Mind the Gap: How BabyLMs Learn Filler-Gap Dependencies

Chi-Yun Chang, Xueyang Huang, Humaira Nasir, Shane Storks, Olawale Akingbade, Huteng Dai

University of Michigan

Introduction

- Humans acquire complex syntactic dependencies such as filler-gap relationships from limited and often noisy input, raising the question of whether artificial neural language models can achieve the same (Gagliardi et al., 2016; Wilcox et al., 2018; Howitt et al., 2024).
- **Filler-gap dependencies** provide a critical test of syntactic learning because they require models to track long-distance relationships and respect island constraints.
- The BabyLM Challenge offers **child-oriented corpora** for training language models on resource-limited input, allowing a developmentally realistic framework to examine syntactic acquisition in smaller-scale models (Hu et al., 2024).

Research Question: Can language models trained on predominantly child-oriented, child-sized input acquire filler-gap dependencies, generalize across constructions, and respect structural constraints such as island effects?

Methodology

This project investigates the acquisition of filler-gap dependencies by GPT-2 models trained on child-language data, following established methods (Wilcox et al., 2018; Ozaki et al., 2022).

Models Trained	GPT-2 (10M tokens and 100M tokens) on BabyLM Challenge corpora (70% child-directed/oriented input)
Additional Models Assessed	ConcreteGPT (10M tokens) (Capone et al., 2024) and BabbleGPT (100M tokens) (Goriely et al., 2024)
Upper Bound Model	A threshold GPT-2 pretrained on a 40GB general corpus
Test Materials	A suite of sentences for four major syntactic constructions
Experimental Design	2×2 factorial setup: manipulate presence of fillers (wh-licensors) and gaps to yield grammatical vs. ungrammatical conditions
Evaluation Metrics	Surprisal at critical regions; wh-licensing scores (filler–gap interaction); flip tests (surprisal reverses with gap presence); grammaticality division tests (surprisal difference: grammatical vs. ungrammatical)
Statistical Analysis	Tests are conducted through mixed-effects linear regression with random intercepts by sentence set to test acquisition of filler–gap dependencies and structural constraints

Table 1. Methodology summary

Construction Types

Here we outline each construction type and what aspect of filler-gap behavior it is designed to test.

Types	Examples
A. Gap Distance	They found out [that / what]filler the baker [who lives nearby / who visits the bakery every Sunday]mod length gave [a free loaf / ___]gap to the customer this morning.
B. Double Gaps	John knows [that / who]filler [the police / ___]gap1 found [the thief / ___]gap2 in the alley.
C. Wh-Islands	You mentioned [that / *what]filler your coworker stated { [whether]complementizer the intern sent [the wrong file / *___]gap to the client }wh-island
D. Adjunct Islands	We found out [that / *what]filler [the parade started after]adjunct pos trigger {the mayor of the city gave [the opening speech / *___]gap in front of the cheering crowd.}adjunct island back

Table 2. Construction types illustrating filler-gap dependencies. Bold, colored spans are manipulated factors: mod length varies modifier length; complementizer varies the word that introduces the embedded clause (e.g., that, whether); adjunct pos trigger varies where the adjunct island occurs; filler and gap indicate the filler and gap sites.

Conclusion and Discussion

- **Larger** GPT-2 models show **stronger** filler–gap learning, but still fail on complex constraints like adjunct islands.
- All models show weaker performance on **long-distance dependencies**, mirroring the late acquisition of such patterns in child language.
- **Flip test** shows that even stronger models do **not** fully capture filler–gap **bijectivity**, suggesting inductive biases are needed for human-like generalization.
- **BabyLM** models **outperform GPT-2-10M** on several constructions but show **mixed results at 100M**, indicating modest gains from specialized training yet persistent difficulty with complex island constraints.

References

Luca Capone, Alessandro Bondielli, and Alessandro Lenci. Concretegpt: A baby gpt-2 based on lexical concreteness. In *BabyLM Challenge, EMNLP 2024*, 2024. Poster presentation.

Annie Gagliardi, Tara M Mease, and Jeffrey Lidz. Discontinuous development in the acquisition of filler-gap dependencies: Evidence from 15-and 20-month-olds. *Language Acquisition*, 23:234–260, 2016.

Zébulon Goriely, Richard Diehl Martinez, Andrew Caines, Paula Buttery, and Lisa Beinborn. From babble to words: Pre-training language models on continuous streams of phonemes. In *BabyLM Challenge, EMNLP 2024*, 2024.

Kevin Howitt, Sunil Nair, Amelia Dods, and Robert M Hopkins. Generalizations across filler-gap dependencies in neural language models. *Proceedings of the 28th Conference on Computational Natural Language Learning*, 2024.

Michael Y. Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and Ethan Gotlieb Wilcox. Findings of the second babylm challenge: Sample-efficient pretraining on developmentally plausible corpora. In *BabyLM Challenge, EMNLP 2024*, 2024.

Satoru Ozaki, Daniel Yurovsky, and Lauren Levin. How well do LSTM language models learn filler-gap dependencies? In *Proceedings of the Society for Computation in Linguistics 2022*, pages 76–88, 2022.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. What do RNN language models learn about filler-gap dependencies? *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2018.

Licensor-Gap Interaction

Each row in **Figure 1** and **Figure 2** represents one construction: (a) `gap-distance-obj`, (b) `gap-distance-PP`, (c) `wh-islands`, (d) `adjunct-islands`. Constructions fully learned with statistical significance (robust to intervening factors and capturing island constraints) are marked by asterisks. **Figure 3** visualizes the wh-licensing scores of double gaps.

- The **threshold GPT-2 model** demonstrates robust acquisition of most filler-gap dependencies and partial sensitivity to island effects.
- **GPT-2-10M** fails to acquire any construction, while **GPT-2-100M** succeeds fully on `gap-distance-obj` and shows global licensing behavior in `wh-islands`, but fails to capture `adjunct-island` constraints.
- **ConcreteGPT** demonstrates global licensing behavior for wh-islands, while **BabbleGPT** acquires `gap-distance-obj` fully, and show licensing effects for `wh-islands` and `adjunct-islands`. However, both models continue to overlook island constraints.

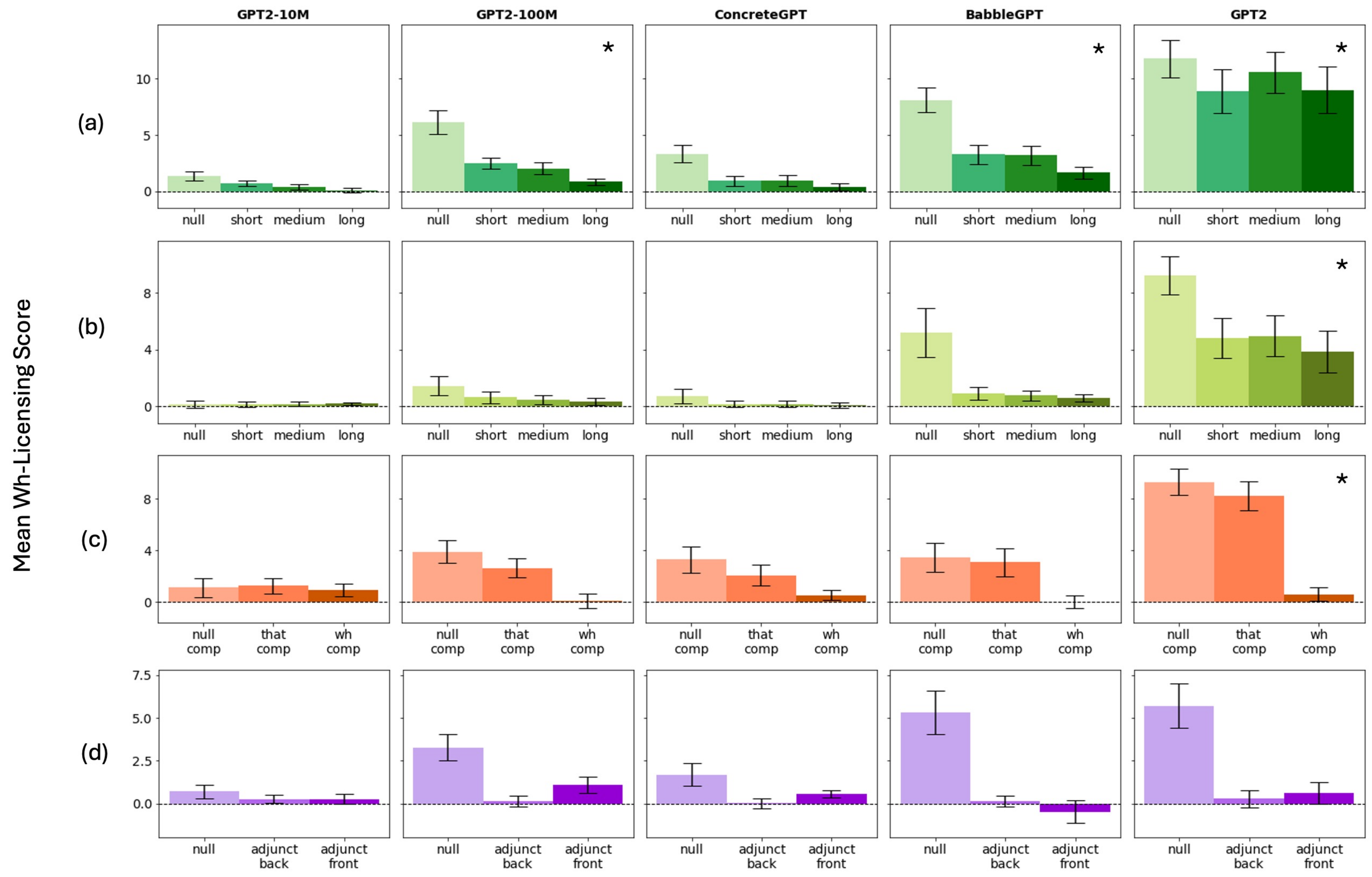

Figure 1. Wh-licensing scores with local surprisals.

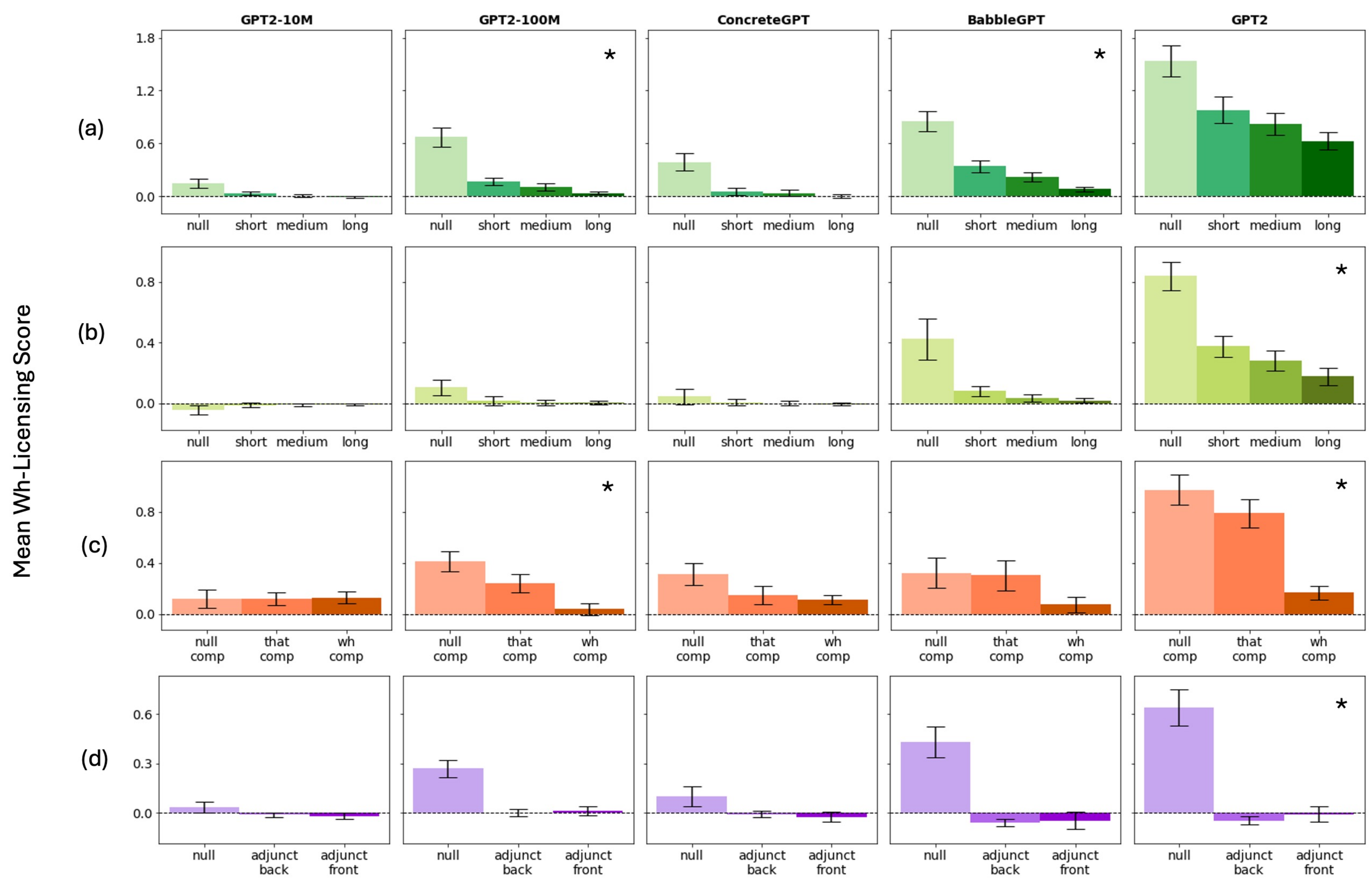

Figure 2. Wh-licensing scores with global surprisals.

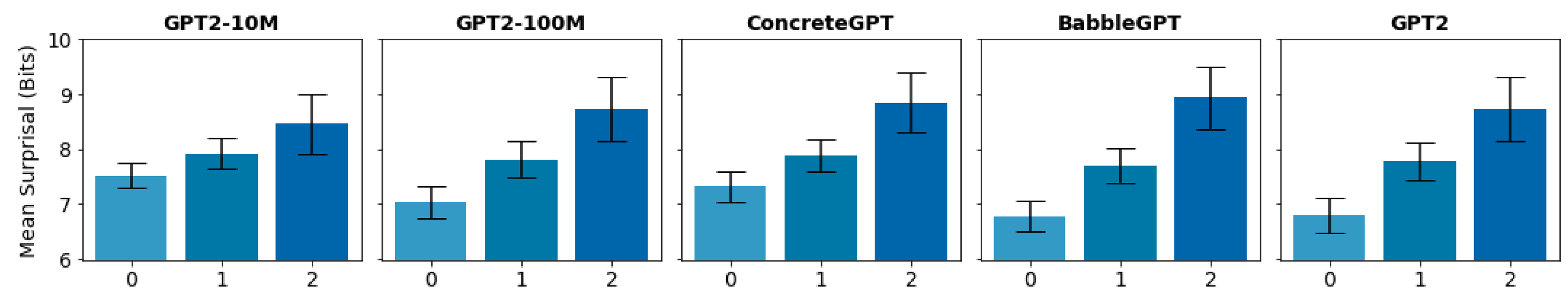

Figure 3. Double gaps: mean surprisal vs. number of illicit gaps.

Flip Test

- The **threshold GPT-2 model** passes most flip tests, capturing both directions of bijectivity and showing island awareness, though adjunct-island results are mixed.
- **GPT-2-10M** captures only one direction in some constructions and misses island effects, while **GPT-2-100M** passes the flip test for local `gap-distance-obj` and local `wh-islands`, but remains inconsistent on other island constructions.
- **ConcreteGPT** shows one-sided flips across constructions without island sensitivity, whereas **BabbleGPT** passes the flip test for local `wh-islands` and local `adjunct-islands`, demonstrating better acquisition of island constraints.

Division by Grammaticality

- The GPT-2 model passes the grammaticality test for all constructions with high statistical significance.
- **GPT-2-10M** passes the test for `double-gaps` and `wh-islands`. **GPT-2-100M** passes the test for `double-gaps`, `wh-islands`, and `adjunct-islands`.
- **ConcreteGPT** passes the test for `double-gaps`, `wh-islands`, and `adjunct-islands`. **BabbleGPT** passes the test for all constructions except for `gap-distance-PP`.

Github: https://github.com/um-cap-lab/EMNLP-2025-submission   Poster Session   {stellacy, xueyangh, humairan, sstorks, olawale, huteng}@umich.edu