-nonpar(ZH)
random init 100M-ZH	0.4889	0.7765		0.2749	0.4988	0.4853	0.5423	0.3557	<u>0.2913</u> *	0.42
CFG+protein+crystal-ZH	0.4885	0.7712		0.2749*	0.4806	<u>0.5309</u> *	0.4915	0.3818*	0.2136	<u>0.55</u> *
CFG+protein-ZH	0.4872	0.7756		<u>0.2759</u> *	0.4674	0.4814	0.5520*	0.3707*	<u>0.2913</u> *	0.45*
protein+crystal-ZH	0.4755	0.7880*		0.2721	0.4930	0.4915+	0.5506*	0.3667*	0.2233	0.51*
CFG+crystal-ZH	0.4966 *	0.7948*		0.2716	0.4922	0.4930*	0.5416	<u>0.3862</u> *	0.2524	0.52*
CFG-ZH	0.4929 *	0.7557		<u>0.2759</u> *	<u>0.5128</u> *	0.4892*	<u>0.5528</u> *	0.3714*	0.2427	0.50*
protein-ZH	0.4816	0.7777*		0.2694	0.4918	0.4807	0.5434*	0.3268	0.2136	0.44*
crystal-ZH	0.4911 *	0.7781*		0.2729	0.5012*	0.4884	0.5443*	0.3614*	0.2427	0.48*
shuffle-dyck-ZH	0.4929 *	0.7951*		0.2695	0.4926	0.4845	0.5496*	0.3659*	0.2718	0.46*
Model (Monolingual EN)	Avg	BLiMP(EN)	MultiBLiMP(EN)	HellaSwag(EN)	Winogrande(EN)	XStoryCloze(EN)		GlobalPIQA(EN)	Piqa-par(EN)	Piqa-nonpar(EN)
rand init 100M-EN	0.5212	0.6598	0.8351	0.2672	<u>0.5144</u>	0.5046		0.3462	0.2524	0.44
CFG+protein+crystal-EN	0.5180	0.6665*	0.8247	0.2642	0.4889	0.5031		0.3608*	0.2816*	0.44*
CFG+protein-EN	0.5200	0.6610*	0.8364*	0.2643	0.4880	0.5093*		0.3612*	0.2524*	0.47*
protein+crystal-EN	0.5203	0.6643*	0.8312	0.2638	0.4872	0.5039		0.3712*	0.2524*	<u>0.49</u> *
CFG+crystal-EN	0.5250 *	<u>0.6752</u> *	0.8299	0.2642	0.4988	0.5062*		0.3759*	0.2718*	0.48*
CFG-EN	0.5230 *	0.6590	<u>0.8390</u> *	<u>0.2681</u> *	0.5128	0.4930		0.3661*	0.2621*	0.47*
protein-EN	0.5262 *	0.6654*	0.8325	0.2676*	0.4979	0.5031		<u>0.3907</u> *	<u>0.2913</u> *	<u>0.49</u> *
crystal-EN	0.5198	0.6554	0.8351*	0.2636	0.5128	0.5054*		0.3464*	0.2427	0.45*
shuffle-dyck-EN	0.5195	0.6619*	0.8208	0.2636	0.5004	<u>0.5138</u> *		0.3562*	0.2524*	0.46*
Model (Monolingual NL)	Avg	BLiMP(NL)	MultiBLiMP(NL)	HellaSwag(NL)	Winogrande(NL)	XStoryCloze(NL)	XCOMPS(NL)	GlobalPIQA(NL)	Piqa-par(NL)	Piqa-nonpar(NL)
random init 100M-NL	0.5532	0.7711	0.9091	0.2674	<u>0.5161</u>	0.4930	0.5187	0.3970	0.2039	0.59
CFG+protein+crystal-NL	0.5614 *	<u>0.7977</u> *	0.8945	0.2666	0.4839	<u>0.5101</u> *	0.5317*	0.4452*	<u>0.3204</u> *	0.57
CFG+protein-NL	0.5474	0.7727*	0.8953	0.2672	0.4856	0.4954*	0.5342*	0.3811	0.2621*	0.5
protein+crystal-NL	0.5614 *	0.7786*	0.9163*	0.2651	0.4872	0.4992*	<u>0.5374</u> *	<u>0.4458</u> *	0.2816*	<u>0.61</u> *
CFG+crystal-NL	0.5571*	0.7818*	0.9232*	<u>0.2686</u> *	0.4856	0.4985*	0.5108	0.4309*	0.2718*	0.59*
CFG-NL	0.5408	0.7371	0.9005	0.2665	0.5021	0.4822	0.5155	0.3818	0.2136*	0.55
CFG (SOV)-NL	0.5442	0.7631	0.9099*	0.2679*	0.4889	0.4930*	0.5153	0.3714	0.2427*	0.50
CFG (SOV & SVO)-NL	0.5577 *	0.7945*	<u>0.9245</u> *	<u>0.2712</u> *	0.4740	0.5023*	0.5210*	0.4164*	0.2427*	0.59*
protein-NL	0.5442	0.7416	0.8846	<u>0.2682</u> *	0.4930	0.4830	0.5137	0.4254*	0.3107*	0.54
crystal-NL	0.5580 *	0.7832*	0.9215*	0.2672	0.5054	0.4907	0.5309*	0.4068*	0.2136*	0.60*
shuffle-dyck-NL	0.5620 *	0.7842*	0.9224*	0.2676*	0.5103	0.4938*	0.5101	0.4457*	0.2913*	0.60*
Model (Monolingual FA)	Avg		MultiBLiMP(FA)	HellaSwag(FA)	Winogrande(FA)	XStoryCloze(FA)	XCOMPS(FA)	GlobalPIQA(FA)	Piqa-par(FA)	Piqa-nonpar(FA)
random init 100M-FA	0.5011	0.7634	0.2691	0.5198	0.5070	<u>0.5455</u>	0.4017	0.2233	<u>0.58</u> *	
CFG+protein+crystal-FA	0.4910	0.7716*	0.2661	0.5049	0.5054	0.5259	0.3720	0.2039	0.54	
CFG+protein-FA	0.4956	0.7720*	0.2650	0.4934	0.5147*	0.5370	0.3914	0.2427*	0.54	
protein+crystal-FA	0.4982	0.7744*	0.2680	0.5082	0.4954	0.5369	0.4061*	0.2621*	0.55	
CFG+crystal-FA	0.4976	0.7611	0.2679	0.4810	0.5132*	0.5367	<u>0.4259</u> *	<u>0.2718</u> *	<u>0.58</u> *	
CFG-FA	0.4947	<u>0.7822</u> *	<u>0.2707</u> *	0.4852	<u>0.5163</u> *	0.5370	0.3767	0.2233*	0.53	
CFG (SOV)-FA	0.4849	0.7630	0.2655	0.4860	0.4985	0.5347	0.3618	0.2136	0.51	
CFG (SOV & SVO)-FA	0.4877	0.7466	0.2666	0.4909	0.5101*	0.5301	0.3816	0.2331*	0.53	
protein-FA	0.4893	0.7262	0.2697*	<u>0.5247</u> *	0.5147*	0.5289	0.3718	0.2136	0.53	
crystal-FA	0.5034 *	0.7779*	0.2690	0.5148	0.5054	0.5427	0.4109*	<u>0.2718</u> *	0.55	

Table 16: Zero-shot evaluation results for monolingual models trained on Chinese (ZH), English (EN), Dutch (NL), or Persian/Farsi (FA) in stage 2. Best models are underlined, and those that beat or tie with the random init baseline are marked with a *.

Model (Monolingual NL)	Avg	BLiMP(NL)	MultiBLiMP(NL)	HellaSwag(NL)	Winogrande(NL)	XStoryCloze(NL)	XCOMPS(NL)	GlobalPIQA(NL)	Piqa-par(NL)	Piqa-nonpar(NL)
sharedtok random init 70M-NL	<u>55.17</u>	<u>78.92</u> +	92.11	<u>26.77</u> +	49.55+	49.23+	51.41	38.17	22.33	54.00
sharedtok random init 100M-NL	<u>55.37</u> †	78.46	<u>92.49</u> †	26.60	48.97	48.07	52.92†	40.11†	<u>26.21</u> †	54.00
sharedtok CFG+protein+crystal-NL	<u>55.34</u> †	78.32	92.02	26.58	48.89+	49.15+	52.21†	40.18†+	21.36	59.00†+
sharedtok CFG-NL	<u>55.80</u> †+	78.47+	92.45†	26.61+	49.46+	48.92+	<u>53.51</u> †+	<u>41.17</u> †+	22.33†	<u>60.00</u> †+
sharedtok protein-NL	<u>55.18</u> †	78.74+	92.28†	26.51	48.97+	<u>49.69</u> †+	51.86†	38.23†	18.45	58.00†+
sharedtok crystal-NL	<u>54.92</u>	78.28	<u>92.75</u> †+	26.38	<u>49.96</u> †+	48.61+	51.75†	36.68	21.36	52.00
sharedtok shuffle-dyck-NL	<u>55.02</u>	78.76+	92.11†	26.36	<u>49.96</u> †+	49.54†+	52.16†	36.24	17.48	55.00†+
Model (Monolingual ZH)	Avg	BLiMP(ZH)		HellaSwag(ZH)	Winogrande(ZH)	XStoryCloze(ZH)	XCOMPS(ZH)	GlobalPIQA(ZH)	Piqa-par(ZH)	Piqa-nonpar(ZH)
sharedtok random init 70M-ZH	<u>48.26</u>	77.14		26.96	47.65	48.07	<u>54.56</u> +	35.20+	20.39	<u>50.00</u> +
sharedtok random init 100M-ZH	<u>48.52</u> †	78.13†		27.17†	48.89†	48.76†	53.52	34.62	<u>25.24</u> †	44.00
sharedtok CFG+protein+crystal-ZH	<u>47.52</u>	76.95		27.04†	47.57	47.29	52.60	33.68	21.36†	46.00+
sharedtok CFG-ZH	<u>48.65</u> †+	77.29†		27.01	<u>49.96</u> †+	47.76	54.28+	35.62†+	<u>25.24</u> †+	46.00+
sharedtok protein-ZH	<u>48.45</u> †	<u>78.29</u> †+		27.08†	47.57	48.14†	53.96+	35.68†+	21.36†	<u>50.00</u> †+
sharedtok crystal-ZH	<u>48.94</u> †+	77.23†		27.04†	<u>49.96</u> †+	<u>48.99</u> †+	53.78	<u>36.65</u> †+	23.30†	<u>50.00</u> †+
sharedtok shuffle-dyck-ZH	<u>48.26</u> †	78.19†+		<u>27.38</u> †+	47.57	47.45	54.29+	34.68+	21.36†	48.00+
Model (Monolingual EN)	Avg	BLiMP(EN)	MultiBLiMP(EN)	HellaSwag(EN)	Winogrande(EN)	XStoryCloze(EN)		GlobalPIQA(EN)	Piqa-par(EN)	Piqa-nonpar(EN)
sharedtok rand init 70M-EN	<u>52.83</u>	68.29	84.55+	26.47	49.55	<u>50.54</u> +		37.61	26.21	49.00
sharedtok rand init 100M-EN	<u>53.38</u> †	<u>69.08</u> †	83.90	26.75†	50.62†	49.85		<u>40.08</u> †	<u>28.16</u> †	52.00†
sharedtok CFG+protein+crystal-EN	<u>52.48</u> †	68.02	83.77	26.40	50.70†	48.37		37.61†	26.21†	49.00†
sharedtok CFG-EN	<u>51.33</u>	66.04	83.51	26.54†	<u>51.85</u> †+	49.30		30.73	18.45	43.00
sharedtok protein-EN	<u>52.16</u>	67.45	81.56	26.28	49.88†	49.69		38.11†	26.21†	50.00†
sharedtok crystal-EN	<u>52.81</u>	66.94	84.16+	26.15	49.63†	50.39+		39.61†	26.21†	<u>53.00</u> †+
sharedtok shuffle-dyck-EN	<u>53.12</u> †	66.91	<u>86.10</u> †+	<u>26.87</u> †+	50.62†+	50.15+		38.08†	<u>28.16</u> †+	48.00

Table 17: Shared tokenizer monolingual zero-shot evaluation results for Dutch (NL, top), Chinese (ZH), English (EN). Best performers are underlined, and those that beat random init (70M) are marked with a † while those that beat random init (100M) are marked with a +.

Model (NL Evals)	Avg	ARC(NL)	Belebele(NL)	BMLama(NL)	MNLI(NL)	Sib200(NL)	TruthfulQA(NL)	Include(NL)	POS(NL)
MissUniverse-p-cr	<u>38.27</u>	<u>26.88</u>	19.89	13.74	40.48	<u>64.50</u>	<u>22.32</u>	25.89	<u>92.50</u>
MissUniverse-c-cr	<u>38.09</u>	24.79	20.45	16.80	42.85	62.00	20.54	25.00	92.25
MissUniverse-c-p-cr	<u>37.18</u>	25.83	<u>21.02</u>	12.67	<u>45.95</u>	52.5	16.96	<u>31.25</u>	91.26
MissUniverse-c-p	<u>36.85</u>	25.83	20.45	<u>17.47</u>	43.92	48.00	17.86	29.46	91.84
Model (ZH Evals)	Avg	ARC(ZH)	Belebele(ZH)	BMLama(ZH)	MNLI(ZH)	Sib200(ZH)	TruthfulQA(ZH)	Include(ZH)	POS(ZH)
MissUniverse-p-cr	<u>40.51</u>	22.71	<u>26.14</u>	<u>17.14</u>	46.96	72.00	27.68	15.18	<u>89.02</u>
MissUniverse-c-cr	<u>39.70</u>	23.96	21.02	13.41	<u>48.48</u>	70.00	<u>28.57</u>	23.21	88.94
MissUniverse-c-p-cr	<u>40.61</u>	<u>27.08</u>	23.30	13.08	45.38	<u>75.00</u>	20.54	<u>28.57</u>	87.88
MissUniverse-c-p	<u>39.86</u>	24.17	25.57	13.99	46.06	66.00	25.00	25.89	88.71
Model (EN Evals)	Avg	ARC(EN)	Belebele(EN)	BMLama(EN)	MNLI(EN)	Sib200(EN)	TruthfulQA(EN)		POS(EN)
MissUniverse-p-cr	<u>42.98</u>	21.46	21.02	<u>19.87</u>	44.99	66.00	27.68		<u>90.32</u>
MissUniverse-c-cr	<u>42.10</u>	26.25	21.02	17.38	44.65	<u>72.00</u>	26.79		90.14
MissUniverse-c-p-cr	<u>42.64</u>	<u>26.88</u>	21.02	13.16	44.09	62.50	<u>38.39</u>		89.31
MissUniverse-c-p	<u>41.94</u>	24.58	<u>24.43</u>	18.96	<u>46.57</u>	62.50	24.11		89.95

Table 18: Multilingual finetuning evaluation results for Dutch (NL, top), Chinese (ZH, middle), and English (EN, bottom). Averages are calculated from the listed evals. Best performers are underlined, and those that beat random init are marked with a +.