

Second language input density can induce age of acquisition effects in language models

Simon Jerome Han

Department of Psychology
Stanford University
sjeromeh@stanford.edu

James L. McClelland

Department of Psychology
Stanford University

Abstract

People who learn a second language early in life are more likely to achieve high levels of grammatical proficiency than those who learn a second language later. Researchers have begun to use language models in an effort to address explanations for this age of acquisition (AOA) effect. We argue that work to date has not considered an important factor that may be relevant to AOA: younger learners may experience a greater density of exposure to a second language than older learners. While earlier work neglecting this density factor in LMs failed to find an AOA effect, we demonstrate empirically and theoretically that language models that receive second language input at lower densities learn to make second language grammatical judgments more slowly and also plateau at lower levels of ability. Crucially, this remains true even in the long run when cumulative second language input is held constant. Our findings highlight the possibility that second language AOA effects in humans as well as models might reflect differences in exposure density that covary with age. This account stands in contrast to classic accounts of second language AOA effects that emphasize the roles of maturation and entrenchment, and adds an important dimension to the effort to use language models to interrogate this effect.

1 Introduction

Children are more likely than adults to achieve native-level proficiency at a second language (Oyama, 1976; Patkowski, 1980; Hartshorne et al., 2018). To a first approximation, classic explanations for this age of acquisition (AOA) effect can be grouped into two categories: maturational accounts, which suggest a biologically timed decline in language-learning ability (Penfield and Roberts, 1959; Lenneberg, 1967; Long, 1990; Johnson and Newport, 1989), and entrenchment accounts, which propose that increasing first language experience

interferes with second language acquisition (Flege, 1995; Piske et al., 2001; Odlin, 2003; Monner et al., 2013; MacWhinney, 2017). Distinguishing between these explanations has historically been difficult because they are deeply confounded: aside from a handful of rare cases, more mature human learners are almost always more experienced learners as well.

Language models (LMs), which can acquire substantial syntactic competence by learning from naturalistic input (Warstadt et al., 2020; Hu et al., 2024), offer a new approach to making progress on this challenge. Because LM training data can be controlled directly, LMs can be used to test whether second language AOA effects can arise from experience alone. Multiple studies have now tested this question, and their conclusions have all been negative: LMs first trained on one language and then on another do not reliably show AOA effects unless additional mechanisms such as age-dependent regularization are introduced (Oba et al., 2023; Muller et al., 2025; Constantinescu et al., 2025). Constantinescu et al. conclude from this that LMs provide evidence more consistent with a maturational than an entrenchment account of human second language AOA effects.

Here we argue that these findings fail to consider an important variable that may vary with age of acquisition. This variable is density – the proportion of linguistic input an individual receives in their second language once the exposure begins. Given important differences between older and younger immigrants in factors such as cultural immersion (Jia and Aaronson, 2003), early second language learners may often receive second language input at greater densities than later learners do. Indeed, later second language learners that receive second language input at high densities have also been documented to achieve higher second language abili-

Code and data are available at: <https://github.com/S-J-HAN/second-language-input-density>

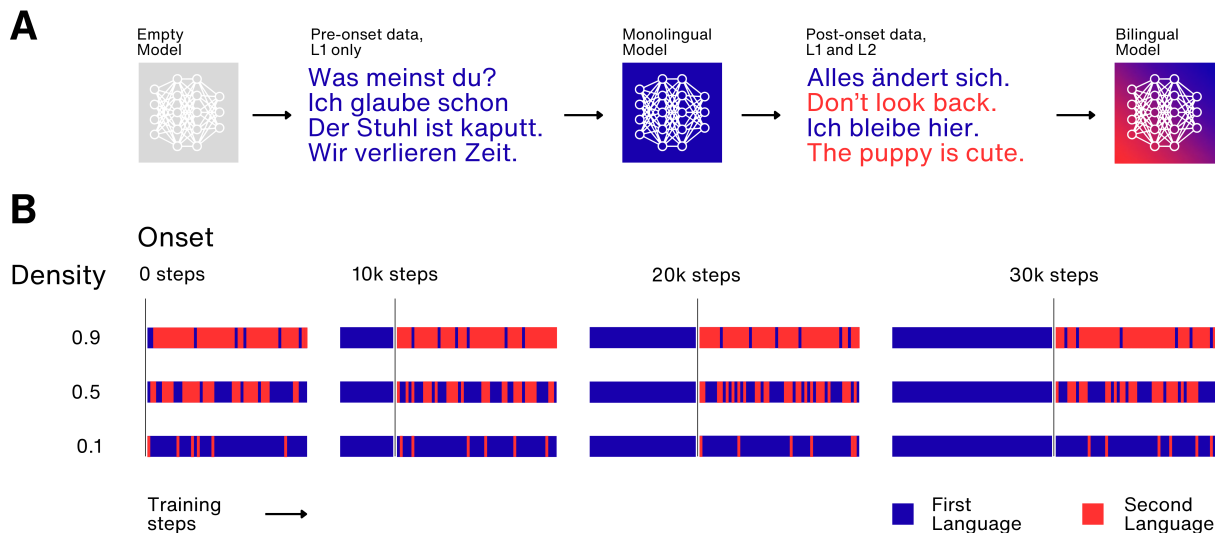


Figure 1: Our experimental design. **A**: Like previous LM second language AOA studies, we train LMs in two stages: a ‘pre-onset’ monolingual stage, where training data consists of only the first language (L1, blue), and a ‘post-onset’ bilingual stage, where training data is a mix of first and second (L2, red) languages. **B**: Different from existing studies, our design contains 3×4 conditions for each L1-L2 pair, which allows us to systematically vary the pre-onset number of L1 training steps and the post-onset proportion of training data allocated to L2.

ties (Ioup et al., 1994). Prior LM studies that operationalize second language learning simply as a shift in training distribution therefore fail to account for this important quality of human experience. With the exception of Matusевич et al. (2017), who consider a similar data parameter in a much simpler computational model from before the LM era, this is true of each of the recent LM second language AOA studies that we mentioned above.

By training LMs using an experimental design that systematically varies second language onset and density (Figure 1), we show that second language AOA effects naturally fall out of LMs under the assumption that onset and density covary (Figure 2). Specifically, we find that the timing of second language onset has little if any effect on an LM’s ability to acquire a second language, which is consistent with prior work. However, we also find that density plays a crucial role in determining the LM’s second language learning speed and ultimate level of syntactic ability. Importantly, this remains true even if we hold constant the cumulative number of second language training steps between density conditions (Figure 3), indicating that it is not due to differences in the quantity of second language training data that each model has access to.

Our results suggest that human second language AOA effects need not be attributed to traditional maturational or entrenchment accounts alone. In-

stead, they point to an alternate hypothesis: that second language AOA effects may be the natural outcome of a statistical learner optimizing a shared prediction objective over a changing mixture of linguistic environments, where learning gradients for first and second languages are in constant competition. In other words, our results are consistent with the perspective that second language AOA effects may in part be driven by differences in the learner’s input rather than changes in the learner itself, and that sufficient second language input density is required for native-level proficiency to emerge, regardless of age. This aligns with an ‘emergentist’ view of second language AOA effects (Caldwell-Harris and MacWhinney, 2023), which attributes such effects to a wide variety of cognitive and environmental factors.

2 Background

We begin by reviewing the empirical literature on second language AOA effects, the major theoretical explanations that have been proposed to explain them, and recent efforts to use neural networks and especially language models to study their origins. We note that what we refer to here as ‘second language AOA effects’ is sometimes referred to by others as ‘second language critical period effects’; we choose the former terminology as it is more theory-neutral. For compactness, we will also refer to first and second languages as ‘L1’ and ‘L2’.

2.1 The empirical observation

The observation that children who immigrate to a new linguistic environment are more likely to achieve native-level L2 abilities than adults who immigrate under similar circumstances is supported by decades of empirical evidence. Classic studies from the late 20th century demonstrate that later AOA is associated with lower levels of ultimate attainment in domains such as phonology and morphosyntax (Oyama, 1976; Patkowski, 1980; Johnson and Newport, 1989), while a recent large-scale study by Hartshorne et al. (2018) finds evidence for a prolonged period of high syntactic learning ability followed by a rapid later decline among a sample of hundreds of thousands of L2 English speakers. Together, these studies establish AOA as one of the most reliable predictors of eventual L2 proficiency, but leave open the question of what exactly causes this effect.

We note that AOA effects for L2 acquisition are related to but distinct from similar effects for L1 acquisition. Evidence from cases of severe linguistic deprivation (Fromkin et al., 1974) and late L1 exposure (Newport, 1990) suggest that L1 AOA effects also exist. Although this observation is important, such effects need not have the same explanation. While the principles that we discuss may generalize in some ways to L1 acquisition, our empirical target in this work is L2 AOA effects, and in particular in effects on L2 syntax acquisition.

2.2 The theoretical explanations

One long-standing explanation for L2 AOA effects is maturational. In particular, Lenneberg (1967) popularized the ‘Critical Period Hypothesis’ by arguing that language acquisition is enabled by a biological window of heightened neural plasticity that narrows with age. This hypothesis is supported by several lines of evidence, including observations that brain regions associated with language processing become increasingly lateralized (Krashen, 1973; Holland et al., 2001), that children recover more readily from aphasia than adults (Bates, 2014), and that early-acquired languages recruit different cortical networks from late-acquired languages (Kim et al., 1997).

An alternate family of explanations emphasizes entrenchment rather than maturation. These accounts suggest that age-related declines in L2 learning arise because later learners acquire the L2 through an existing L1 system. As L1 represen-

tations become increasingly stable, L2 is subsequently filtered through these representations, making them harder to acquire (Flege et al., 1999; Odlin, 2003). This is consistent with a broader connectionist view that prior learning changes the representational state of the learner in ways that can be inhibiting (Ellis and Lambon Ralph, 2000; MacWhinney, 2017).

Finally, age effects may also reflect differences in the circumstances under which learners acquire an L2. Children and adults often differ in their cultural immersion within a new linguistic environment, and these factors can shape both the quantity and qualities of L2 input that individuals receive. For example, while child immigrants may be placed in schools and form peer relationships that gradually shift them towards L2-dominant patterns of language use, older immigrants may continue to work and socialize using their L1 (Jia and Aaronson, 2003; Flege and Liu, 2001). Similarly, individual motivations and attitudes toward L2 acquisition can influence how much learners use it in daily life (Dornyei and Ryan, 2015). These considerations suggest that age effects may be related not only to biological maturation and L1 entrenchment, but also to the specific qualities of the learner’s input data and how they engage with it. This may help explain why there are cases where individuals who begin learning an L2 as adults can still achieve native-level L2 abilities (Ioup et al., 1994; Birdsong, 1992), as well as cases where individuals who begin learning an L2 very early in life nevertheless converge at abilities that are below native-level (Montrul, 2005; Prela et al., 2024).

2.3 Neural network models

Because their training regimes can be controlled directly, neural network models make it possible to ask whether AOA-like effects can arise from characteristics of the training regime alone. A notable example of this is Achille et al. (2018), who show that the CNNs they trained on visual tasks remained permanently impaired if they received degraded input early in training. Another, by Dohare et al. (2024), demonstrates that deep networks naturally lose plasticity over the course of continual training in a wide variety of settings, achieving better performance on earlier tasks than later tasks. These studies support the entrenchment view by demonstrating that there are at least some settings in which neural networks do exhibit AOA-like effects.

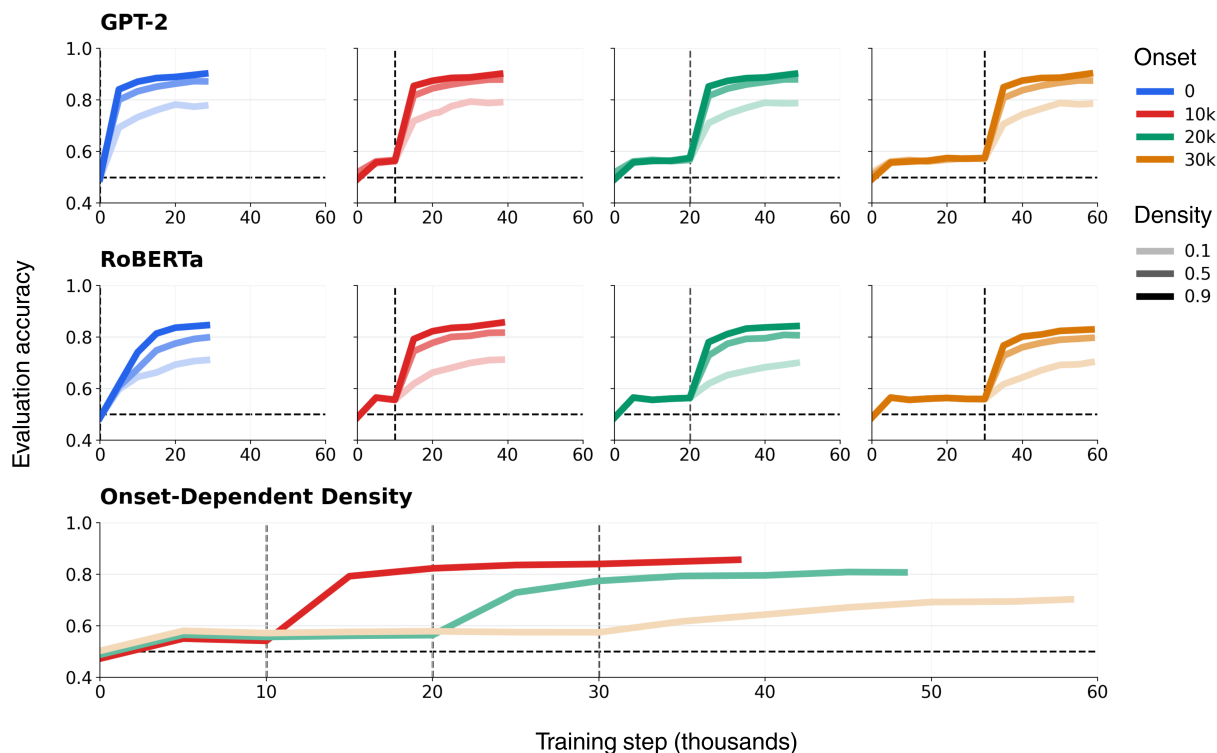


Figure 2: **Top row:** L2 syntax accuracy as a function of training steps for a 19M parameter GPT-2 model. Training steps before the vertical dashed line are L1-only, while training steps after the dashed line are sampled from an L1-L2 mix, with L2 density denoted by the darkness of each line. Horizontal dashed line denotes chance performance. **Middle row:** Similar results for a 19M parameter RoBERTa model. **Bottom row:** RoBERTa results from a hypothetical setting where density decreases with onset. Here, a human-like AOA effect emerges naturally.

Yet, for language acquisition more specifically, prior studies using LMs have thus far failed to produce evidence that they exhibit AOA-like effects for L2 acquisition. Oba et al. (2023), Constantinescu et al. (2025), and Muller et al. (2025) each report variations of an experiment where an LM is first trained on an L1 and then trained on an L2, and they each find that L1 training has very little if any impact on L2 attainment compared to baselines without prior L1 training. Oba et al. conclude that LMs are therefore ‘non-humanlike’ with respect to L2 acquisition, and Constantinescu et al. suggest that additional mechanisms such as age-dependent regularization are needed to explain AOA effects exhibited by people. However, these studies do not explore the potential impact of co-variation in the human L1-L2 exposure distribution with AOA. Our simulations highlight the possibility that such co-variation could help to explain L2 AOA effects.

3 Experiment

The central motivation for our experiment is the idea that age-dependent changes to the learner’s input data rather than changes to the learner itself

may be an important driver of L2 AOA effects. This perspective remains underrepresented in the human L2 AOA literature, and to our knowledge it is completely missing in the LM literature. Figure 1 illustrates our experimental design. Like Constantinescu et al. and other similar studies, we examine whether LMs trained on an L1 followed by an L2 exhibit an AOA effect. Unlike these studies, we introduce two parameters that manipulate the LM’s training distribution: **onset**, which determines *when* the L2 is introduced to the LM’s training data, and **density**, which determines the *proportion* of training data allocated to the L2 once it is introduced.

3.1 Varying L2 onset and density

For any given L1-L2 pair in our experiments, we systematically vary both onset and density. We use onset values of 0, 10000, 20000, and 30000, and density values of 0.1, 0.5, and 0.9. This means that the models in our experiments receive between 0 and 30000 L1 training steps before the L2 is introduced to the training distribution, and between 10 and 90% of their training data in their L2 once

it is introduced. This is implemented by sampling each microbatch with probability p from the L2 dataset and probability $1 - p$ from the L1 dataset, where p is a given density value.

Together, onset and density give us a 4×3 experimental design with 12 conditions. We run five uniquely seeded runs for each condition, L1-L2 pair, and model architecture that we test, giving us a total of 720 training runs. For each run, we train our model until we reach the equivalent of 3 epochs on our L2 dataset (ie. 3 full passes through the L2 dataset, or around 25000 steps). This means that on average we run 28000 post-onset training steps in the 0.9 density condition, 50000 steps in the 0.5 condition, and 254000 steps in the 0.1 condition, with the difference in step counts coming from additional L1 training steps.

3.2 Languages and dataset

We run our onset-density experiment across six L1-L2 pairs. In each pair, we use English as the L2, and one of German, Spanish, Turkish, Arabic, Russian, or Greek as the L1. We use English as the L2 because this allows us to use the same evaluation benchmark for all L1-L2 pairs, and we choose this particular set of L1s because they are relatively high resource languages that capture a diverse spread of languages in terms of their similarity to English.

For training data we use OpenSubtitles (Tiedemann, 2009), which provides naturalistic dialogue data for each language. This gives us datasets with 139M tokens for English, 162M for German, 196M for Spanish, 276M for Turkish, 269M for Arabic, 185M for Russian, and 105M for Greek.

3.3 Models

We collect results for two model architectures, GPT-2 (Radford et al., 2019) and RoBERTa (Liu et al., 2019), to represent autoregressive decoder and bidirectional encoder architectures respectively. This also aligns our work with Constantinescu et al., who use the same architectures. We initialize both architectures with 8 layers, 8 attention heads, a residual stream dimension of 256, an MLP hidden layer dimension of 1024, and a context window of 130. This gives us 19M parameter models for both architectures.

For tokenization, we use the pretrained tokenizers for GPT-2 and Roberta on Huggingface. During initial testing on English data alone, we found that using these tokenizers led to higher performance

on our evaluation suite than training our own tokenizers on our training data. To train our models, we use constant learning rates of $1e-3$ for GPT-2 and $5e-4$ for RoBERTa, and for both models we use a batch size of 32, 4 gradient accumulation steps, a weight decay of 0.01 and dropout of 0.1. This means that each training step processes 16,640 tokens for our GPT-2 models and 16,384 tokens for our RoBERTa models, with the difference coming from RoBERTa’s padding offset. We chose these hyperparameters based on a hyperparameter sweep conducted on each model architecture using English data alone, with performance on our evaluation suite as a target.

3.4 Evaluation

For model evaluation we focus on the ability to make grammaticality judgements, in line with previous LM L2 AOA studies (Oba et al., 2023; Constantinescu et al., 2025; Muller et al., 2025). Specifically, we use Zorro (Huebner et al., 2021), a dataset of 46000 English minimal pairs distributed across 23 syntactic categories. We use Zorro because we find that our models and datasets are too small to achieve compelling accuracy (e.g. $>75\%$) on a more substantial minimal pairs benchmark such as BLiMP (Warstadt et al., 2020). Consequently, we report Zorro results in our main text and BLiMP results, which align with our Zorro results, in Appendix A. To evaluate models on Zorro we compare sentence log probabilities for each sentence pair for GPT-2, and pseudo-log probabilities (calculated as the mean log probability assigned to each token when it is individually masked) for RoBERTa. Zorro results are then reported as the percentage of sentence pairs where the grammatical sentence had a higher probability rating than the ungrammatical sentence. During training, we evaluate on a fixed subsample of Zorro consisting of 200 items per category every 5000 training steps, and at the end of training we run evaluation on the full Zorro dataset.

4 Results

Across all conditions, we achieve an average final accuracy of 87.8% on Zorro (SD 2.8%; Min 79.3%; Max 91.8%) for GPT-2 and 78.3% (SD 7.6%; Min 58.6%; Max 89.4%) for RoBERTa, indicating that all of our models surpass chance performance on Zorro, and that on average they achieve meaningfully strong performance on Zorro as well.

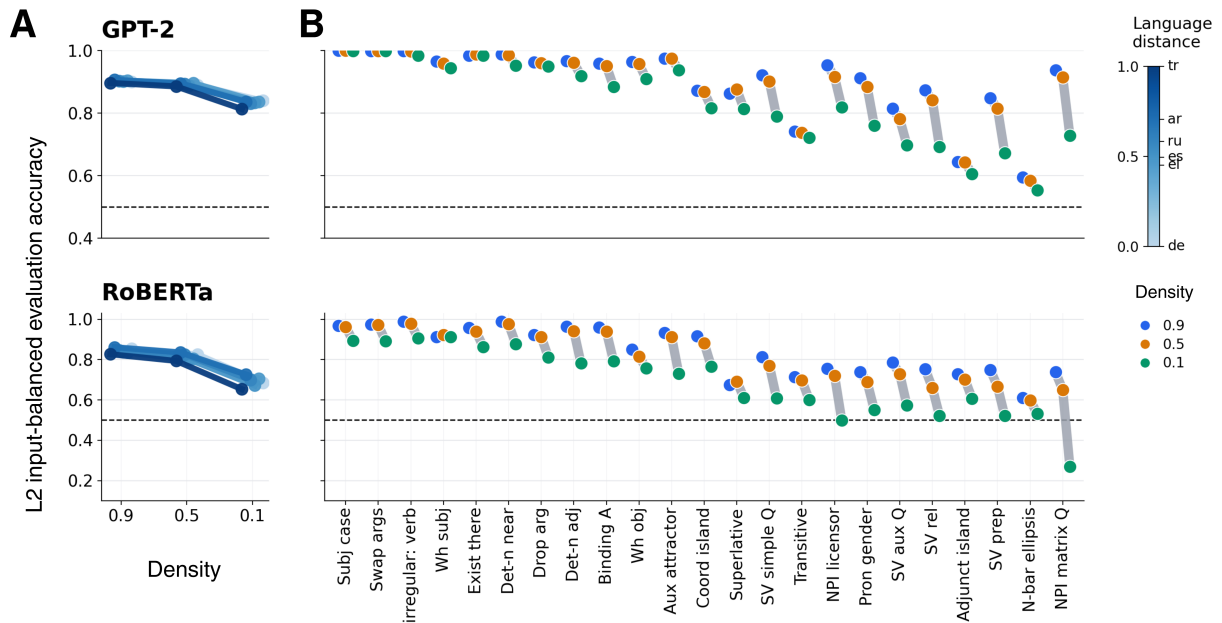


Figure 3: **A:** Final overall Zorro accuracies for different input densities, aggregated across onset values and controlling for total number of L2 training steps. Line color denotes L1. **B:** Final Zorro accuracies for different Zorro syntactic categories, aggregated across onset and L1 values. Point color denotes input density. A = Adj; SV = subject-verb; Q = question, N = Noun; NPI = negative polarity item.

4.1 Onset has very little effect on L2 acquisition

Consistent with prior work, we first find that L2 onset has at most very small effects on either architecture’s final performance on Zorro. Figure 2 shows Zorro performance as a function of training steps for GPT-2 and RoBERTa. A matched-block fixed-effect OLS comparison tested whether adding categorical onset improved prediction of final Zorro accuracy after blocking on each $L1 \times \text{density} \times \text{seed cell}$. Adding onset did not significantly improve model fit both GPT-2 ($F(3, 267) = 2.05, p = .107, \Delta R^2 = 0.002, \eta_p^2 = 0.023$), but did significantly improve model fit for RoBERTa ($F(3, 267) = 4.77, p = .003, \Delta R^2 = 0.006, \eta_p^2 = 0.051$) by a small amount. Importantly, mean accuracy did not consistently increase or decrease with later L2 onset, and performance differences between the earliest and latest onset conditions were within one percentage point of one another. Our results therefore demonstrate that initial L1 training interferes very little with later L2 training in the models that we test.

4.2 Input density affects both L2 learning speed and ultimate attainment

Unlike onset, we find that input density has large effects on final Zorro performance. Figure 2 demon-

strates that both GPT-2 and RoBERTa learn L2 much more quickly when density is 0.9 than when it is 0.5 or 0.1. This is intuitive if the only question is learning speed, as the model receives L2 input $9\times$ more frequently when density is 0.9 than when it is 0.1.

The more important result is that both models plateau at lower levels of final Zorro performance when density is lower, even when total L2 exposure is held constant. When we run the 0.5 and 0.1 density conditions until they each receive the same number of L2 training batches as the 0.9 condition, we find that GPT-2 achieves a final accuracy on Zorro of 84.3% ($SD = 1.5\%$) in the 0.1 condition and 89.1% ($SD = 0.8\%$) in the 0.5 condition, compared to 90.1% ($SD = 0.7\%$) in the 0.9 condition. Likewise, RoBERTa achieves 69.0% ($SD = 4.0\%$) in the 0.1 condition and 81.4% ($SD = 4.0\%$) in the 0.5 condition, compared to 84.3% ($SD = 3.2\%$) in the 0.9 condition. The difference in final accuracy between the 0.1 and 0.9 conditions held across all L1s, ranging from 4.6 percentage points for German to 7.4 points for Turkish for GPT-2, and from 13.0 points for Greek to 17.3 points for Turkish for RoBERTa. These results are illustrated in Figure 3A. They are further supported by fixed-effects regressions comparing density within matched $L1 \times \text{onset} \times \text{seed blocks}$, where adding density

significantly improves model fit for both GPT-2 ($F(2, 238) = 1156.93$, $p < .001$, $\Delta R^2 = .847$, $\eta_p^2 = .907$) and RoBERTa ($F(2, 238) = 952.05$, $p < .001$, $\Delta R^2 = .763$, $\eta_p^2 = .889$).

4.3 Onset and L1 similarity have almost no effect on density effects

We next ask whether density effects are moderated by onset or by the degree of similarity between L1 and L2. To test this, we convert each matched L1/onset/seed block into a paired density-advantage score, defined as the final Zorro accuracy difference between two density conditions, and regress this score against categorical onset and L1–L2 similarity as measured by Lang2Vec distance (Littell et al., 2017). For RoBERTa, onset does not reliably predict the density advantage ($F(3, 111) = 2.45$, $p = .067$, $\Delta R^2 = .061$), and Lang2Vec similarity also does not predict this advantage ($F(1, 111) = 0.01$, $p = .942$, $\Delta R^2 = .000$). There is also no evidence that the relationship between onset and density depends on Lang2Vec similarity ($F(3, 108) = 0.68$, $p = .563$, $\Delta R^2 = .017$). For GPT-2, onset does not reliably predict the density advantage ($F(3, 111) = 1.69$, $p = .174$, $\Delta R^2 = .030$), and there is no evidence for an onset-by-similarity interaction ($F(3, 108) = 0.47$, $p = .703$, $\Delta R^2 = .008$). However, Lang2Vec similarity does reliably predict the GPT-2 density advantage ($F(1, 111) = 36.70$, $p < .001$, $\Delta R^2 = .216$). Thus, density effects are broadly stable across onset conditions, and while L1–L2 similarity does not moderate the density effect in RoBERTa, it does explain meaningful variation in the magnitude of the GPT-2 density advantage.

4.4 Density effects are primarily driven by subject-verb agreement and negative polarity item licensing

Finally, to understand which syntactic phenomena drive the density effect, we decompose the final Zorro accuracy gap between 0.9 and 0.1 density conditions by their syntactic categories, and we compare matched L1/onset/seed cells and then average category effects over L1s. For GPT-2, we find that the mean category-level gap between 0.9 and 0.1 density conditions was 5.9%, with the largest drops being subject verb agreement across relative clauses (SV Rel; 16.8%, 95% CI [14.1, 19.6]), negative polarity item licensing matrix questions (NPI Matrix Q; 16.1%, 95% CI [12.3, 19.3]),

and subject verb agreement across prepositional phrases (SV Prep; 15.7%, 95% CI [12.1, 19.2]). For RoBERTa, we find that the mean category-level gap was 15.3%, with the largest category drops being NPI Matrix Q (46.9%, 95% CI [43.2, 49.8]), negative polarity item licensing with only NPI licensors (NPI Licensor; 25.6%, 95% CI [19.2, 31.5]), and SV Rel (23.1%, 95% CI [20.8, 25.5]). To account for potential ceiling-effects, we also repeat our analysis using empirical-logits in Appendix B and report similar results there. For both models, subject-verb agreement and NPI licensing accounted for the greatest differences at the family level: subject-verb agreement accounted for 39.7% of the gap and NPI licensing accounted for another 20.3% for GPT-2, while they accounted for 24.8% and 20.6% of the gap for RoBERTa.

These results hold across L1s. For example, when we calculate differences between the 0.9 and 0.1 density conditions across individual L1 and syntax-category cells, we find that the 0.9 density condition has a higher accuracy than the 0.1 density condition in 127/138 of the cells ($138 = 6 \text{ L1s} * 23 \text{ syntax categories}$) for our GPT-2 models, and in 132/138 of the cells for our RoBERTa models. These positive effects of density are distributed fairly evenly across L1s: for our GPT-2 models, the proportion of syntax categories where the 0.9 density condition has a higher accuracy than the 0.1 density condition ranges from 20/23 for German to 22/23 for Arabic and Greek, and for our RoBERTa models the proportion is also at least 20/23 for all L1s. Additionally, we find that if we rank syntax categories for each L1 by the difference between their 0.9 and 0.1 density condition accuracies, we can obtain a mean pairwise Spearman correlation between rankings of 0.79 for our GPT-2 models and 0.75 for our RoBERTa models. Of these syntax categories, SV across relative clauses is among the five largest differences for all L1s for our GPT-2 models, while NPI matrix questions is among the five largest differences for five of the six L1s. For our RoBERTa models, NPI matrix questions is also among the five largest differences for all L1s, and SV across relative clauses is among the five largest effects for five of the six L1s.

These results are further illustrated in Figure 3B. To summarize, the density effect occurs primarily within categories involving subject-verb agreement and NPI licensing for both models, and is broadly consistent in its distribution across syntax categories for all L1s.

5 Why does input density affect L2 ultimate attainment?

The result that L2 Zorro performance plateaus at lower levels given lower L2 input densities but equal L2 training steps is counterintuitive when L2 learning is viewed as a function of the number of L2 training steps. If the number of L2 training steps were all that mattered, then a model trained with low L2 input density should eventually catch up to a model trained with high L2 input density once it has seen the same amount of L2 data. Why, then, does L2 performance plateau at a lower level when density is 0.1?

This result can straightforwardly be interpreted as the consequence of gradient competition between L1 and L2 within a single shared model. When input density is high, L2 updates occur with enough frequency to reinforce one another and allow the model to stay within parameter regions that are conducive to L2 syntax. However, when density is low, each L2 update becomes embedded within a much larger stream of L1 updates, which can then undo or prevent L2-specific progress and push the model towards parameter regions that are better suited for L1. Final L2 performance therefore depends more on the density rather than the eventual number of L2 training steps.

This interpretation can be formalized with a simple model. Let p denote post-onset L2 input density, and let $z_t \in [0, 1]$ denote a scalar proxy for the model’s L2 competence at training step t . If we assume that L2 training batches increase z_t with diminishing returns such that each L2 update increases z_t in proportion to $1 - z_t$, and if we also assume that L1 training batches interfere with L2 competency such that they decrease z_t in proportion to z_t , then the expected update rule is

$$z_{t+1} = z_t + \eta p(1 - z_t) - \eta \lambda(1 - p)z_t \quad (1)$$

where η is the learning rate and λ controls the strength of L1 interference. The fixed point of this update rule is then

$$z^*(p) = \frac{p}{p + \lambda(1 - p)} \quad (2)$$

Thus, final L2 competence in this simple setting will always depend on input density, which not only determines L2 learning speed but also changes the fundamental equilibrium at which it converges, and which determines this equilibrium regardless of how much L2 input the model itself

ultimately receives. This also accounts for why final L2 performance is only very slightly lower when density is 0.5 compared to 0.9, but much lower when density is 0.1. For example, when $\lambda = 0.03$, $z^*(0.1) = 0.79$, $z^*(0.5) = 0.97$, and $z^*(0.9) = 0.99$: here, a moderate input density still provides a sufficiently dense L2 signal for L2-supporting parameters to stabilize, whereas a low input density succumbs to the overwhelming dominance of the L1 training context. The 0.1 density condition is therefore not only a slower learning version of the 0.5 and 0.9 density conditions, but also a qualitatively different learning environment.

6 Discussion

Since [Lenneberg](#) first introduced the Critical Period Hypothesis for L2 AOA effects, theories for why children tend to achieve greater L2 proficiency than adults do have tended to emphasize changes to the learner itself. This tradition has permeated into present day studies that seek to use language models to advance such theories. [Oba et al. \(2023\)](#), [Constantinescu et al. \(2025\)](#), and [Muller et al. \(2025\)](#) each test whether prior L1 experience can naturally inhibit later L2 acquisition in neural language models; when they find that this is not the case, they each then draw the conclusion that their results support the hypothesis that L2 AOA effects are not primarily driven by experience, or that L2 acquisition in language models is not humanlike. [Constantinescu et al.](#) in particular suggest that age-dependent decreases in neural plasticity, implemented in their case using elastic weight consolidation ([Kirkpatrick et al., 2017](#)), might instead be necessary to account for the difference in AOA effects between language models and people. Explanations for L2 AOA effects have thus passed back and forth from one candidate mechanism internal to the learner to another.

Here, we argue that explanations for L2 AOA effects need not be attributed to these internal changes alone. Although we do not deny that there could be such changes and that they could play a role, we explore an even simpler idea: that age-dependent changes to a learner’s *input*, rather than to the learner itself, may drive a meaningful portion of the AOA effect. Using a large set of language model training runs on six different L1-L2 pairs, we find in our simulations that varying input density, the proportion of training steps allocated to the L2 once it has been introduced, can lead the model

to plateau at very different levels of L2 syntax ability. Indeed, out of L1-L2 similarity, L2 onset, and L2 density, we show that density is broadly the only parameter that matters for determining these plateaus. Thus, if we assume that density tends to decrease with age, then we find that L2 AOA effects naturally fall out of language models as well.

Translating this result back to the human literature requires the assumption that input density also covaries with age in people. There are good reasons to believe that this is true, and our work is related to prior work that promotes or otherwise hints at these ideas. For example, [Hartshorne et al. \(2018\)](#) find that within a large sample of hundreds of thousands of second language English speakers, the L2 AOA effect is empirically best modeled as a long plateau in learning ability followed by a sharp and continuous decline at around the age of 17. This point of decline is significantly later than what classic theories may have predicted, and it coincides quite conspicuously with the end of secondary education, and therefore the closing of an environment where dense L2 exposure is particularly easy to come by. [Jia and Aaronson \(2003\)](#) follow Chinese child and adolescent immigrants to America over a period of three years, and find that the younger children quickly switched over to L2 dominant language use while the older adolescents were more likely to maintain L1 dominant language use; amongst several environmental features, [Jia and Aaronson](#) note the greater ratio of L2/L1 input that the younger children received. Finally, work such as that of [Paradis \(2019\)](#) and [Ioup \(1989\)](#) suggest that easy L2 acquisition is not a given for children – sufficient input is required, and individual variation is large – while work such as that of [Flege et al. \(1999\)](#) finds that AOA effects on morphosyntax can go away when confounding variables such as education and self-estimates of L1 and L2 use are controlled for.

Translating this result back to the human literature also requires a willingness to abstract away other important differences between humans and language models – the underlying transformer-based architectures of the language models that we study are of course different in many ways from the language network of the human brain. For example, it seems unlikely that one would find the equivalent of a pretrained multilingual tokenizer within the human language network, even though the existence and quality of such a tokenizer is crucial for enabling our language models to achieve meaningful performance. It is also certainly true that

our language models lack countless other mechanisms, both known and unknown, that exist within the human language network. Our intention here is not to make specific claims about the biological and mechanistic correspondence between humans and language models as language learners. Instead, we hope to make two more general points. First, language models *can* exhibit L2 AOA effects under plausibly human-like conditions. They should therefore not be used to support maturational accounts or refute experience-based accounts of human L2 AOA effects. Second, language models offer a compelling existence proof for the idea that data parameters such as input density can drive learning outcomes that resemble L2 AOA effects almost entirely on their own. This aligns with the spirit of other works that have argued for the value that language models have as model systems for modern linguistics ([Futrell and Mahowald, 2026](#)).

In summary, our work here suggests that language models can exhibit L2 AOA effects if we allow for systematic, age-dependent changes in their input data. This presents an alternate perspective from [Constantinescu et al.](#), who instead suggest that age-dependent regularization may make up the difference between humans and language models with respect to AOA effects. To build on this work, one natural next step would be to run simulations on models that operate on raw speech rather than tokenized text. This would allow us to sidestep the question of tokenizer choice. It may also allow us to begin answering questions about AOA effects for other components of language such as morphology and phonology that are impossible to study with a tokenizer-based system. Another step would be to interrogate more thoroughly whether real-world human data aligns with the density-based account that we have proposed. This would involve confirming that for people, L2 input density does tend to decrease with age. It would also involve confirming that density is the driver of AOA effects rather than a mere co-variate of age. For example, individuals with early onset but low density should be relatively likely to have low L2 proficiency, while individuals with late onset but high density should be relatively likely to have high L2 proficiency. This aligns with prior work that examines heritage speakers ([Montrul, 2005](#); [Prela et al., 2024](#)) and immersion speakers ([Birdsong, 1992](#); [Ioup et al., 1994](#)) of an L2, and it would be valuable to reexamine similar cases with the motivation of using input density as an explicit independent variable.

Limitations

Our work has several limitations. First, although we have shown that LMs can exhibit AOA effects if we allow L2 onset to covary with input density, we cannot yet categorically state that this is true for people. The existing literature on L2 acquisition suggests that it is highly likely, but an important next step would be to collect data that explicitly makes the case that people tend to receive lower density L2 input as they age. Second, we only test relatively small instantiations of two language model architectures, and it remains to be seen whether the gradient competition dynamics for mixed L1/L2 training continue to apply to models with larger parameter counts. We do however suggest that the brain itself is a capacity-limited system, and so our exploration with these capacity-limited models remains informative. Third, our experiments use only a single language, English, as the L2; future work should confirm our finding on a range of L2s. Fourth, our models use the pretrained tokenizers for GPT-2 and RoBERTa respectively. On the one hand we believe that this is acceptable because our evaluation metric is calculated with respect to a single L2, English, which these tokenizers are primarily trained on. On the other, we recognize that the use of tokenizers that separate L1 and L2 into such different input spaces for the model represents a striking departure from human language perception. Thus, another interesting future direction would be to consider more carefully how different tokenization schemes may impact L2 acquisition.

Acknowledgments

We thank Carol Dweck, Ellen Markman, Nastasia Klevak, and Satchel Grant for helpful feedback and discussion on various drafts of this project.

AI usage statement

We used GPT-5.5 via OpenAI Codex to provide light editing suggestions for clarity, and for help with coding up our experiments and analysis. We also used GPT-5.5 to provide the initial layout for why input density affects L2 ultimate attainment detailed in Section 5.

References

Alessandro Achille, Matteo Rovere, and Stefano Soatto. 2018. Critical learning periods in deep networks. In

International conference on learning representations.

Elizabeth Bates. 2014. Plasticity, localization, and language development 1. In *Biology and Knowledge Revisited*, pages 205–253. Routledge.

David Birdsong. 1992. Ultimate attainment in second language acquisition. *Language*, 68(4):706–755.

Catherine L Caldwell-Harris and Brian MacWhinney. 2023. Age effects in second language acquisition: Expanding the emergentist account. *Brain and language*, 241:105269.

Ionut Constantinescu, Tiago Pimentel, Ryan Cotterell, and Alex Warstadt. 2025. Investigating critical period effects in language acquisition through neural language models. *Transactions of the Association for Computational Linguistics*, 13:96–120.

Shibhansh Dohare, J Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A Rupam Mahmood, and Richard S Sutton. 2024. Loss of plasticity in deep continual learning. *Nature*, 632(8026):768–774.

Zoltan Dornyei and Stephen Ryan. 2015. *The psychology of the language learner revisited*. Routledge.

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.

James E Flege. 1995. Second language speech learning: Theory, findings, and problems. *Speech perception and linguistic experience: Issues in cross-language research*, 92(1):233–277.

James Emil Flege and Serena Liu. 2001. The effect of experience on adults’ acquisition of a second language. *Studies in second language acquisition*, 23(4):527–552.

James Emil Flege, Grace H Yeni-Komshian, and Serena Liu. 1999. Age constraints on second-language acquisition. *Journal of memory and language*, 41(1):78–104.

Victoria Fromkin, Stephen Krashen, Susan Curtiss, David Rigler, and Marilyn Rigler. 1974. The development of language in genie: A case of language acquisition beyond the “critical period”. *Brain and language*, 1(1):81–107.

Richard Futrell and Kyle Mahowald. 2026. How linguistics learned to stop worrying and love the language models. *Behavioral and Brain Sciences*, 49:e198.

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.

- Scott K Holland, Elena Plante, Anna Weber Byars, Richard H Strawsburg, Vince J Schmithorst, and William S Ball Jr. 2001. Normal fmri brain activation patterns in children performing a verb generation task. *Neuroimage*, 14(4):837–843.
- Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, 121(36):e2400917121.
- Philip A Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. Babyberta: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th conference on computational natural language learning*, pages 624–646.
- Georgette Ioup. 1989. Immigrant children who have failed to acquire native english. *Variation in second language acquisition*, 2:160–175.
- 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.
- Gisela Jia and Doris Aaronson. 2003. A longitudinal study of chinese children and adolescents learning english in the united states. *Applied psycholinguistics*, 24(1):131–161.
- 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.
- Karl HS Kim, Norman R Relkin, Kyoung-Min Lee, and Joy Hirsch. 1997. Distinct cortical areas associated with native and second languages. *Nature*, 388(6638):171–174.
- James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and 1 others. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.
- Stephen D Krashen. 1973. Lateralization, language learning, and the critical period: Some new evidence. *Language learning*, 23(1):63–74.
- Eric H Lenneberg. 1967. The biological foundations of language. *Hospital Practice*, 2(12):59–67.
- Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 8–14.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Michael H Long. 1990. Maturational constraints on language development. *Studies in second language acquisition*, 12(3):251–285.
- Brian MacWhinney. 2017. Entrenchment in second-language learning. In *Entrenchment and the psychology of language learning: How we reorganize and adapt linguistic knowledge.*, pages 343–366. De Gruyter Mouton.
- Yevgen Matushevych, Afra Alishahi, and Ad Backus. 2017. The impact of first and second language exposure on learning second language constructions. *Bilingualism: Language and Cognition*, 20(1):128–149.
- Derek Monner, Karen Vatz, Giovanna Morini, So-One Hwang, and Robert DeKeyser. 2013. A neural network model of the effects of entrenchment and memory development on grammatical gender learning. *Bilingualism: Language and Cognition*, 16(2):246–265.
- Silvina Montrul. 2005. Second language acquisition and first language loss in adult early bilinguals: Exploring some differences and similarities. *Second language research*, 21(3):199–249.
- Dyana Muller, Koquium Li Lin, Charles Kemp, and Francis Mollica. 2025. Probing for experience-driven critical period effects in a large language model. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47.
- Elissa L Newport. 1990. Maturational constraints on language learning. *Cognitive science*, 14(1):11–28.
- 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, Toronto, Canada. Association for Computational Linguistics.
- Terence Odlin. 2003. Cross-linguistic influence. *The handbook of second language acquisition*, pages 436–486.
- Susan Oyama. 1976. A sensitive period for the acquisition of a nonnative phonological system. *Journal of psycholinguistic research*, 5(3):261–283.
- Johanne Paradis. 2019. English second language acquisition from early childhood to adulthood: The role of age, first language, cognitive, and input factors. In *Proceedings of the BUCLD*, volume 43, pages 11–26.
- Mark S Patkowski. 1980. The sensitive period for the acquisition of syntax in a second language 1. *Language learning*, 30(2):449–468.

Wilder Penfield and Lamar Roberts. 1959. *Speech and brain mechanisms*. Princeton University Press.

Thorsten Piske, Ian RA MacKay, and James E Flege. 2001. Factors affecting degree of foreign accent in an l2: A review. *Journal of phonetics*, 29(2):191–215.

Leonarda Praela, Ewa Dąbrowska, and Miquel Llompart. 2024. Beyond age: exploring ultimate attainment in heritage speakers and late l2 learners. *Frontiers in Psychology*, 15:1419116.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Jörg Tiedemann. 2009. *News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces*, volume V, pages 237–248.

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.

A BLiMP Results

In addition to Zorro, we also run L2 syntax evaluation using BLiMP (Warstadt et al., 2020). Because BLiMP is much more challenging than Zorro, our models tend to achieve final accuracies that are below 75%. Broadly, our BLiMP results are similar to our Zorro results, except with lower accuracies across all conditions. Figures 5 and 6 illustrate BLiMP versions of Figures 2 and 3.

Quantitatively, we observe that on average GPT-2 achieves 69.2% accuracy on BLiMP ($SD = 2.0\%$; Min 63.8%; Max 72.2%) while RoBERTa achieves 63.9% accuracy ($SD = 3.8\%$; Min 54.0%; Max 69.9%). When we match L1/onset/seed blocks, GPT-2 achieves an average accuracy of 66.7% ($SD = 1.0\%$) in the 0.1 density condition, 70.0% ($SD = 0.7\%$) in the 0.5 condition, and 71.0% ($SD = 0.6\%$) in the 0.9 condition. Similarly, RoBERTa achieves an average accuracy of 59.3% ($SD = 1.7\%$) in the 0.1 condition, 65.3% ($SD = 2.1\%$) in the 0.5 condition, and 67.0% ($SD = 1.6\%$) in the 0.9 condition. Adding density significantly improves matched-block model fit for both GPT-2 ($F(2, 238) = 1565.54$, $p < .001$, $\Delta R^2 = 0.851$, $\eta_p^2 = 0.929$) and RoBERTa ($F(2, 238) = 901.72$, $p < .001$, $\Delta R^2 = 0.765$, $\eta_p^2 = 0.883$). By contrast, onset effects are not statistically significant for GPT-2 ($\Delta R^2 = 0.001$) and are statistically significant but small for RoBERTa

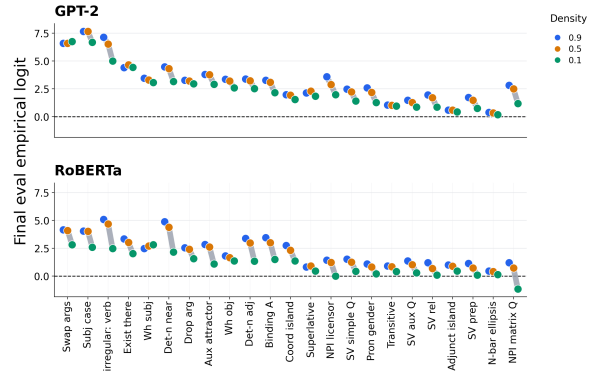


Figure 4: Final Zorro empirical-logits for different Zorro syntactic categories, aggregated across onset and L1 values. Point color denotes input density.

($\Delta R^2 = 0.007$). Our BLiMP results therefore support the same broad conclusion as our Zorro results that input density is the primary driver of final L2 syntactic proficiency.

B Ceiling-effects for Zorro subcategories

Our decomposition of density effects on Zorro subcategories is reported in the main text using raw accuracy scores. However, our models achieve very high accuracy on several subcategories even when density is 0.1, raising concerns about ceiling effects. To address these concerns, we repeat our decomposition on an empirical logit-scale, where we reuse the L1/onset/seed/category cells reported in main text but convert accuracy scores to $\log((c + \epsilon)/(n - c + \epsilon))$, where n is the number of items in the subcategory, c is the number of correct items, and ϵ is a small pseudo-count, set to 0.5, that keeps transformed values finite. Results are visualized in Figure 4. Using the same 0.9-0.1 density contrast, we find that subject-verb agreement and NPI licensing remain amongst the largest contributors to the empirical-logit category gap for both GPT-2, where subject-verb agreement accounts for 21.1% of the gap and NPI licensing for 17.1% of the gap, and RoBERTa, where subject-verb agreement accounts for 15.1% of the gap and NPI licensing for 13.2% of the gap. The main difference from the raw accuracy analysis is that the largest contributor to the gap for RoBERTa is instead determiner-noun agreement, at 16.5%. Thus, although the raw accuracy scores suggest that near-ceiling points should be interpreted with a degree of caution, the empirical-logit analysis indicates that our conclusion is not an artifact of those ceiling effects.

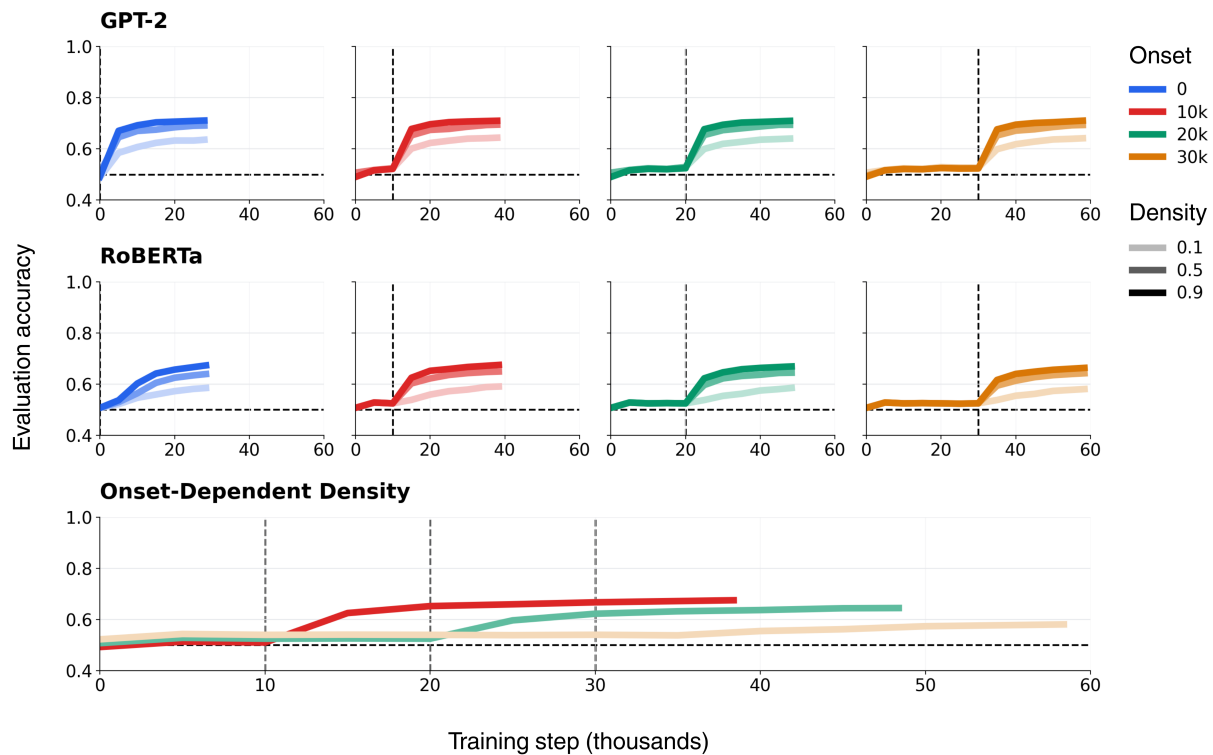


Figure 5: **Top row:** L2 syntax accuracy, as measured by BLiMP rather than Zorro, as a function of training steps for a 19M parameter GPT-2 model. **Middle row:** Similar results for a 19M parameter RoBERTa model. **Bottom row:** RoBERTa results from a hypothetical setting where density decreases with onset. Again, a human-like AOA effect emerges naturally.

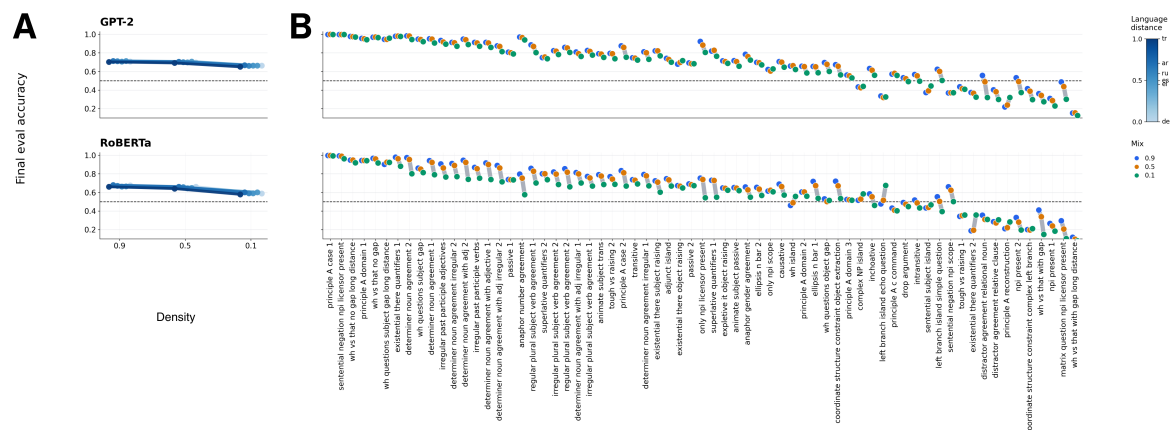


Figure 6: **A:** Final overall BLiMP accuracies for different input densities, aggregated across onset values and controlling for total number of L2 training steps. Line color denotes L1. **B:** Final BLiMP accuracies for different BLiMP syntactic categories, aggregated across onset and L1 values. Point color denotes input density.