Investigating Critical Period Effects in Language Acquisition through Neural Language Models

Ionut Constantinescu **Tiago Pimentel** **Ryan Cotterell** **Alex Warstadt**
ETH Zürich, Switzerland
ionut.constantinescu@alumni.ethz.ch tiago.pimentel@inf.ethz.ch
ryan.cotterell@inf.ethz.ch awarstadt@ethz.ch

Abstract

Humans appear to have a critical period (CP) for language acquisition: Second language (L_2) acquisition becomes harder after early childhood, and ceasing exposure to a first language (L_1) after this period (but not before) typically does not lead to substantial loss of L_1 proficiency. It is unknown whether these CP effects result from innately determined brain maturation or as a stabilization of neural connections naturally induced by experience. In this study, we use language models (LMs) to test the extent to which these phenomena are peculiar to humans, or shared by a broader class of language learners. We vary the age of exposure by training LMs on language pairs in various experimental conditions, and find that LMs, which lack any direct analog to innate maturational stages, do not show CP effects when the age of exposure of L_2 is delayed. Our results contradict the claim that CP effects are an inevitable result of statistical learning, and they are consistent with an innate mechanism for CP effects. We show that we can reverse-engineer the CP by introducing a regularizer partway through training to simulate a maturational decrease in plasticity. All in all, our results suggest that L_1 learning on its own may not be enough to induce a CP, and additional engineering is necessary to make language models more cognitively plausible.

1 Introduction

The tension between nature and nurture is central to questions surrounding how humans acquire language. The **Critical Period (CP)** for language acquisition is no exception. Around the onset of adolescence, humans exhibit a loss in ability to acquire a second language through immersion and a tendency not to forget their first language under deprivation (Penfield and Roberts, 1959; Lenneberg, 1967; Johnson and Newport, 1989; Pallier et al., 2003). Scholars of human development have long debated whether these phenomena are predetermined by innately encoded developmental changes in the maturing brain (Penfield, 1965; Chomsky, 1965; Pinker, 1994), or natural consequences of increased experience that any typical statistical learner would be subject to (Elman et al., 1996; Seidenberg and Zevin, 2006; Thiessen et al., 2016).

Until recently, it was difficult to differentiate these two hypotheses; as we could only observe one kind of statistical learner (i.e., humans), we could not identify which properties of its learning process were responsible for its behavior. Recent improvement in neural language modeling upends this state of affairs (Warstadt and Bowman, 2022). **Language Models (LMs)** can learn to simulate many native-like grammatical judgments (Warstadt et al., 2020; Zhang et al., 2021; Hu et al., 2020)—long regarded as one of the main behavioral measures of native speaker knowledge (Chomsky, 1957; Johnson and Newport, 1989)—and, like humans, they acquire this knowledge from unstructured input without the need for negative evidence. However, their learning algorithm and structure differ from those of humans in a number of ways. LMs can thus provide additional information about which phenomena are likely to be typical of general language learners, and which are peculiar to humans.

In this work, we use language models to study the CP for language acquisition, focusing on second language acquisition and first language attrition. Our experiments test for **CP effects**[1] in LMs by training them from scratch on bilingual data, varying only the age of exposure to L_2. We

[1]The term **critical period** is sometimes used to refer to a specific biological construct. For clarity, we use the term **CP effects** to refer to the characteristic observable effects of age of exposure on L_1 and L_2 performance.

Transactions of the Association for Computational Linguistics, vol. 13, pp. 96–120, 2025. https://doi.org/10.1162/tacl_a_00725
Action Editor: Afra Alishahi. Submission batch: 6/2024; Revision batch: 9/2024; Published 1/2025.

test whether, like humans, LMs learn L_2 more easily when exposed to it simultaneously with L_1 from the beginning of training, rather than when exposed to it only after learning L_1. Similarly, we test whether they fail to forget L_1 after extensive training on it. Experimentally, we find that LMs are unlike humans in *both* respects.[2] Thus, our results contradict the view that CP effects are an expected consequence of statistical learning, and they are consistent with (but only provide weak evidence for) the view that the CP in humans is a biologically programmed developmental stage.

The benefits of studying the CP in neural models, however, go beyond just discerning the two hypotheses above. If LMs do differ from humans, it may be useful to attempt to reverse-engineer those learning properties exhibited by humans (Dupoux, 2016). Furthermore, minimizing differences between LMs and humans is a necessary step in enabling their use more broadly as models of human language acquisition (Warstadt and Bowman, 2022). Thus, we also attempt to reverse-engineer a CP by simulating a loss of neural plasticity using **Elastic Weight Consolidation** (**EWC**; Kirkpatrick et al., 2017), a Bayesian regularizer used in machine learning to mitigate catastrophic forgetting. Our experiments show that both of the CP effects emerge in tandem when the model's plasticity is explicitly controlled in this way. Our findings demonstrate the utility of LMs as tools for theories about human language acquisition, and they suggest a path forward to making LMs more developmentally plausible models of human language acquisition.

2 Background: The Critical Period

The proposition that there is a critical period for language learning has long been prominent in language acquisition research (Penfield and Roberts, 1959; Lenneberg, 1967). Discussions around the CP, however, typically cluster a number of related observations; these must be teased apart in order to be properly understood (Singleton, 2005; Mayberry and Kluender, 2018). The critical period can be divided into 3 main phenomena, namely: a CP for L_1 acquisition, a CP for L_2 acquisition, and a CP for L_1 attrition. We focus on the latter two here.[3]

2.1 Critical Period for L_2 Acquisition

CP effects for L_2 acquisition consist of greater difficulty in learning a second language and worse learning outcomes as the age of exposure increases. As humans vary greatly in the beginning of L_2 exposure, this is perhaps the most well-known of the CP phenomena. The effects of age of exposure on phonetics and phonology (i.e., one's accent) are part of folk knowledge and were a key piece of evidence in the first works to propose a neurological mechanism for these (Lenneberg, 1967). Numerous studies also show that age of exposure correlates with worse L_2 performance on morphological and syntactic acceptability judgment tasks (Johnson and Newport, 1989; Hartshorne et al., 2018). While the exact nature and reliability of these effects has been questioned at times (Ioup et al., 1994), the existence of age-of-exposure effects is generally accepted. We refer the reader to several thorough reviews of the relevant evidence (Singleton, 2005; Thiessen et al., 2016; Mayberry and Kluender, 2018).

In the realm of computational learners, no prior work has tested this CP in a controlled manner. In a more general form, however, L_2 acquisition has been studied in depth (Dufter and Schütze, 2020; Chen et al., 2023, *inter alia*). Most related to our work, Oba et al. (2023) trained a number of language models on an L_1 and then fine-tuned these models on both L_1 and L_2; they find that, unlike humans, this two-step training improves LMs' L_2 performance. This already suggests that LMs may not show CP effects for L_2 learning. However, they do make L_2 learning relatively easier by fine-tuning their models on L_1 and L_2 simultaneously within the same bidirectional transformer context, which we contend is not cognitively plausible.

Beyond Oba et al.'s (2023) study, weak evidence against the existence of a CP for L_2 in neural networks is suggested by a large body of work on transfer learning which fine-tunes pretrained neural networks to perform new tasks (e.g., Devlin

[2] We note that our results thus align with the well-known fact that language models: (i) are prone to catastrophic forgetting; and (ii) are good at transfer learning. Our experiments, though, support the novel conclusion that transfer performance (training in L_1 followed by L_2) leads to similar or better results than jointly training on both languages.

[3] Strong evidence of a CP for L_1 acquisition has been demonstrated in late L_1 learners from the deaf community (Mayberry and Fischer, 1989; Newport, 1990). However, as simulating late L_1 exposure requires training on non-linguistic data, we consider it beyond the scope of this work.

et al., 2019; Liu et al., 2019; Driess et al., 2023). These studies indicate that a neural model can achieve superior performance on a fine-tuned task compared to training on it from scratch. However, these studies do not manipulate age-of-exposure while controlling for the total amount of input, and their main focus is on tasks other than language modeling itself, such as classification.

2.2 Critical Period for L_1 Attrition

The CP for L_1 attrition refers to a loss of proficiency in L_1 due to a lack of exposure to it. This phenomenon is largely constrained to earlier ages (Pallier, 2007), as adults who emigrate from their L_1 community do not typically forget their L_1 entirely. However, profound attrition is possible if L_1 exposure ceases during childhood. For example, Pallier et al. (2003) studied Korean-born adoptees in France who had no recognition of their L_1, despite living in Korea for as long as eight years.

In the computational domain, language attrition relates to another large body of work in life-long and continual learning. In short, a large number of works have shown that neural networks are prone to catastrophic forgetting (McCloskey and Cohen, 1989; French, 1999), losing most of their proficiency in their original training domain when fine-tuned on another. Continual learning mitigates catastrophic forgetting through the use of adapters (Houlsby et al., 2019; Pfeiffer et al., 2020), regularizers (Kirkpatrick et al., 2017; Pan et al., 2020), or further training in the original domain.

2.3 Theories

Critical period effects in humans are typically interpreted as evidence that neural plasticity in the language centers of the brain decreases as the brain matures (Newport, 1990). However, the cause of this decrease in plasticity is a matter of debate, with much of the divergence among theories stemming from whether they emphasize **innate** or **experiential** mechanisms as responsible for this decrease.[4]

Innate accounts of the CP argue that this loss in plasticity is driven by properties which are specific to how humans acquire language. Some of these accounts are based on the hypothesis that children—but not adults—are equipped with a specialized language acquisition device such as Universal Grammar (Chomsky, 1965; Newport, 1990). On this view, the CP occurs when Universal Grammar is (wholly or partially) lost, displaced, or dismantled as we age, which would explain why adults struggle with language acquisition (Chomsky, 1965, p. 207; Borer and Wexler, 1987; Bley-Vroman et al., 1988; Schachter, 1988; Pinker, 1994, p. 294).[5] Other innate accounts are not language-specific, especially those with an explicit neurobiological basis. For example, humans (and other mammals; Paolicelli et al., 2011) go through a phase of synaptic pruning peaking in late childhood and adolescence (Huttenlocher, 1979, 1990) during which disused neuronal connections are reduced (Hensch, 2005). Monolingual brains show signs of more extensive pruning than bilingual ones (Mechelli et al., 2004), suggesting that early in life abundant synapses provide the necessary plasticity to acquire a second language with ease, and that later these synapses may be pruned if not yet recruited (De Bot, 2006). Beyond synaptic pruning, other neurobiological process such as myelination (Pulvermüller and Schumann, 1994; Pujol et al., 2006) and lateralization (Lenneberg, 1967) are also correlated with a loss in plasticity as we age.[6]

By contrast, experiential accounts of the CP argue that a loss in plasticity is a consequence of learning itself (Munro, 1986; Elman et al., 1996, p. 283; Ellis and Lambon Ralph, 2000; Zevin and Seidenberg, 2002; Seidenberg and Zevin, 2006; Thiessen et al., 2016; Achille et al., 2019). Early experiments on **connectionist** models found that stages associated with human development sometimes fall out naturally during the training of low-bias neural networks (McClelland, 1989). Connectionist word learning simulations found

[4] With exceptions, most views are quite diverse and many scholars advocate for a nuanced view with multiple causes (e.g., Newport, 1990; Thiessen et al., 2016; Singleton, 2005). Other explanations for CP involve social factors, such as a decrease in willingness to experiment, in motivation to fit into one's community, or in the likelihood of being immersed in the target language (Hartshorne et al., 2018).

[5] Child and adult language acquisition are seen as driven by different mechanisms on this view (Thiessen et al., 2016), supported by evidence that general analytic ability predicts adult—but not child—L_2 learning outcomes (DeKeyser, 2000).

[6] Often, these processes have experiential correlates as well, i.e., their outcomes are modulated by experiences during development (Mechelli et al., 2004; Cheng et al., 2019). This is consistent, however, with the timing and onset of the process being biologically determined.

an effect of age of acquisition on learning outcomes (Ellis and Lambon Ralph, 2000; Zevin and Seidenberg, 2002). Numerous scholars explain this loss of plasticity—sometimes referred to as **entrenchment**—as a natural consequence of the training dynamics of networks that lead to convergence (Munro, 1986; Elman et al., 1996; Ellis and Lambon Ralph, 2000; Seidenberg and Zevin, 2006). As Ellis and Lambon Ralph (2000, p. 1108) argue, in a model with random weights (e.g., after initialization) the activations of individual units tend towards intermediate values, leading to large weight changes, but as training proceeds, the units' activations tend towards extreme values making them less prone to change, even if the prediction loss is large.

3 The Role of LMs in Studying the CP

Computational models have the potential to be a powerful tool for informing debates about language acquisition, as they enable a degree of control over the learning mechanism and environment not possible with human subjects; their relevance to questions about human learning, however, is hampered by their numerous differences from human learners (Warstadt and Bowman, 2022). Nonetheless, there are some theoretical claims that current LMs can provide strong or even conclusive evidence about. Not surprisingly, these models are increasingly being used to test theories of language acquisition (McCoy et al., 2020; Lavechin et al., 2021; Wilcox et al., 2023; Warstadt et al., 2023). Language models can, for example, refute some poverty of the stimulus claims by providing existence proofs about language learnability (Clark and Lappin, 2011, p. 30).

In general, theories of CP effects are rather diverse and nuanced. However, we can identify two strong claims which are echoed in many of the accounts above and about which LMs can provide evidence: the strong innate claim and the strong experiential claim.

Strong Innate Claim. *Innate learning constraints are necessary to explain critical period effects.*

The strong innate claim is implicit in the argument that the mere existence of CP effects counts as evidence in favor of an innate mechanism like Universal Grammar (see, e.g., Schachter, 1988). This argument depends on the premise that a change in learning ability as extreme as what is seen in L_2 acquisition could not (or would be very unlikely to) arise from a single domain-general learning mechanism. This premise, and thus the argument, is simple to refute by finding a counterexample, that is, an instance of a low-bias learner that does show CP effects. Transformer-based LMs, while not bias-free, have proven to be effective learners for vision (Dosovitskiy et al., 2021), protein folding (Jumper et al., 2021), and many other types of data, suggesting that they are sufficiently domain-general to refute strong innate claim if they do show CP effects.

Strong Experiential Claim. *Critical period effects are a necessary consequence of successful statistical learning.*

The strong experiential claim has been argued to follow from a mathematical understanding of the training dynamics of connectionist networks (Munro, 1986; Ellis and Lambon Ralph, 2000). Seidenberg and Zevin (2006) speak of a **paradox of success**, whereby successful generalization creates the conditions for a loss in plasticity. This claim is similarly simple to refute, by finding a successful connectionist learner that fails to show CP effects.

Studying the CP in LMs serves an additional purpose for our understanding of human language acquisition. Dupoux (2016) argues that reverse-engineering properties of human language acquisition can give insights into the mechanisms behind those properties at the algorithmic or implementational level (Marr and Poggio, 1976). While we endorse this view, and do attempt to reverse-engineer CP effects, our efforts are at the computational level. The resulting models, however, do more closely resemble human learners in the relevant property, which makes results obtained from them more likely to generalize to humans (Warstadt and Bowman, 2022).

4 Research Questions and Methodology

To study the claims above, we now put forward two research questions which we investigate with the help of language models.

RQ 1. *Can we find evidence of a critical period for* L_2 *learning in language models?*

RQ 2. *Can we find evidence of a critical period for* L_1 *attrition in language models?*

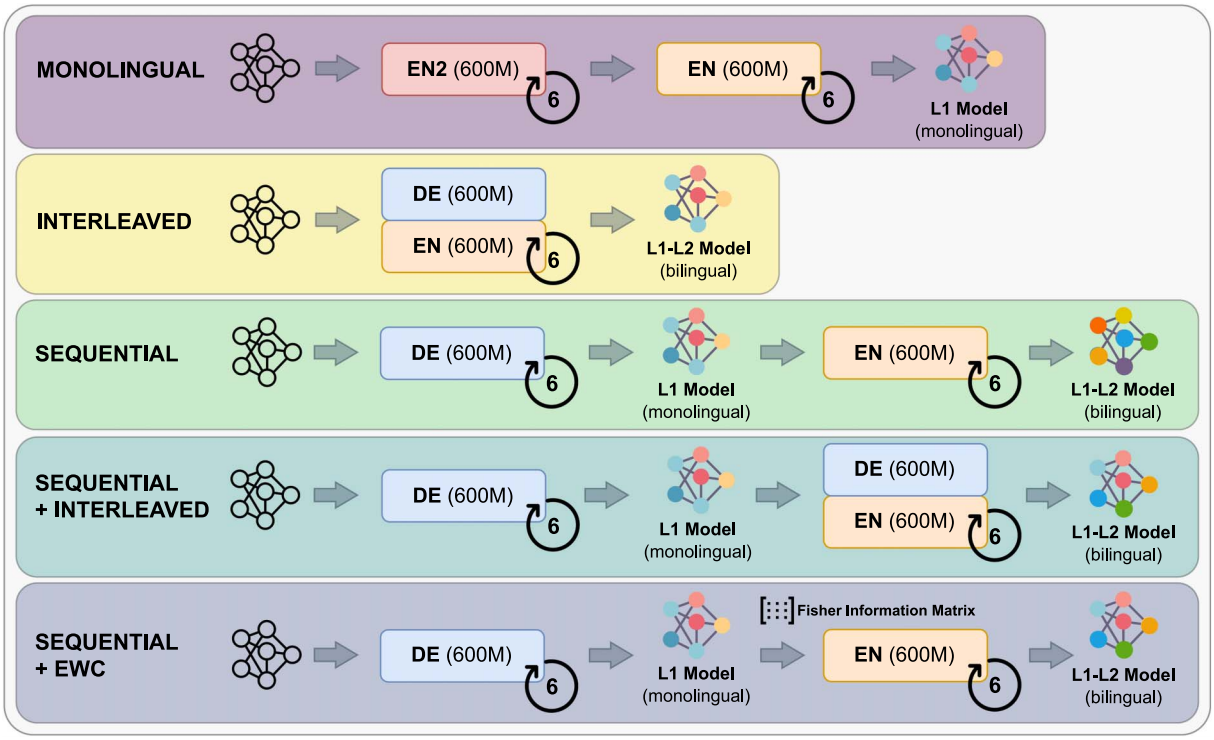

Figure 1: A visualization of the training conditions, using $L_1 = \texttt{de}, L_2 = \texttt{en}, S = 600M, E = 6$.

These are the main questions that we want to investigate, and on which we focus our experiments. We analyze them by first training LMs in various multilingual setups—while altering the ages (epochs) at which L_1 and L_2 are acquired—and then evaluating our models on L_1 and L_2. Importantly, we do *not* make any modifications to the LMs' architecture or learning objectives in these experiments.

Beyond these main questions, we also explore two other research question here.

RQ 3. *Does reducing plasticity in language models induce human-like critical period effects?*

RQ 4. *Are critical period effects in language models dependent on* L_1 *and* L_2*'s similarity?*

Investigating RQ 3 serves a dual purpose. First, it allows us to test whether the critical period effects associated with L_2 acquisition and L_1 attrition arise in tandem as a result of a loss in plasticity, or whether these phenomena can (at least in one case) be decoupled. Second, if we are successful at reverse engineering these CP effects, we obtain a model that more closely resembles human learners and that might be useful for future work. We test this question by evaluating the linguistic performance on both L_1 and L_2 while training a model whose learning objective has an extra regularizer which enforces a reduction in plasticity.

Similarly, RQ 4 also serves a dual purpose. First, it informs us about the relationship between language similarity and CP effects. Second, it implicitly assesses how sensitive our results are to a specific choice of language pair, providing us with a notion of how robust our experiments are to this choice. We test this question by performing the above analysis in a number of language pairs which differ in their similarity.

4.1 Training Conditions

Our goal is to train a language model $\mathbb{L}$ from scratch on pairs of languages (L_1 and L_2) while manipulating two independent variables: (i) the ages (epochs) of exposure to L_1 and L_2, and (ii) the level of programmed plasticity. In this section our focus will be on providing the methodology for variable (i), whereas the methodology for variable (ii) is left to be presented in §4.2.

The obvious way to manipulate age of acquisition in the case of language modeling is to alter the training data schedule. As visualized in Figure 1, we consider five schedules, which we will refer to as ''training conditions'' throughout

the paper. Across all conditions, the datasets remain unchanged (for a given language pair) and the size of the training data per language is kept consistent (we denote this quantity by S). The total number of training iterations per example, also known as epochs, is also a constant number (we refer to this as E). As a consequence, the amount of exposure to L_1 and L_2 is the same across conditions, with the inevitable exceptions of the MONOLINGUAL and SEQUENTIAL-INTERLEAVED conditions.

MONOLINGUAL. This condition simulates a monolingual human learner exposed to only one language during their lifetime. The simplest approach in this condition would be to just train $\mathbb{L}$ for $2 \cdot \mathrm{E}$ epochs on a monolingual dataset. However, this would mean that $\mathbb{L}$ would be trained only on half the number of tokens (S) compared to the other conditions ($2 \cdot \mathrm{S}$). To account for this, we create a second monolingual dataset of the same size and train $\mathbb{L}$ on the two datasets in a sequential manner.

INTERLEAVED. This condition aims to replicate a simultaneous bilingual human learner exposed to two different languages from birth. It is also an implementation of typical multitask learning. We train $\mathbb{L}$ for a total of E epochs on an interleaved bilingual dataset. Throughout training, $\mathbb{L}$ encounters batches of fixed size that alternate between L_1 and L_2. As the bilingual dataset is double in size compared to the monolingual datasets, E epochs provide the same amount of training steps per language as with the other conditions.

SEQUENTIAL. This condition represents the experience of a late L_2 learner who changes linguistic communities, losing exposure to L_1 entirely. It can also be seen as a typical implementation of transfer learning. In this condition, $\mathbb{L}$ is trained for E epochs exclusively on L_1, and then subsequently trained for another E epochs on L_2. The shift from L_1 to L_2 occurs abruptly, with a complete halt in L_1 exposure rather than a gradual transition.

SEQUENTIAL-INTERLEAVED. This condition is closely related to the previous sequential condition, but recreates an experience more common among human bilinguals where L_1 exposure continues during L_2 acquisition. It is also an implementation of one approach to continual learning. After the initial stage of L_1 learning, $\mathbb{L}$ is trained on L_1 interleaved with L_2. We continue to use the same L_1 dataset from the initial training stage.

SEQUENTIAL-EWC. This condition attempts to emulate in LMs an innate reduction in plasticity like that proposed for humans. The L_2 models are trained from the same L_1 checkpoints as for the normal SEQUENTIAL condition, but with EWC regularization added to the loss function. The reduction in plasticity is not progressive, but only changes once, after L_1 has been fully trained.

4.2 Enforcing Plasticity

There are several methods to simulate a computational reduction in plasticity in LMs. We have chosen Elastic Weight Consolidation (Kirkpatrick et al., 2017) due to its popularity and simplicity. EWC introduces a Bayesian-inspired regularization term on a LM's loss partway through training; this term penalizes deviations from a prior distribution over the parameter space (defined in terms of L_1 training), simulating the end of the critical period. The modified loss function is defined as

$$\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}_{\mathrm{L}_2}(\boldsymbol{\theta}) + \lambda \cdot \mathcal{R}_{EWC}(\boldsymbol{\theta}) \qquad (1)$$

We introduce an additional hyperparameter $\lambda \in \mathbb{R}_{\geq 0}$ to control the strength of the EWC regularization term. In the trivial case when $\lambda = 0$, there is no programmed decrease in plasticity. We provide the complete derivation of EWC in Appendix E.

5 Experimental Setup

This section provides a comprehensive description of the experimental setup for this project, including: languages, datasets, model architectures, and evaluation methods.[7]

Languages. Our experiments consist of training LMs from scratch on language pairs. In this work, we rely on `English` (`en`) data for evaluation, as it has a wealth of well-studied resources for assessing language proficiency (see §5). Therefore, we justify the selection of the other languages in this study according to their relatedness to `English`. To reduce the computational overhead, we restrict most of our experiments to only two language pairs: `German`–`English`

[7]The code is released at `https://github.com/iconstantinescu/lm-critical-period`.

and Finnish–English. We choose these languages because they are both (relatively) high-resource languages that are well-represented in our chosen data domains. Furthermore, German (de) is in the same language family as English (Indo-European/Germanic), while Finnish (fi) is unrelated to both (Finno-Ugric).

Nonetheless, to improve the generalizability of our results, we also run one experiment with an extended set of languages as L_1, using English as L_2 (see Experiment 4 in §6). We select languages from various language families (Indo-European (IE) or not) using different scripts (Latin (L) or not). To represent the extreme endpoints, we also use a different corpus of English as the most closely related L_1, and a corpus in a programming language (Java) as the least closely related L_1. The complete list is as follows (from most to least related):

- Same language: English2 (en2)
- Germanic: German (de), Dutch (nl)
- IE, L: Spanish (es), Polish (pl)
- IE, Non-L: Greek (el), Russian (ru)
- Non-IE, L: Finnish (fi), Turkish (tr)
- Non-IE, Non-L: Arabic (ar), Korean (ko)
- Programming language: Java (java)

Datasets. Ideally, we would train on data that closely resembles the kinds of language children encounter during language acquisition. This typically involves natural speech and narratives. Unfortunately, there are no existing datasets in any language that are fully representative of the type and volume of language input a child is exposed to during learning, let alone in a variety of languages. As a consequence, in this work, we construct a customized mix of multilingual training data sourced from three complementary domains: **spoken**, **literature,** and **non-fiction**. We select the *OpenSubtitles* (Lison and Tiedemann, 2016) corpus for the spoken domain, the *Gutenberg* (Gerlach and Font-Clos, 2020) collection for literature, and the *Wikipedia* content for non-fiction.[8] While these choices of datasets are not fully developmentally plausible, we judge them to have a good balance of diversity, quality, and quantity among the limited set of publicly available multilingual datasets. See Appendix A for preprocessing details and Tables 1, 2, and 3 in Appendix B for data statistics.

Models. We run our experiments with both autoregressive and masked language models. To represent these two categories, we consider RoBERTa (Liu et al., 2019), an encoder-only transformer trained to predict masked tokens, and GPT-2 (Radford et al., 2019), a decoder-only transformer trained to predict next tokens. We rely on implementations from the HuggingFace Transformers library (Wolf et al., 2020), namely, roberta-base with 125M parameters and gpt2 with 137M parameters. We choose these two models due to their availability, size, and scientific relevance. These models are then trained according to the conditions defined in §4.1. In all three sequential conditions, L_2 training starts from the final L_1 checkpoints, but with a new optimizer.

Hyper-parameters. We run a Bayesian hyper-parameter search using the Weights & Biases Sweeps API (Biewald, 2020) to identify good model configurations for our training data and methodology. We extract several model configurations (see Appendix C) with which we run the various experiments from §6.

EWC Implementation. We estimate the Fisher Information Matrix according to Eq. (9), using $K = 10$ samples per input. The model is trained using the loss in Eq. (1).[9] We choose the regularization strength λ such that L_1 performance matches L_2 performance at the end of training, which we identify experimentally to be $\lambda = 20$ for GPT-2 and $\lambda = 150$ for RoBERTa. We discuss the details and reasoning behind the choice of λ in §6.

Evaluation. We use Perplexity (PPL), BLiMP, and GLUE to assess models' language proficiency throughout our experiments. To account for the differences in tokenization across language pairs, we report the PPL per character.[10]

[8] To perform Experiment 4 with a broader range of languages, we exclude the *Gutenberg* dataset due to insufficient data availability across languages. Further, for the special Java (programming) language, we use an additional corpus called *The Stack* (Kocetkov et al., 2023).

[9] While we use EWC's loss for training, we still report the cross-entropy loss for evaluations to be consistent.

[10] In addition, PPL per character being smaller in magnitude than the more familiar PPL per token, differences between PPL scores *relative to the total magnitude* are smaller as well.

BLiMP (Warstadt et al., 2020) is a dataset of minimal pairs targeting contrasts in acceptability for a variety of grammatical phenomena in English. Evaluation is performed in a zero-shot setting by comparing LM surprisal on a grammatical and ungrammatical sentence. GLUE and superGLUE[11] are compilations of semantic, commonsense, and syntactic tasks (Wang et al., 2018, 2019a). Evaluation is performed by fine-tuning all parameters of the model as well as a randomly initialized classifier MLP. While similar benchmarks (some even using the names BLiMP and GLUE) are available for many of the languages we study, these are inherently different datasets which are not mutually comparable. Thus, throughout this work, we use only `English` as the evaluation language. This is far from optimal from the perspective of linguistic representation, but we prioritize controlled evaluation, rather than obtaining a large set of exploratory results.

6 Experiments

This section describes our four experiments, designed to provide evidence for answering the research questions detailed in §4.

Experiment 1: Regular Training.
$\mathrm{L_1} \in \{\texttt{de}, \texttt{fi}\}$, $\mathrm{L_2} \in \{\texttt{en}\}$, $\mathrm{S} = \mathbf{600M}$, $\mathrm{E} = \mathbf{6}$.

The goal of this setup is to study the CP for $\mathrm{L_2}$ learning, which relates to RQ 1 and RQ 3. We run this experiment on both GPT-2 and RoBERTa models, using `German` and `Finnish` data as a first language and `English` data as a second language. The dataset for each language has the same size of 600 million tokens, a factor which was limited by the availability of the `Finnish` data. The models are trained with a limited budget of 6 epochs per language. This number has been empirically determined (after preliminary exploration) to provide a good trade-off between computational costs and model performance (i.e., the models perform sufficiently well after training for 6 epochs and the learning slows down). Finally, in order to introduce more variability and provide more result samples, the trainings are run with three different configurations ($\mathbf{C_1}, \mathbf{C_2}, \mathbf{C_3}$, see Tables 4 and 5 in Appendix C).

[11] We henceforth use GLUE to refer to a combination of tasks from GLUE and superGLUE (see Appendix D) implemented in the BabyLM evaluation pipeline (Warstadt et al., 2023).

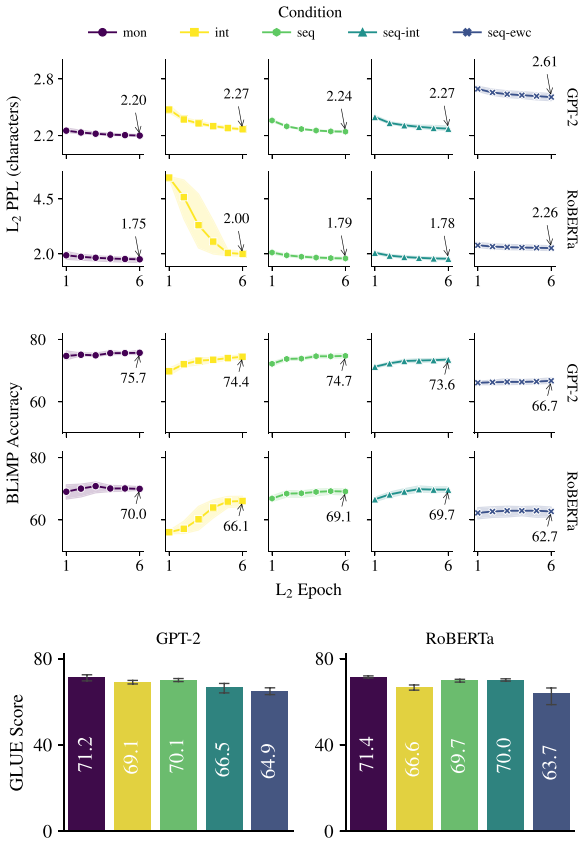

Figure 2: $\mathrm{L_2}$ (`en`) results for regular training (6 epochs). Results are aggregated across model configuration and $\mathrm{L_1}$ (`de` and `fi`). **Top:** PPL per character on $\mathrm{L_2}$ (`en`) during training on $\mathrm{L_2}$. **Middle:** Accuracy on BLiMP during training on $\mathrm{L_2}$. **Bottom:** Performance on GLUE at the end of training.

We illustrate the results for this experiment in Figure 2. In general, the learning patterns through epochs are similar across conditions (except for INTERLEAVED), with variations observed mostly in the final performance. As expected, the MONOLINGUAL performance is the best among all conditions. We notice that the INTERLEAVED condition performs slightly worse than the SEQUENTIAL condition, although the difference is more noticeable in RoBERTa models than in GPT-2 models. Both achieve lower scores than the MONOLINGUAL training (the baseline for native-level language proficiency). Results from the SEQUENTIAL-INTERLEAVED condition differ based on the model architecture. On GPT-2 it is worse compared to SEQUENTIAL results (i.e., keeping $\mathrm{L_1}$ exposure hinders $\mathrm{L_2}$ learning), while on RoBERTa it is better (i.e., keeping $\mathrm{L_1}$ exposure helps $\mathrm{L_2}$ learning). Lastly, the SEQUENTIAL-EWC condition shows clear differences, with much worse $\mathrm{L_2}$ proficiency. Unsurprisingly, the reduction in plasticity through

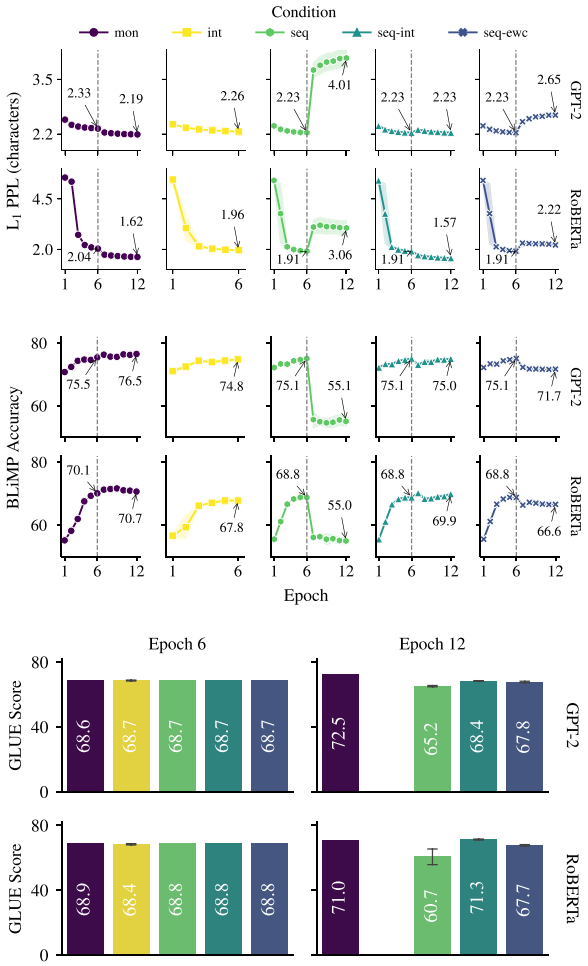

Figure 3: L_1 (en) results when the language order is reversed (6 + 6 epochs). Results are aggregated across L_2 (de and fi). **Top:** PPL per character on the L_1 (en) validation set during training. **Middle:** Accuracy on BLiMP during training. **Bottom:** Performance on GLUE at the end of training on L_1 and L_2.

regularization has a significant (negative) impact on L_2's learning outcome.

Experiment 2: Reversing the Language Order. $L_1 \in \{en\}$, $L_2 \in \{de, fi\}$, $S = 600M$, $E = 6$.

The purpose of this experiment is to study the CP for L_1 attrition, which relates to RQ 2 and RQ 3. The experimental setup differs from the previous one mainly in that the order of the languages is reversed: English is used as a first language and German and Finnish as second languages. This swap allows us to use all three evaluation benchmarks to track L_1 performance across the entire training process. Furthermore, each run is performed with a single configuration (C_1).

The results for this experiment are displayed in Figure 3. In general, we observe a smaller drop in GLUE scores compared to BLiMP scores after exposure to L_1 ceases (i.e., comparing epoch 6 to 12). This is most probably caused by the conceptual difference between evaluations in a zero-shot setting and evaluations that require fine-tuning. When fine-tuning on GLUE, the LM is once again allowed to learn from L_1 data. In the MONOLINGUAL condition, L_1 performance keeps improving until the end of training, even though it slows down in the second stage. In the INTERLEAVED condition, models reach the same level of L_1 and L_2 proficiency (when comparing these results to the ones from Figure 2).[12] The more notable result comes from the SEQUENTIAL learning. In this condition, LMs rapidly lose the knowledge acquired from L_1 learning after L_2 exposure is started. The final L_1 perplexity values indicate that, in the end, the LMs forget almost everything that they have learned before. It looks like the second language completely replaces the first one. However, the loss of L_1 is completely mitigated in the SEQUENTIAL-INTERLEAVED condition. When L_1 exposure is prolonged *without* any reduction in plasticity, models are able to retain all the prior L_1 knowledge.[13]

When a computational regularization method is introduced in the SEQUENTIAL-EWC condition, L_1 knowledge is successfully preserved to a certain degree. However, as we have seen from Figure 2, L_2 learning is also harmed in this case. To explore this trade-off, we vary the λ values and test the models' performance on both L_1 and L_2 at the end of training. The results are illustrated in Figure 5. We see that the regularization strength λ highly influences both L_1 and L_2 learning outcomes. When high λ values are used (strong EWC regularization), L_1 knowledge can be predictably maintained at the initial levels. However, L_2 learning will be almost completely impaired. On the other side, when lower λ values are used (weak EWC regularization), L_2 learning is affected less, but L_1 will not be preserved. It is also noticeable that in this case, there is a higher variance on the final L_1 outcomes. As mentioned in §5, we have selected the λ values for all our experiments as the point where L_1 and L_2 performance is roughly equivalent (intersection lines are marked on the plots). Thus, we do not favor either very strong

[12] Note that the dataset size is 2S in this condition, so the model is only trained for E epochs.

[13] RoBERTa continues learning L_1 during the second stage. We believe this is due to undertraining.

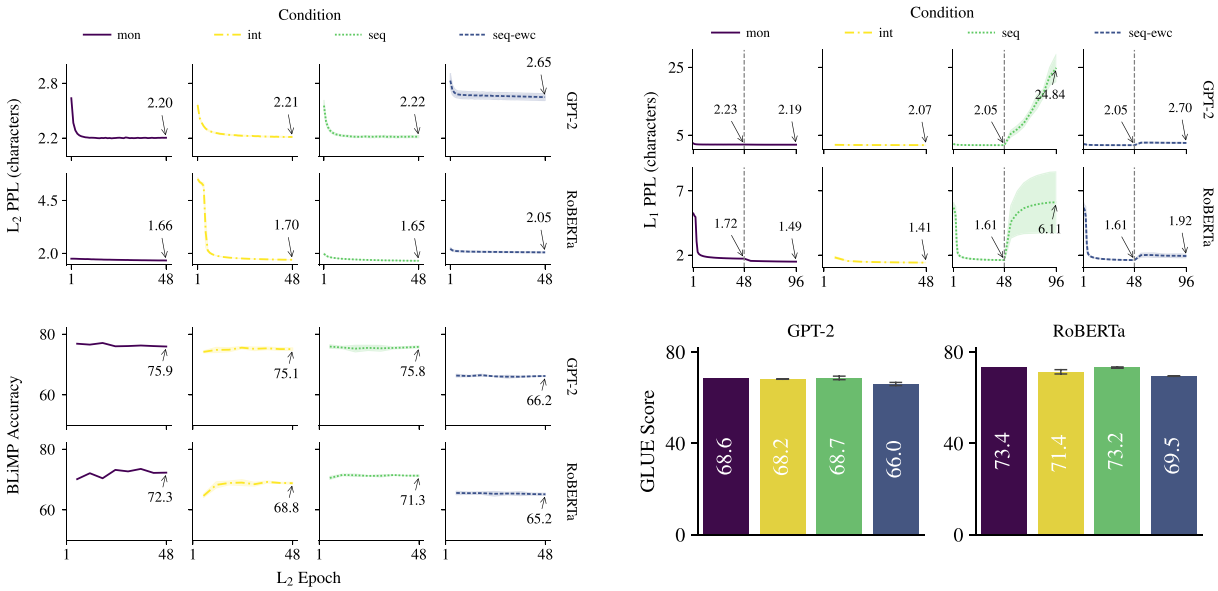

Figure 4: Summary of the L_2 (`en`) evaluation results for the convergence training (48 epochs). Results are aggregated across L_1 (`de` and `fi`). **Top left:** PPL per character on the L_2 (`en`) validation set during training on L_2. **Top right:** PPL per character on the L_1 (`de`, `fi`) validation set during training. **Bottom left:** Accuracy on BLiMP during L_2 training. **Bottom right:** Performance on GLUE at the end of L_2 training.

or very weak regularization, and also to match the behavior exposed by the INTERLEAVED models.

Experiment 3: Training to Convergence.
$L_1 \in \{\texttt{de}, \texttt{fi}\}$, $L_2 \in \{\texttt{en}\}$, S = **600M**, E = **48**.

This experiment is motivated by the observation that CP effects can become stronger with a later age of exposure. We extend the training time for each language (thus postponing the start of L_2 exposure) to 48 epochs, allowing the model weights to better converge. As this is more computationally demanding, we do only one run per condition and we also omit the SEQUENTIAL-INTERLEAVED condition.

The results are provided in Figure 4, which shows that the final results are more uniform across the first three conditions, especially between INTERLEAVED and SEQUENTIAL. We still see a substantial loss in L_1 performance in the SEQUENTIAL condition, indicating that longer L_1 training does not lead to entrenchment, though these PPL scores are for `fi` and `de`, and thus are not directly comparable to the `en` from other experiments. The performance of the INTERLEAVED setting shows a slight improvement compared to regular training, indicating that a longer training duration was beneficial (and necessary for convergence). The learning pattern for the SEQUENTIAL-EWC condition

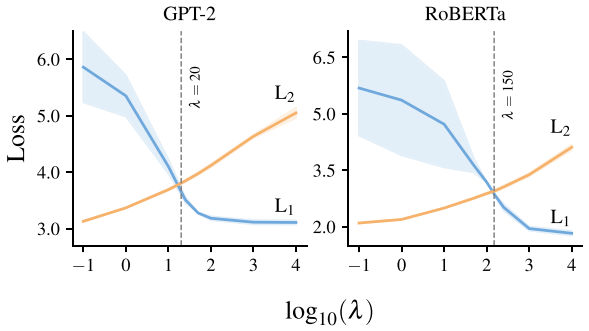

Figure 5: Trade-off between L_1 and L_2 performance (CE) at the end of training as a function of λ (EWC strength). Results are aggregated across L_1 (`de` and `fi`).

does not change, i.e., the final L_2 performance does not significantly improve with additional training. It also appears the EWC regularization does not simply slow down L_2 learning, but rather acts as a lower bound: L_2 knowledge can never improve past a certain point for a given choice of λ.

Experiment 4: Diversifying the Language Pool.
$L_1 \in \{\texttt{ar}, \texttt{de}, \texttt{el}, \texttt{es}, \texttt{fi}, \texttt{ko}, \texttt{nl}, \texttt{pl}, \texttt{ru}, \texttt{tr}, \texttt{java}\}$, $L_2 \in \{\texttt{en}\}$, S = **100M**, E = **6**.

This experiment both addresses RQ 4 and provides more diversity in the results; the latter is

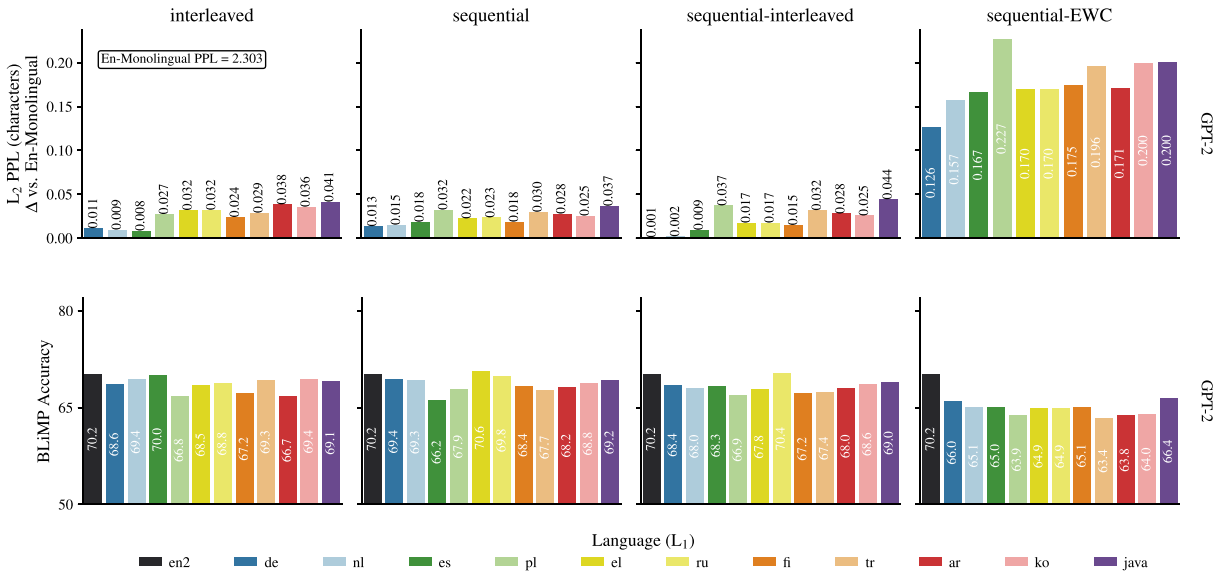

Figure 6: L_2 (en) results for each language pair at the end of training (6 epochs). **Top:** PPL per character on L_2 (en) at the end of training minus PPL per character from the MONOLINGUAL condition. **Bottom:** Accuracy on BLiMP at the end of training.

especially important considering the limited selection of languages for the previous experiments. The main focus here is to find whether the choice of languages or, more concretely, the degree of similarity between L_1 and L_2, has any impact on the CP. For this, we consider a wider range of languages based on their relatedness to English (see §5). To accommodate for the increase in computational demand, we consider only the GPT-2 model architecture (with configuration $\mathbf{C_5}$), we reduce the size of the training data to 100 million tokens, we only train with English as L_2, and we only run PPL and BLiMP evaluations.

We present the findings in Figure 6. Considering L_2 PPL, we find the expected pattern, i.e., PPL generally increases (gets worse) when L_1 is less closely related to it. However, the effect size is quite small (hence we plot the difference with respect to MONOLINGUAL condition): only 0.1–0.3 in most cases, while the total PPL ranges from about 2.3–2.5. However, the BLiMP results do not support the same conclusion. Mostly, BLiMP scores vary seemingly at random, and indeed these differences could reflect random noise due to model initialization. As expected, L_2 performance on BLiMP *is* greatest when L_1 and L_2 are just different corpora of English. But curiously, Java pretraining aids BLiMP performance more than most natural languages. The only condition where we can make the strongest case for an effect is SEQUENTIAL-EWC, which suggests that perhaps relatedness effects exist at the beginning of transfer but are wiped out by extensive L_2 training.

7 Discussion

We first address our four research questions from §4. We then explore our results' implications for the theoretical claims introduced in §3.

7.1 Research Questions

RQ 1. This RQ concerned L_2 CP effects. We find evidence against a CP for L_2 learning in language models. In experiments 1 and 3 (Figures 2 and 4), we consistently find that final performance in L_2 is *worse* (or at least not different) in the INTERLEAVED than in the SEQUENTIAL condition. This is the opposite of the pattern found in humans, where bilingual learners have greater proficiency in L_2 with earlier exposure to it (Johnson and Newport, 1989). Comparing the INTERLEAVED and MONOLINGUAL conditions also informs this question. For humans, early L_2 learners resemble monolinguals in their L_2 performance. However, MONOLINGUAL models outperform INTERLEAVED ones on all three metrics. Looking at the entire learning trajectory, INTERLEAVED models differ markedly from both MONOLINGUAL and SEQUENTIAL models, showing more gradual and delayed improvements.

RQ 2. This RQ concerned the CP for L_1 attrition. Humans show few signs of L_1 attrition after the CP, even if L_1 exposure decreases or ceases. Our results in experiment 2 provide strong evidence against a similar phenomenon in typical LMs. Instead, we find that L_1 performance worsens rapidly and to a large degree in the SEQUENTIAL condition after L_1 exposure ceases. This is expected given the susceptibility of neural networks to catastrophic forgetting. The loss of L_1 proficiency is prevented by continuing L_1 exposure in the SEQUENTIAL-INTERLEAVED condition, but such continued exposure is not necessary in humans.[14]

RQ 3. This RQ concerned the trade-off between L_2 learning and preventing L_1 attrition when explicitly reducing plasticity partway during learning. We find strong evidence for such a trade-off when comparing the SEQUENTIAL-EWC condition to the SEQUENTIAL condition. The value of λ we selected preserved L_1 performance substantially compared to the SEQUENTIAL models, at the cost of harming final performance in L_2. The L_2 learning curves also converge relatively quickly when plasticity is reduced. Our exploration of different values of λ (Figure 5) shows that preserving L_1 to monolingual levels harms L_2 acquisition by roughly 1.5 nats of L_2 performance, but we do not directly compare grammatical performance of our models to that of human late L_2 learners. Thus, our attempt at reverse engineering shows a broadly human-like learning pattern when using EWC, but we cannot say quantitatively and at a high level of granularity whether the result is human-like.

RQ 4. This RQ concerned the impact of language similarity on CP effects. Our results showed that the language family of L_1 and its script has an impact on L_2 learning in the expected way for only a subset of evaluations. Based on earlier findings from Papadimitriou and Jurafsky (2020) and Oba et al. (2023), we had expected L_2 performance would be greater when L_1 is more closely related. However, only our results for PPL support this prior conclusion. Our results for BLiMP do not unless EWC is applied, suggesting that models are ordinarily plastic enough to learn the grammar of L_2 regardless of relatedness, unless plasticity is specifically reduced. We note that Papadimitriou and Jurafsky (2020) reduce plasticity by freezing the model weights before transferring to L_2, but we do not venture an explanation why Oba et al. (2023) seemingly find models to be less plastic than we do.

[14]This ignores self-talk as a potential source of L_1 exposure; self-talk may continue even if external L_1 exposure ceases.

7.2 Theoretical Implications

Our results show that CP effects are not naturally arising in LMs in a typical training regime. At the same time, we are able to suggest a methodology to reverse engineer human-like learning patterns by artificially reducing plasticity later in training. As discussed in §3, results like these are relevant to certain specific claims in the critical period literature. Specifically, they refute the strong experiential claim, which states that all successful learning algorithms will show CP effects, and they are consistent with (but do not provide strong evidence for) the strong innate claim, which states that innate maturational stages are necessary to produce CP effects.

The strong experiential claim incorrectly predicts that our LMs will naturally show CP effects. This may come as a surprise, given this claim was based on previous studies on connectionist models finding evidence for the phenomenon of entrenchment (Munro, 1986; Ellis and Lambon Ralph, 2000). One explanation for this discrepancy in results is the difference in models' capacity. These earlier works trained extremely small models by today's standards, whereas our LMs may be over-parameterized and therefore have sufficient capacity to train to convergence without entrenchment.

On the other hand, the strong innate claim correctly predicts that our LMs will not show CP effects, unless we introduce an innate loss in plasticity. Our introduction of EWC part-way through training is akin to an innate loss in plasticity. However, our experiments are not strong evidence either that the strong innate claim is correct or that EWC is a plausible algorithmic mechanism underlying the loss in plasticity in humans. The possibilities remain that other statistical learners will show CP effects as a natural consequence of experience, and that humans could be one such learner.

Finally, many scholars of human development advocate for a complex explanation of CP effects involving both innate and experiential mechanisms (e.g., Newport, 1990; Thiessen et al., 2016;

Singleton, 2005). We consider this nuanced view to be likely correct, but our results suggest that the role of statistical learning should not be assumed or overstated without evidence from humans or more cognitively plausible models.

7.3 Limitations and Future Work

From a Bayesian epistemological point of view, results from GPT-2 and RoBERTa should affect our priors about general learners and humans in *some* ways, but Transformer LMs are inherently limited as models of human learners. Assuming a sort of Copernican Principle for cognitive modeling, humans and LMs should both be unextraordinary relative to the theoretical class of language learners. So without other evidence or *a priori* reasoning, we should assume they share properties: i.e., the property that we identified in LMs that the ordinary mechanism of learning fails to lead to CP effects. However, there are many differences that could lead us *a priori* to hypothesize differences in how humans and LMs learn. More generalizable evidence could be obtained by considering more cognitively-inspired models and learning algorithms, and learning environments drawing on more developmentally plausible linguistic data and multimodal input (Warstadt and Bowman 2022). For those who want to defend the position that CP effects fall out in humans—but not LMs—from ordinary learning, these or other differences must be identified as the cause. Our experiments seek a compromise given today's resources by choosing architectures and procedures that maximize human-like learning *outcomes*, while still using a transcribed-speech training corpus and human-scale data. Future work should revisit this compromise as better resources for cognitive modeling are developed.

One caveat to the rejection of the strong experiential claim is that the claim only applies to learners that adequately acquire L_1. While this condition is not rigorously defined, one might argue our models do not qualify. For example, our GPT-2 models in the SEQUENTIAL and INTERLEAVED conditions achieve $75 \pm 1\%$ accuracy on BLiMP. This is substantially better than chance (50%) and is comparable to the strongest baseline model from the BabyLM Challenge (Warstadt et al., 2023), but falls well below human agreement with BLiMP (89%). No LMs currently achieve fully human-like performance on BLiMP, and higher performance generally requires compromising on developmental plausibility, though more data-efficient architectures exist (Samuel et al., 2023; Warstadt et al., 2023). We do not expect our results to be qualitatively different if our experiments are run with more effective Transformer-based LMs, but this is something future work should confirm.

We must also acknowledge the theoretical implications of learning rate decay and the optimizer. The learning rate impacts the magnitude of changes to model weights during training, and so it is directly related to the plasticity of the model. Extensive work in machine learning has found that the learning rate should decrease in the later stages of training (see, e.g., Gotmare et al., 2019). Thus, one could take this as evidence that a predetermined loss in plasticity is necessary for successful learning, though the implementation may not necessarily result in CP effects. As our goal with our experiments was to reproduce typical LM training pipelines, we reduce the learning rate in all conditions, but we also restart the learning rate in the sequential conditions at the beginning of L_2 training, as is standard in fine-tuning (Howard and Ruder, 2018). However, this may be interpreted as artificially *increasing* the plasticity of our models, which could contribute to the lack of CP effects. Learning rate schedules, their interaction with successful learning, and their impact on critical periods should be the focus of future work.

We identify several additional avenues for future work: First, the regularization method we use, EWC (Kirkpatrick et al., 2017), can be viewed only as a computational-level model of an innate biological CP; future work could consider other regularizers such as Memory Aware Synapsis (Aljundi et al., 2018) that arguably model what happens in humans at the algorithmic level. Second, there is neurolinguistic evidence for some degree of modularity in how human bilinguals process different languages (Hernandez et al., 2005), but our models use completely shared parameters for languages. Future work can explore LM architectures that encourage or directly build in modularity, such as XLM (Conneau and Lample, 2019) or X-MOD (Pfeiffer et al., 2022). Third, our models learn in a text-only environment, but for humans L_1 and L_2 are both grounded in the same non-linguistic stimuli. Training multimodal models can lead to more realistic simulations, as well as enable testing of CP effects for

L_1 learning, which requires non-linguistic experience to precede L_1 acquisition.

8 Conclusion

There are many obvious ways in which LMs differ from humans when learning language. Our work reveals another important point of divergence, namely, that LMs remain far more plastic later into the learning process than human learners. Even though humans and models differ substantially, comparing the learning trajectories of human learners to those of computational ones tells us something about humans: Those features that we do share are more likely to be natural properties of language learning, while those that we do not are more likely to require idiosyncratic innate mechanisms. Our results provide strong evidence against the hypothesis that CP effects are necessarily induced solely by experience, and they are consistent with, but only provide weak evidence in favor of, the view that innate mechanisms are necessary to explain CP phenomena. It will be important to replicate these results in other artificial learners, including ones which resemble humans more closely in other respects. Our study thus constitutes early progress towards integrating modern LMs into the study of human language acquisition.

Acknowledgments

We thank the editors and reviewers whose feedback has led to substantial improvements in this work. TP and AW are supported by ETH Postdoctoral Fellowships.

References

Alessandro Achille, Matteo Rovere, and Stefano Soatto. 2019. Critical learning periods in deep networks. In *7th International Conference on Learning Representations*.

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. 2018. Memory aware synapses: Learning what (not) to forget. In *15th European Conference*, volume 11207 of *Lecture Notes in Computer Science*, pages 144–161. https://doi.org/10.1007/978-3-030-01219-9_9

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second PASCAL recognising textual entailment challenge. *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

Luisa Bentivogli, Bernardo Magnini, Ido Dagan, Hoa Trang Dang, and Danilo Giampiccolo. 2009. The fifth PASCAL recognizing textual entailment challenge. In *Proceedings of the Second Text Analysis Conference*.

Peter J. Bickel and Kjell A. Doksum. 2001. *Mathematical Statistics: Basic Ideas and Selected Topics*, 2 edition. CRC Press.

Lukas Biewald. 2020. Experiment tracking with Weights and Biases. Software available from wandb.com.

Robert W. Bley-Vroman, Sascha W. Felix, and Georgette L. Loup. 1988. The accessibility of Universal Grammar in adult language learning. *Interlanguage Studies Bulletin*, 4(1):1–32. https://doi.org/10.1177/026765838800400101

Hagit Borer and Kenneth Wexler. 1987. The maturation of syntax. In *Parameter Setting*, pages 123–172. D. Reidel Publishing Company. https://doi.org/10.1007/978-94-009-3727-7_6

Yihong Chen, Kelly Marchisio, Roberta Raileanu, David Ifeoluwa Adelani, Pontus Lars Erik Saito Stenetorp, Sebastian Riedel, and Mikel Artetxe. 2023. Improving language plasticity via pretraining with active forgetting. In *Advances in Neural Information Processing Systems 36*.

Qi Cheng, Austin Roth, Eric Halgren, and Rachel I. Mayberry. 2019. Effects of early language deprivation on brain connectivity: Language pathways in deaf native and late first-language learners of American Sign Language. *Frontiers in Human Neuroscience*, 13:320. https://doi.org/10.3389/fnhum.2019.00320, PubMed: 31607879

Noam Chomsky. 1957. *Syntactic Structures*. MIT Press. https://doi.org/10.1515/9783112316009

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press. https://doi.org/10.21236/AD0616323

Alexander Clark and Shalom Lappin. 2011. *Linguistic Nativism and the Poverty of the Stimulus*. Wiley-Blackwell. https://doi.org/10.1002/9781444390568

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2924–2936. https://doi.org/10.18653/v1/N19-1300

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32*, pages 7057–7067.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190. https://doi.org/10.1007/11736790_9

Kees De Bot. 2006. The plastic bilingual brain: Synaptic pruning or growth? Commentary on Green, et al. *Language Learning*, 56(s1):127–132. https://doi.org/10.1111/j.1467-9922.2006.00358.x

Robert M. DeKeyser. 2000. The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition*, 22(4):499–533. https://doi.org/10.1017/S0272263100004022

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional Transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16×16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations*.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. 2023. PaLM-E: An embodied multimodal language model. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 8469–8488.

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4423–4437. https://doi.org/10.18653/v1/2020.emnlp-main.358

Emmanuel Dupoux. 2016. Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *CoRR*, abs/1607.08723.

Andrew W. Ellis and Matthew A. Lambon Ralph. 2000. Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: Insights from connectionist networks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(5):1103–1123. https://doi.org/10.1037/0278-7393.26.5.1103, PubMed: 11009247

Jeffrey Elman, Elizabeth Bates, Mark H. Johnson, Annette Karmiloff-Smith, Domenico Parisi, and Kim Plunkett. 1996. *Rethinking Innateness: A Connectionist Perspective on Development*. MIT Press. https://doi.org/10.7551/mitpress/5929.001.0001

Robert M. French. 1999. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135. https://doi.org

/10.1016/S1364-6613(99)01294-2, PubMed: 10322466

Martin Gerlach and Francesc Font-Clos. 2020. A standardized Project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *Entropy*, 22(1):126. https://doi.org/10.3390/e22010126, PubMed: 33285901

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL at ACL 2007 Workshop on Textual Entailment and Paraphrasing*, pages 1–9. https://doi.org/10.3115/1654536.1654538

Akhilesh Gotmare, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2019. A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation. In *7th International Conference on Learning Representations*.

Joshua K. Hartshorne, Joshua B. Tenenbaum, and Steven Pinker. 2018. A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition*, 177:263–277. https://doi.org/10.1016/j.cognition.2018.04.007, PubMed: 29729947

Takao K. Hensch. 2005. Critical period plasticity in local cortical circuits. *Nature Reviews Neuroscience*, 6(11):877–888. https://doi.org/10.1038/nrn1787, PubMed: 16261181

Arturo Hernandez, Ping Li, and Brian MacWhinney. 2005. The emergence of competing modules in bilingualism. *Trends in Cognitive Sciences*, 9(5):220–225. https://doi.org/10.1016/j.tics.2005.03.003, PubMed: 15866148

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 328–339. https://doi.org/10.18653/v1/P18-1031

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744. https://doi.org/10.18653/v1/2020.acl-main.158

Peter R. Huttenlocher. 1979. Synaptic density in human frontal cortex – developmental changes and effects of aging. *Brain Research*, 163(2):195–205. https://doi.org/10/0006-8993(79)90349-4, PubMed: 427544

Peter R. Huttenlocher. 1990. Morphometric study of human cerebral cortex development. *Neuropsychologia*, 28(6):517–527. https://doi.org/10.1016/0028-3932(90)90031-I, PubMed: 2203993

Georgette Ioup, Elizabeth Boustagui, Manal El Tigi, and Martha Moselle. 1994. Reexamining the critical period hypothesis: A case study of successful adult SLA in a naturalistic environment. *Studies in Second Language Acquisition*, 16(1):73–98. https://doi.org/10.1017/S0272263100012596

Jacqueline S. Johnson and Elissa L. Newport. 1989. Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, 21(1):60–99. https://doi.org/10.1016/0010-0285(89)90003-0, PubMed: 2920538

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. 2021. Highly accurate

protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589. https://doi.org/10.1038/s41586-021-03819-2

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 252–262. https://doi.org/10.18653/v1/N18-1023

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526. https://doi.org/10.1073/pnas.1611835114, PubMed: 28292907

Denis Kocetkov, Raymond Li, Loubna Ben Allal, Jia Li, Chenghao Mou, Yacine Jernite, Margaret Mitchell, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Dzmitry Bahdanau, Leandro von Werra, and Harm de Vries. 2023. The Stack: 3 TB of permissively licensed source code. *Transactions on Machine Learning Research*, 2023.

Frederik Kunstner, Philipp Hennig, and Lukas Balles. 2019. Limitations of the empirical Fisher approximation for natural gradient descent. In *Advances in Neural Information Processing Systems*, volume 32.

Marvin Lavechin, Maureen de Seyssel, Lucas Gautheron, Emmanuel Dupoux, and Alejandrina Cristia. 2021. Reverse-engineering language acquisition with child-centered long-form recordings. *preprint, PsyArXiv*. https://doi.org/10.31234/osf.io/pt9xq

Eric H. Lenneberg. 1967. The biological foundations of language. *Hospital Practice*, 2(12):59–67. https://doi.org/10.1080/21548331.1967.11707799

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2011. The Winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, volume 46, page 47.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 923–929.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

David Marr and Tomaso Poggio. 1976. From understanding computation to understanding neural circuitry. *AI Memo*, 357.

Rachel I. Mayberry and Susan D. Fischer. 1989. Looking through phonological shape to lexical meaning: The bottleneck of non-native sign language processing. *Memory & Cognition*, 17(6):740–754. https://doi.org/10.3758/BF03202635, PubMed: 2811671

Rachel I. Mayberry and Robert Kluender. 2018. Rethinking the critical period for language: New insights into an old question from American Sign Language. *Bilingualism: Language and Cognition*, 21(5):886–905. https://doi.org/10.1017/S1366728917000724, PubMed: 30643489

James L. McClelland. 1989. Explorations in the microstructure of cognition: Psychological and biological models. In *Parallel Distributed Processing*, volume 2, pages 8–45. MIT Press.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In Gordon H. Bower, editor, *Psychology of Learning and Motivation*, volume 24, pages 109–165. Academic Press. https://doi.org/10.1016/S0079-7421(08)60536-8

R. Thomas McCoy, Robert Frank, and Tal Linzen. 2020. Does syntax need to grow on trees? Sources of hierarchical inductive bias in sequence-to-sequence networks. *Transactions of the Association for Computational Linguistics*, 8:125–140. https://doi.org/10.1162/tacl_a_00304

Andrea Mechelli, Jenny T. Crinion, Uta Noppeney, John O'Doherty, John Ashburner, Richard S. Frackowiak, and Cathy J. Price. 2004. Structural plasticity in the bilingual brain. *Nature*, 431(7010):757–757. https://doi.org/10.1038/431757a, PubMed: 15483594

Paul W. Munro. 1986. State-dependent factors influencing neural plasticity: A partial account of the critical period. In *Parallel Distributed Processing, Volume 2: Explorations in the Microstructure of Cognition: Psychological and Biological Models*, pages 471–502. Bradford Books.

Elissa L. Newport. 1990. Maturational constraints on language learning. *Cognitive Science*, 14(1):11–28. https://doi.org/10.1207/s15516709cog1401_2

Miyu Oba, Tatsuki Kuribayashi, Hiroki Ouchi, and Taro Watanabe. 2023. Second language acquisition of neural language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13557–13572. https://doi.org/10.18653/v1/2023.findings-acl.856

Christophe Pallier. 2007. Critical periods in language acquisition and language attrition. In Barbara Köpke, Monika S. Schmid, Merel Keijzer, and Susan Dostert, editors, *Studies in Bilingualism*, volume 33, pages 155–168. John Benjamins Publishing Company. https://doi.org/10.1075/sibil.33.11pal

Christophe Pallier, Stanislas Dehaene, Jean-Baptiste Poline, Denis LeBihan, Anne-Marie Argenti, Emmanuel Dupoux, and Jacques Mehler. 2003. Brain imaging of language plasticity in adopted adults: Can a second language replace the first? *Cerebral Cortex*, 13(2):155–161. https://doi.org/10.1093/cercor/13.2.155, PubMed: 12507946

Pingbo Pan, Siddharth Swaroop, Alexander Immer, Runa Eschenhagen, Richard E. Turner, and Mohammad Emtiyaz Khan. 2020. Continual deep learning by functional regularisation of memorable past. In *Advances in Neural Information Processing Systems 33*.

Rosa C. Paolicelli, Giulia Bolasco, Francesca Pagani, Laura Maggi, Maria Scianni, Patrizia Panzanelli, Maurizio Giustetto, Tiago Alves Ferreira, Eva Guiducci, Laura Dumas, Davide Ragozzino, and Cornelius T. Gross. 2011. Synaptic pruning by microglia is necessary for normal brain development. *Science*, 333(6048):1456–1458. https://doi.org/10.1126/science.1202529, PubMed: 21778362

Isabel Papadimitriou and Dan Jurafsky. 2020. Learning music helps you read: Using transfer to study linguistic structure in language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6829–6839. https://doi.org/10.18653/v1/2020.emnlp-main.554

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32.

Wilder Penfield. 1965. Conditioning the uncommitted cortex for language learning. *Brain*, 88(4):787–798. https://doi.org/10.1093/brain/88.4.787, PubMed: 5856079

Wilder Penfield and Lamar Roberts. 1959. *Speech and Brain Mechanisms*. Princeton University Press.

Jonas Pfeiffer, Naman Goyal, Xi Victoria Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular Transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495. https://doi.org/10.18653/v1/2022.naacl-main.255

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A framework for adapting Transformers. In *Proceedings of the*

2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 46–54. https://doi.org/10.18653/v1/2020.emnlp-demos.7

Steven Pinker. 1994. *The Language Instinct*. William Morrow and Company. https://doi.org/10.1037/e412952005-009

Jesus Pujol, Carles Soriano-Mas, Hector Ortiz, Nuria Sebastián-Gallés, Josep M. Losilla, and Joan Deus. 2006. Myelination of language-related areas in the developing brain. *Neurology*, 66(3):339–343. https://doi.org/10.1212/01.wnl.0000201049.66073.8d, PubMed: 16476931

Friedemann Pulvermüller and John H. Schumann. 1994. Neurobiological mechanisms of language acquisition. *Language Learning*, 44(4):681–734. https://doi.org/10.1111/j.1467-1770.1994.tb00635.x

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. https://doi.org/10.18653/v1/D16-1264

David Samuel, Andrey Kutuzov, Lilja Øvrelid, and Erik Velldal. 2023. Trained on 100 million words and still in shape: BERT meets British National Corpus. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1909–1929. https://doi.org/10.18653/v1/2023.findings-eacl.146

Jacquelyn Schachter. 1988. Second language acquisition and its relationship to Universal Grammar. *Applied Linguistics*, 9(3):219–235. https://doi.org/10.1093/applin/9.3.219

Mark S. Seidenberg and Jason D. Zevin. 2006. Connectionist models in developmental cognitive neuroscience: Critical periods and the paradox of success. In Yuko Munakata and Mark H. Johnson, editors, *Processes of Change in Brain and Cognitive Development*, pages 585–612. Oxford University Press. https://doi.org/10.1093/oso/9780198568742.003.0025

David Singleton. 2005. The critical period hypothesis: A coat of many colours. *International Review of Applied Linguistics in Language Teaching*, 43(4):269–285. https://doi.org/10.1515/iral.2005.43.4.269

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

Erik D. Thiessen, Sandrine Girard, and Lucy C. Erickson. 2016. Statistical learning and the critical period: How a continuous learning mechanism can give rise to discontinuous learning. *WIREs Cognitive Science*, 7(4):276–288. https://doi.org/10.1002/wcs.1394, PubMed: 27239798

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32*, pages 3261–3275.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355. https://doi.org/10.18653/v1/W18-5446

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations*. https://doi.org/10.18653/v1/W18-5446

Alex Warstadt and Samuel R. Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In Shalom Lappin and Jean-Philippe Bernardy, editors, *Algebraic*

Structures in Natural Language, pages 17–60. CRC Press. https://doi.org/10.1201/9781003205388-2

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34. https://doi.org/10.18653/v1/2023.conll-babylm.1

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392. https://doi.org/10.1162/tacl_a_00321

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641. https://doi.org/10.1162/tacl_a_00290

Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. 2023. Using computational models to test syntactic learnability. *Linguistic Inquiry*, pages 1–44.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1112–1122. https://doi.org/10.18653/v1/N18-1101

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Jason D. Zevin and Mark S. Seidenberg. 2002. Age of acquisition effects in word reading and other tasks. *Journal of Memory and Language*, 47(1):1–29. https://doi.org/10.1006/jmla.2001.2834

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. 2021. When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 1112–1125. https://doi.org/10.18653/v1/2021.acl-long.90

A Data Preprocessing

Cleaning. We remove extra spaces, non-breaking spaces and dataset-specific characters such as dialogue lines and music symbols (in *OpenSubtitles*) and paragraph delimiters (in *Gutenberg*). For the `Java` code extracted from *The Stack*, we remove all docstrings and comments.

Unifying. We downsample the original data sources to align the data quantity and distribution of domains for all languages. A fixed sampling ratio of $2:1:1$ is established for *OpenSubtitles* : *Gutenberg* : *Wikipedia* in order to ensure an even distribution of transcribed and written text within the training data. As mentioned previously, there is an exception to this rule for Experiment 4, where only data from *OpenSubtitles* and *Wikipedia* is sampled in equal parts. To mix the data from different domains while limiting the number of context breaks, large blocks of 10000 lines are uniformly sampled from each dataset, and then randomly shuffled. The resulting unified dataset is then split as follows: 83% train, 8.5% validation, and 8.5% test.

Interleaving. Interleaved datasets are created for each language pair (`en` plus a second language). The same blocks of texts sampled in the previous unifying step are simply interleaved while maintaining their ordering (e.g., L_1 block 1, L_2 block 1, L_1 block 2, L_2 block 2, etc.). In this way, the data from the two languages is presented

in the same order to the model during the INTERLEAVED training as it is during the SEQUENTIAL training.

Size Alignment. We uniformize the training dataset sizes across languages. We quantify dataset sizes in terms of the number of tokens obtained using a BPE tokenizer; see Appendix A. Beyond the fact that tokens are the true unit of input to the LMs, BPE is a compression algorithm, so while the amount of information per word might be highly language-specific, the amount of information per token is more comparable across languages.

Tokenization. We train bilingual byte-Level BPE tokenizers.[15] Given the size of our training datasets, we set a vocabulary size of 32,000 and a minimum frequency of 2. When tokenizing the training data, all the text lines are concatenated and the resulting tokenized dataset is split into fixed-size blocks of 512 tokens which are used as input for both GPT-2 and RoBERTa.

B Data

Dataset	Lang.	Size (GB)	Lines (M)	Words (M)
OpenSubtitles	ar	1.6	39	177
	de	1.2	34	202
	el	6.5	126	650
	en	11.0	316	2112
	es	6.4	213	1144
	fi	1.4	45	191
	ko	0.1	3	8
	nl	3.1	105	600
	pl	6.8	236	1055
	ru	2.2	45	214
	tr	5.1	173	698
Wikipedia	ar	2.1	8	209
	de	6.4	25	931
	el	1.1	2	92
	en	15.0	60	2543
	es	4.5	21	767
	fi	0.8	3	93
	ko	0.8	5	85
	nl	1.9	12	303
	pl	2.0	11	275
	ru	7.4	24	604
	tr	0.7	4	86
Gutenberg	en	19.0	59	3417
	de	0.6	2	103
	fi	0.7	3	92
The Stack	java	3.8	110	337

Table 1: Statistics of the collected data.

[15] With the exception of the MONOLINGUAL condition, which uses a monolingual byte-Level BPE tokenizer instead.

Train Dataset	Size (GB)	Words (M)	Tokens (M)
en	2.2	408	601
en2	2.2	402	592
de	2.4	378	600
fi	2.4	311	596

Table 2: Main experiments' dataset sizes.

Train Dataset	Size (GB)	Words (M)	Tokens (M)
ar	0.6	62	104
de	0.4	63	100
el	0.7	61	102
en	0.4	68	99
en2	0.4	66	96
es	0.4	66	100
fi	0.4	51	100
ko	0.5	47	102
nl	0.4	65	98
pl	0.4	57	104
ru	0.6	54	99
tr	0.4	55	101
java	0.3	30	99

Table 3: Experiment 4 dataset sizes.

C Hyperparameters

Hyperarameter	C_1	C_2	C_3	C_4	C_5
Learning rate ($\times 10^{-3}$)	1.00	1.00	8.00	1.00	1.00
Warmup ratio	7%	9%	10%	7%	7%
Gradient accum. steps	16	32	32	16	4

Table 4: Configurations for GPT-2 models.

Hyperarameter	C_1	C_2	C_3	C_4	C_5
Learning rate ($\times 10^{-4}$)	4.75	3.88	3.90	3.00	4.00
Warmup ratio	5%	10%	9%	1%	1%
Gradient accum. steps	32	32	16	32	1
MLM probability	0.3	0.15	0.3	0.3	0.3

Table 5: Configurations for RoBERTa models.

D Evaluation Tasks

The BLiMP tasks that our models are evaluated on are examplified in Table 7. The subset of GLUE[16] tasks that our models are evaluated on are the following.

CoLA. The Corpus of Linguistic Acceptability consists of around 10K English sentences

[16] https://gluebenchmark.com.

Hyperparameter	Value	
	GPT-2	RoBERTa
n_head	12	12
n_layer	12	12
n_positions	1,024	512
n_embd	768	768
activation_function	''gelu_new''	''gelu''
optimizer	''adamw_hf''	''adamw_hf''
lr_scheduler	''linear''	''linear''
device_train_batch_size	4	8
adafactor	False	False
adam_beta1	0.9	0.9
adam_beta2	0.999	0.999
adam_epsilon	0.00000001	0.00000001
max_grad_norm	1	1
layer_norm_epsilon	0.00001	0.000000000001
weight_decay	0	0
dropout	0.1	0.1
fp16	True	True

Table 6: Fixed hyperparameters for GPT-2 and RoBERTa models.

gathered from published linguistics literature that have been annotated by experts for binary acceptability (grammaticality) judgments (Warstadt et al., 2019). The evaluation metric is Matthews Correlation Coefficient.

SST-2. The Binary Stanford Sentiment Treebank includes 215K unique phrases extracted from the parse trees of movie reviews sentences and that have been fully labeled as having either a positive or negative sentiment by three human judges (Socher et al., 2013). The evaluation metric is Accuracy.

MRPC. The Microsoft Research Paraphrase Corpus contains 5,800 pairs of sentences sourced from a large corpus of news data, each labeled with a binary judgment indicating whether the pair represents a paraphrase or not (Dolan and Brockett, 2005). The evaluation metric is F1 score.

QQP. The Quora Question Pairs is a corpus of over 400K question pairs from the Quora website which are annotated with a binary value denoting whether the questions are paraphrases of each other. The evaluation metric is F1 score.

MNLI. The Multi-Genre Natural Language Inference is a crowd-sourced collection of 433K sentence pairs annotated with textual entailment information, covering a wide range of genres of both spoken and written text (Williams et al., 2018). The evaluation metric is Accuracy.

MNLI-MM. The Multi-Genre Natural Language Inference Mismatched dataset is the mismatched version of the MNLI (matched) dataset, where the dev and test sets use out-of-domain data that does not closely resemble anything seen at training time (Williams et al., 2018). The evaluation metric is Accuracy.

QNLI. The Question-answering Natural Language Inference dataset is automatically derived from the Stanford Question Answering Dataset v1.1 (SQuAD) (Rajpurkar et al., 2016). It consists of question–paragraph pairs with one sentence in each paragraph, sourced from Wikipedia, containing the answer to the corresponding question, written by an annotator (Wang et al., 2019b). The evaluation metric is Accuracy.

RTE. The Recognizing Textual Entailment dataset is compiled from a series of textual entailment challenges: RTE1 (Dagan et al., 2005), RTE2 (Bar-Haim et al., 2006), RTE3 (Giampiccolo et al., 2007), and RTE5 (Bentivogli et al., 2009). The task requires to recognize whether the meaning of a text fragment can be inferred from the other text. The evaluation metric is Accuracy.

The subset of SuperGLUE[17] tasks that the models are evaluated on are the following.

BoolQ. The Boolean Questions dataset is a reading comprehension dataset that consists of almost 16K naturally occurring yes–no questions generated in unprompted and unconstrained settings (Clark et al., 2019). The evaluation metric is Accuracy.

MultiRC. The Multi-Sentence Reading Comprehension dataset contains around 10K questions that can be answered by combining information from a multi-sentence paragraph (Khashabi et al., 2018). The evaluation metric is F1 score.

WSC. The Winograd Schema Challenge dataset is a corpus of sentence pairs that differ in only one or two words and contain an ambiguity that can be resolved using world knowledge and reasoning (Levesque et al., 2011). The evaluation metric is Accuracy.

[17] https://super.gluebenchmark.com.

Field	Phenomenon	Acceptable example	Unacceptable example
Morphology	*Anaphor Agreement*	*Many girls insulted themselves.*	*Many girls insulted herself.*
	Determiner-Noun Agreement	*Rachelle had bought that chair.*	*Rachelle had bought that chairs.*
	Irregular Forms	*Aaron broke the unicycle.*	*Aaron broken the unicycle.*
	Subject-Verb Agreement	*These casseroles disgust Kayla.*	*These casseroles disgusts Kayla.*
Syntax	*Argument Structure*	*Rose wasn't disturbing Mark.*	*Rose wasn't boasting Mark.*
	Ellipsis	*Jill hides one orange chair and Tammy hides more.*	*Jill hides one chair and Tammy hides more orange.*
	Filler-Gap	*Brett knew what many waiters find.*	*Brett knew that many waiters find.*
	Island Effects	*Which bikes is John fixing?*	*Which is John fixing bikes?*
	Subject-Auxiliary Inversion	*Was the steak he is cooking fresh?*	*Is the steak he cooking was fresh?*
Semantics	*NPI Licensing*	*The truck has clearly tipped over.*	*The truck has ever tipped over.*
	Quantifiers	*No boy knew fewer than six guys.*	*No boy knew at most six guys.*
	Hypernym	*He has a chihuahua, so he has a dog.*	*He has a chihuahua, so he has a cat.*
Syntax & Semantics	*Binding*	*Carlos said that Lori helped him.*	*Carlos said that Lori helped himself.*
	Control/Raising	*There was bound to be a fish escaping.*	*There was unable to be a fish escaping.*
Discourse	*Q-A Congruence (easy)*	*A: Who is sleeping? B: David.*	*A: Who is sleeping? B: Eggs.*
	Q-A Congruence (tricky)	*A: Who studies? B: David.*	*A: Who studies? B: Science.*
	Turn-taking	*A: Did you arrive? B: No, we didn't.*	*A: Did you arrive? B: No, you didn't.*

Table 7: Examples of BLiMP minimal pairs (Warstadt et al., 2020).

E Derivation of EWC Regularization for Language Modeling

Let Σ be an alphabet; furthermore, define $\bar{\Sigma} \stackrel{\text{def}}{=} \Sigma \cup \{\text{EOS}\}$. We assume we have access to a collection of strings $\mathcal{D}_{\text{L}_1} = \{\boldsymbol{x}^{(n)}\}_{n=1}^{N} \subset \Sigma^*$ in L_1 and another $\mathcal{D}_{\text{L}_2} = \{\boldsymbol{y}^{(m)}\}_{m=1}^{M} \subset \Sigma^*$ in L_2. Additionally, let $\Theta \subset \mathbb{R}^d$ be a compact set of possible parameters. We take a Bayesian approach and construct a posterior density over possible parameter vectors $\boldsymbol{\theta}$. Let $\pi(\boldsymbol{\theta})$ be a prior density over Θ, and consider the following likelihood

$$p(\mathcal{D}_{\text{L}_1}, \mathcal{D}_{\text{L}_2} \mid \boldsymbol{\theta}) = \underbrace{\prod_{n=1}^{N} p(\boldsymbol{x}^{(n)} \mid \boldsymbol{\theta})}_{\stackrel{\text{def}}{=} p(\mathcal{D}_{\text{L}_1} \mid \boldsymbol{\theta})} \underbrace{\prod_{m=1}^{M} p(\boldsymbol{y}^{(m)} \mid \boldsymbol{\theta})}_{\stackrel{\text{def}}{=} p(\mathcal{D}_{\text{L}_2} \mid \boldsymbol{\theta})} \tag{2a}$$

$$= \prod_{n=1}^{N} \prod_{t=1}^{T_n} p(\text{EOS} \mid \boldsymbol{x}^{(n)}, \boldsymbol{\theta}) p(x_t^{(n)} \mid \boldsymbol{x}_{<t}^{(n)}, \boldsymbol{\theta}) \prod_{m=1}^{M} \prod_{t=1}^{T_m} p(\text{EOS} \mid \boldsymbol{y}^{(m)}, \boldsymbol{\theta}) p(y_t^{(m)} \mid \boldsymbol{y}_{<t}^{(m)}, \boldsymbol{\theta}), \tag{2b}$$

where, as the notation states, our model assumes conditional independence between data instances given the model's parameters. Note that our language model $p(\cdot \mid \boldsymbol{\theta})$ is *unusual* in that it generates sentences from both L_1 and L_2. By Bayes' rule, we have the following posterior

$$p(\boldsymbol{\theta} \mid \mathcal{D}_{\text{L}_1}, \mathcal{D}_{\text{L}_2}) \propto p(\mathcal{D}_{\text{L}_1}, \mathcal{D}_{\text{L}_2} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \tag{3a}$$

$$= p(\mathcal{D}_{\text{L}_2} \mid \boldsymbol{\theta}) p(\mathcal{D}_{\text{L}_1} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \tag{3b}$$

$$\propto p(\mathcal{D}_{\text{L}_2} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{D}_{\text{L}_1}), \tag{3c}$$

where the transition from Eq. (3a) to Eq. (3b) follows from the conditional independence assumption of Eq. (2). There is no known general-purpose algorithm to compute $p(\boldsymbol{\theta} \mid \mathcal{D}_{\mathrm{L}_1}, \mathcal{D}_{\mathrm{L}_2})$. Thus, we construct a Gaussian approximation of $p(\boldsymbol{\theta} \mid \mathcal{D}_{\mathrm{L}_1})$ by computing a second-order Taylor approximation to the log-likelihood around the likelihood mode, i.e.,

$$\boldsymbol{\theta}^*_{\mathrm{L}_1} = \arg\max_{\boldsymbol{\theta} \in \Theta} \log p(\mathcal{D}_{\mathrm{L}_1} \mid \boldsymbol{\theta}). \tag{4}$$

Applying the Taylor approximation yields

$$\log\, p(\mathcal{D}_{\mathrm{L}_1} \mid \boldsymbol{\theta}) \approx \log p(\mathcal{D}_{\mathrm{L}_1} \mid \boldsymbol{\theta}^*_{\mathrm{L}_1}) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*_{\mathrm{L}_1})^\intercal \nabla^2_{\boldsymbol{\theta}} \log p(\mathcal{D}_{\mathrm{L}_1} \mid \boldsymbol{\theta}^*_{\mathrm{L}_1})(\boldsymbol{\theta} - \boldsymbol{\theta}^*_{\mathrm{L}_1}). \tag{5}$$

We have omitted the first term in the Taylor expansion since it is zero precisely because we have expanded $\log p(\mathcal{D}_{\mathrm{L}_1} \mid \boldsymbol{\theta})$ around a local optimum. If the prior is set to be a zero-centered, spherical Gaussian with variance σ^2, i.e.,

$$\pi(\boldsymbol{\theta}) \propto \exp -\frac{1}{2\sigma^2}\boldsymbol{\theta}^\intercal\boldsymbol{\theta}, \tag{6}$$

applying Bayes' rule gives:

$$\log p(\boldsymbol{\theta} \mid \mathcal{D}_{\mathrm{L}_1}) = \log p(\mathcal{D}_{\mathrm{L}_1} \mid \boldsymbol{\theta}) + \log \pi(\boldsymbol{\theta}) - \log p(\mathcal{D}_{\mathrm{L}_1}). \tag{7}$$

As $\log p(\mathcal{D}_{\mathrm{L}_1})$ is constant with respect to $\boldsymbol{\theta}$, it does not influence the optimization problem. By replacing $\log p(\mathcal{D}_{\mathrm{L}_1} \mid \boldsymbol{\theta})$ with the approximation from Eq. (5) we obtain:[18]

$$\log p(\boldsymbol{\theta} \mid \mathcal{D}_{\mathrm{L}_1}) \approx \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*_{\mathrm{L}_1})^\intercal \nabla^2_{\boldsymbol{\theta}} \log p(\mathcal{D}_{\mathrm{L}_1} \mid \boldsymbol{\theta}^*_{\mathrm{L}_1})(\boldsymbol{\theta} - \boldsymbol{\theta}^*_{\mathrm{L}_1}) - \frac{1}{2\sigma^2}\boldsymbol{\theta}^\intercal\boldsymbol{\theta} \tag{8a}$$

$$= \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*_{\mathrm{L}_1})^\intercal \left(\sum_{n=1}^{N} \left(\nabla^2_{\boldsymbol{\theta}} \log p(\textsc{eos} \mid \boldsymbol{x}^{(n)}, \boldsymbol{\theta}^*_{\mathrm{L}_1}) + \sum_{t=1}^{T_n} \nabla^2_{\boldsymbol{\theta}} \log p(x_t^{(n)} \mid \boldsymbol{x}^{(n)}_{<t}, \boldsymbol{\theta}^*_{\mathrm{L}_1}) \right) \right) (\boldsymbol{\theta} - \boldsymbol{\theta}^*_{\mathrm{L}_1}) - \frac{1}{2\sigma^2}\boldsymbol{\theta}^\intercal\boldsymbol{\theta} \tag{8b}$$

$$\approx -\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*_{\mathrm{L}_1})^\intercal \left(\sum_{n=1}^{N} \sum_{t=1}^{T_n+1} \mathop{\mathbb{E}}_{\bar{x} \sim p(\cdot \mid \boldsymbol{x}_{<t}, \boldsymbol{\theta}^*_{\mathrm{L}_1})} \nabla^2_{\boldsymbol{\theta}} - \log p(\bar{x} \mid \boldsymbol{x}^{(n)}_{<t}, \boldsymbol{\theta}^*_{\mathrm{L}_1}) \right) (\boldsymbol{\theta} - \boldsymbol{\theta}^*_{\mathrm{L}_1}) - \frac{1}{2\sigma^2}\boldsymbol{\theta}^\intercal\boldsymbol{\theta}, \tag{8c}$$

where the last approximation replaces $\nabla^2_{\boldsymbol{\theta}} \log p(x_t^{(n)} \mid \boldsymbol{x}^{(n)}_{<t}, \boldsymbol{\theta}^*_{\mathrm{L}_1})$ with its expectation $\mathbb{E}_{\bar{x} \sim p(\cdot \mid \boldsymbol{x}_{<t}, \boldsymbol{\theta}^*_{\mathrm{L}_1})} \nabla^2_{\boldsymbol{\theta}} \log p(\bar{x} \mid \boldsymbol{x}^{(n)}_{<t}, \boldsymbol{\theta}^*_{\mathrm{L}_1})$ and performs a sign manipulation.

A Fast Approximation. Exact computation of the matrix $\mathbb{E}_{\bar{x} \sim p(\cdot \mid \boldsymbol{x}_{<t}, \boldsymbol{\theta}^*_{\mathrm{L}_1})} \nabla^2_{\boldsymbol{\theta}} - \log p(\bar{x} \mid \boldsymbol{x}^{(n)}_{<t}, \boldsymbol{\theta}^*_{\mathrm{L}_1})$ is impractical for multiple reasons. First, it requires second-order automatic differentiation, which is

[18] Note that $\log p(\mathcal{D}_{\mathrm{L}_1} \mid \boldsymbol{\theta}^*_{\mathrm{L}_1})$ in Eq. (5) is also constant w.r.t. θ, so we do not add it.

relatively expensive and not a standard feature in common automatic differentiation toolkits, e.g., PyTorch (Paszke et al., 2019). Second, recall that $p(\cdot \mid \boldsymbol{x}_{<t}, \boldsymbol{\theta}^*_{\mathrm{L}_1})$ is the next-symbol distribution of the language model, i.e., a distribution over $\bar{\Sigma}$, so $\mathbb{E}_{\bar{x}\sim p(\cdot|\boldsymbol{x}_{<t},\boldsymbol{\theta}^*_{\mathrm{L}_1})} \nabla^2_{\boldsymbol{\theta}} - \log p(\bar{x} \mid \boldsymbol{x}^{(n)}_{<t}, \boldsymbol{\theta}^*_{\mathrm{L}_1})$ could be computed in $\mathcal{O}(|\bar{\Sigma}|)$ time. While linear, $\bar{\Sigma}$ is often large in practice in modern language models. Thirdly, the matrix contains $\mathcal{O}(d^2)$ unique entries.[19] When d is large, as it is in our case, we cannot compute all d^2 entries easily. A classic identity involving the Fisher information matrix from Bickel and Doksum (2001, pg. 185) (see also Kunstner et al. (2019, Eq. 4)) allows us to derive the following approximation:

$$
\begin{aligned}
&\sum_{n=1}^{N}\sum_{t=1}^{T_n+1} \mathop{\mathbb{E}}_{\bar{x}\sim p(\cdot|\boldsymbol{x}_{<t},\boldsymbol{\theta}^*_{\mathrm{L}_1})} \nabla^2_{\boldsymbol{\theta}} - \log p(\bar{x} \mid \boldsymbol{x}^{(n)}_{<t}, \boldsymbol{\theta}^*_{\mathrm{L}_1}) \\
&\qquad \approx \sum_{n=1}^{N}\sum_{t=1}^{T_n+1} \mathop{\mathbb{E}}_{\bar{x}\sim p(\cdot|\boldsymbol{x}_{<t},\boldsymbol{\theta}^*_{\mathrm{L}_1})} \nabla_{\boldsymbol{\theta}} \log p(\bar{x} \mid \boldsymbol{x}^{(n)}_{<t}, \boldsymbol{\theta}^*_{\mathrm{L}_1}) \nabla_{\boldsymbol{\theta}} \log p(\bar{x} \mid \boldsymbol{x}^{(n)}_{<t}, \boldsymbol{\theta}^*_{\mathrm{L}_1})^\intercal && \text{(9a)} \\
&\qquad \approx \sum_{n=1}^{N}\sum_{t=1}^{T_n+1} \frac{1}{K}\sum_{k=1}^{K} \nabla_{\boldsymbol{\theta}} \log p(\bar{x}^{(k)} \mid \boldsymbol{x}^{(n)}_{<t}, \boldsymbol{\theta}^*_{\mathrm{L}_1}) \nabla_{\boldsymbol{\theta}} \log p(\bar{x}^{(k)} \mid \boldsymbol{x}^{(n)}_{<t}, \boldsymbol{\theta}^*_{\mathrm{L}_1})^\intercal && \text{(9b)} \\
&\qquad \stackrel{\text{def}}{=} \widetilde{\mathbf{F}}_{\boldsymbol{\theta}^*_{\mathrm{L}_1}} && \text{(9c)}
\end{aligned}
$$

Eq. (9b) is a standard Monte Carlo approximation where $\bar{x}^{(k)} \sim p(\cdot \mid \boldsymbol{x}^{(n)}_{<t}, \boldsymbol{\theta}^*_{\mathrm{L}_1})$. When $K \ll |\bar{\Sigma}|$, the sample-based approximation results in a significant speed-up. Finally, to avoid $\mathcal{O}(d^2)$ computation time, we only approximate the *diagonal* of $\widetilde{\mathbf{F}}_{\boldsymbol{\theta}^*_{\mathrm{L}_1}}$, which has $\mathcal{O}(d)$ entries.

A Simple Regularizer. Synthesizing the above, we arrive at a simple regularizer that should promote a language model, previously trained on L_1 data, to retain its knowledge during training on L_2 data

$$
\mathcal{R}(\boldsymbol{\theta}) = \underbrace{\frac{1}{2}\sum_{i=1}^{d} \left(\widetilde{\mathbf{F}}_{\boldsymbol{\theta}^*_{\mathrm{L}_1}}\right)_{ii} \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}^*_{\mathrm{L}_1})^2_i}_{\mathcal{R}_{EWC}} + \underbrace{\frac{1}{2\sigma^2}\boldsymbol{\theta}^\intercal\boldsymbol{\theta}}_{\mathcal{R}_{L2}} \tag{10}
$$

The second term corresponds exactly to the well-known L_2 regularization term resulting from the prior over the parameters $\boldsymbol{\theta}$. In practice, we generalize the coefficient $\frac{1}{2}$ into a tunable regularization coefficient λ and the coefficient $\frac{1}{2\sigma^2}$ into a tunable regularization coefficient μ. We tune λ on held-out data, but set $\mu = 0$ throughout the experiments and therefore omit it from the main text.[20]

[19] By Schwarz's lemma, if a function is twice *continuously* differentiable, its Hessian is symmetric—hence, the big-$\mathcal{O}$.

[20] Based on empirical observations this hyperparameter did not have an effect on the results.