

# Monolingual and Bilingual Language Acquisition in Language Models

Mihir Sharma, Ryan Ding, Raj Sanjay Shah, Sashank Varma

Georgia Institute of Technology



## Motivation

- Children can simultaneously acquire multiple languages.
- With approximately the same vocabulary size across languages.
- Children acquiring one language outperform children acquiring two languages on monolingual language understanding evaluations.

- Expressed as a <3-month lag at age 2;1 (increases with age)

★ Can we model this rate of language development using Monolingually and Bilingually developing language models?

## Languages and Ordering

For two languages L1 and L2, we run the following simulation types for the pretraining step:

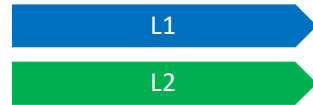


Language Model

BabyLlama

Parameters: 16 Mil  
Vocab Size: 16000 tokens  
Sequence Len: 256

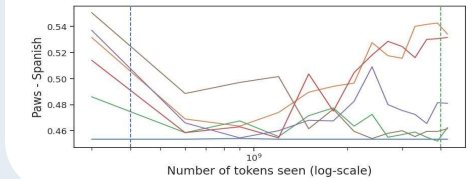
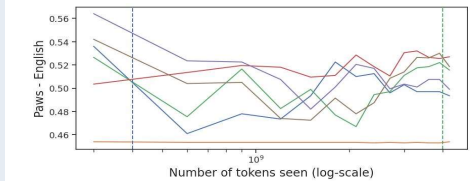
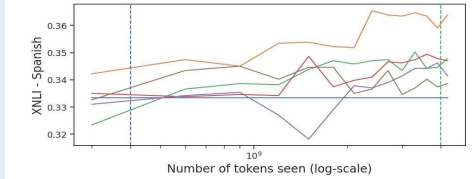
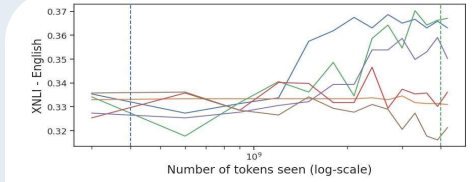
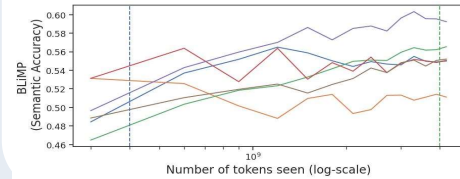
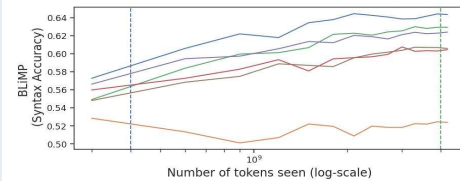
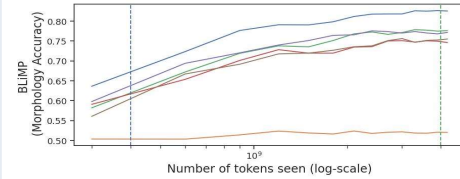
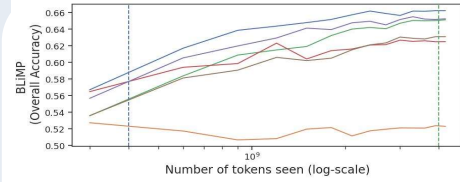
Monolingual



Bilingual



## Results



— English — Seq Spanish-English --- 1 Epoch  
— Spanish — Random English-Spanish --- 10 Epochs  
— Seq English-Spanish — Random English-Spanish-German

## Evaluation

BLiMP

Semantics

She might not ever argue.  
She might probably ever argue.

Syntax

Noah approached.  
Noah works with.

★ Use English-only tasks

Surprisal based tasks

XNLI

Does one sentence entail the other?

PAWS-X

Does one sentence paraphrase the other?

★ Simple Multi-lingual tasks

## Key Takeaways



Baby models can be used to understand Monolingual and Bilingual language acquisition in humans.



Performance  
Monolingual > Bilingual > Trilingual models.



Ordering effects are hard to evaluate in a multi-epoch training setup.



Find us!  
Scan the QR code

