

What Exactly do Children Receive in Language Acquisition?

A Case Study on CHILDES with Automated Detection of Filler-Gap Dependency



paper



dataset

Zhenghao Herbert Zhou & William Dai & Maya Viswanathan &
Simon Charlow & R. Thomas McCoy & Robert Frank

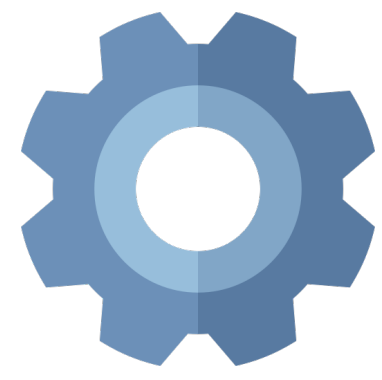
Department of Linguistics & Department of Computer Science, Yale University

Yale

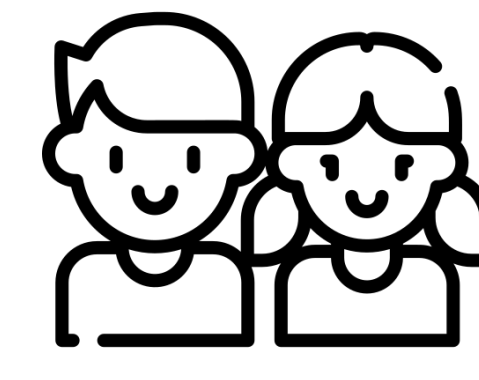
TL;DR

- **Fine-grained input matters:** children's filler-gap dependency (FGD) exposure differs by **construction** and **extraction site**, which shape *linguistic generalization* in acquisition. In the paper:

- * A detection tool for fine-grained FGD, combining constituency and dependency parsing.



- * A large-scale analysis of FGD distribution on CHILDES (with dataset).



- * A BabyLM filtered-corpus training study on generalization.



MOTIVATIONS

- Do children learn an abstract filler-gap dependency that generalizes across constructions, or do they first learn construction-specific, lexically anchored patterns?
- We need *fine-grained, large-scale* data to answer those questions!
- Previous hand annotations are hard to scale, and we provide an **automated, fine-grained FGD detection tool**.

TARGET CONSTRUCTIONS

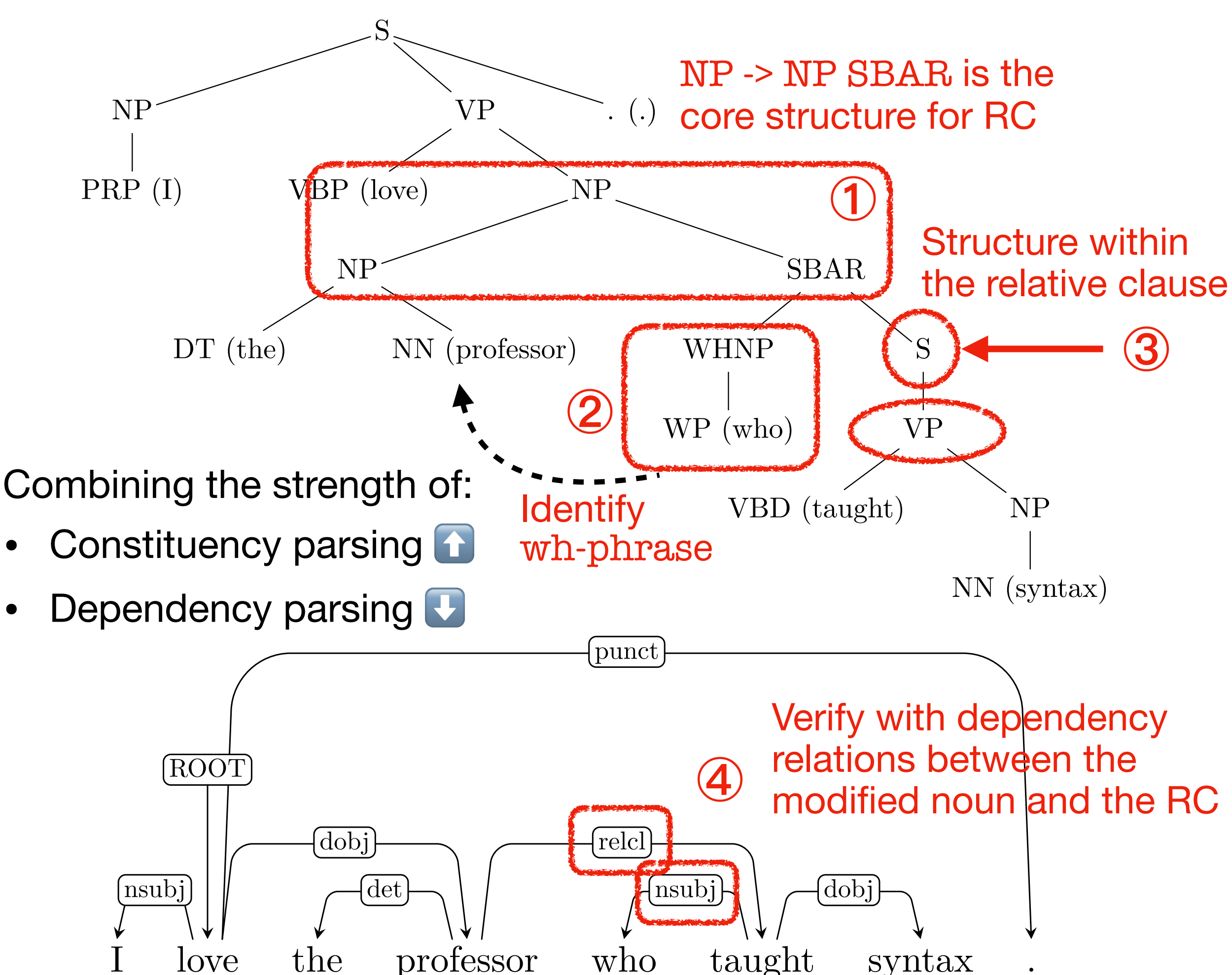
Matrix Questions
Subject: *Who* _ praised the students?
Object: *Who* did the professor praise _?
Adjunct: *When* did the professor praise the student _?

Embedded Questions
Subject: I wonder [*who* _ praised the student].
Object: I wonder [*who* the professor praised _].
Adjunct: I wonder [*why* the professor praised the student _].

Relative Clause
Subject: The professor [*who* _ praised the student] smiled.
Object: The professor [*who* the student praised _] smiled.
Adjunct: The day [*when* the professor praised the student _] was memorable.

- For extraction site, we also detected polar *wh*-questions (yes/no questions) and reduced SRC and ORC (whiz-deletion, contact RC).

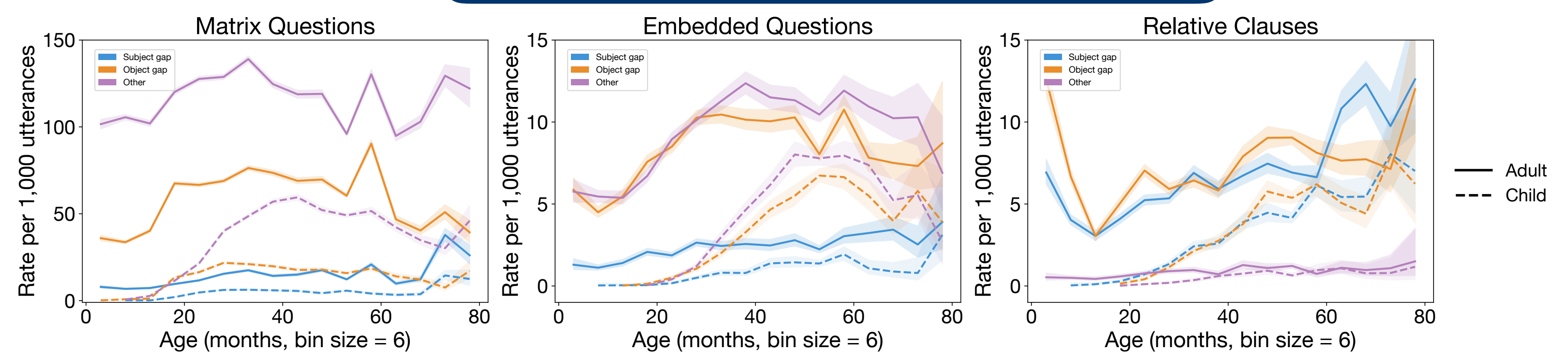
DETECTION STEPS



- **Step 1: Core structure detector**
 - Recursively look for NP -> NP SBAR structure for RC;
- **Step 2: Wh-category Identification**
 - Look for a wh-category immediately dominated by SBAR and retrieve the span of the wh-parse; here, WHNP (*who*);
- **Step 3: Extraction Site Inference**
 - Look for the lowest S node under SBAR, check whether there is an NP preceding a VP; here no means likely subject-extracted.
- **Step 4: Dependency Validation**
 - Verify the relcl relation between the embedded verb and the modified noun, as well as the nsubj relation between embedded verb and the wh-head.

* Verified against human annotated treebanks from Pearl & Sprouse (2013) and achieved > 0.85 F1 scores for most constructions.

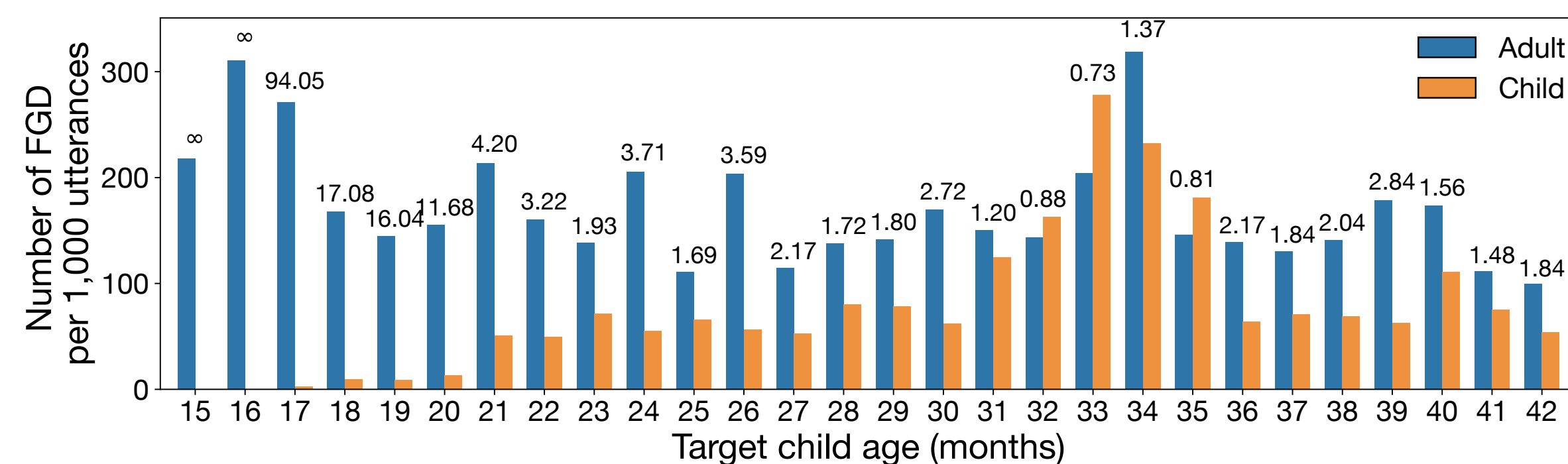
DEVELOPMENTAL DISTRIBUTIONS



- Children's FGD productions broadly track adult input across age, while matrix questions are far more frequent than embedded questions and relative clauses.

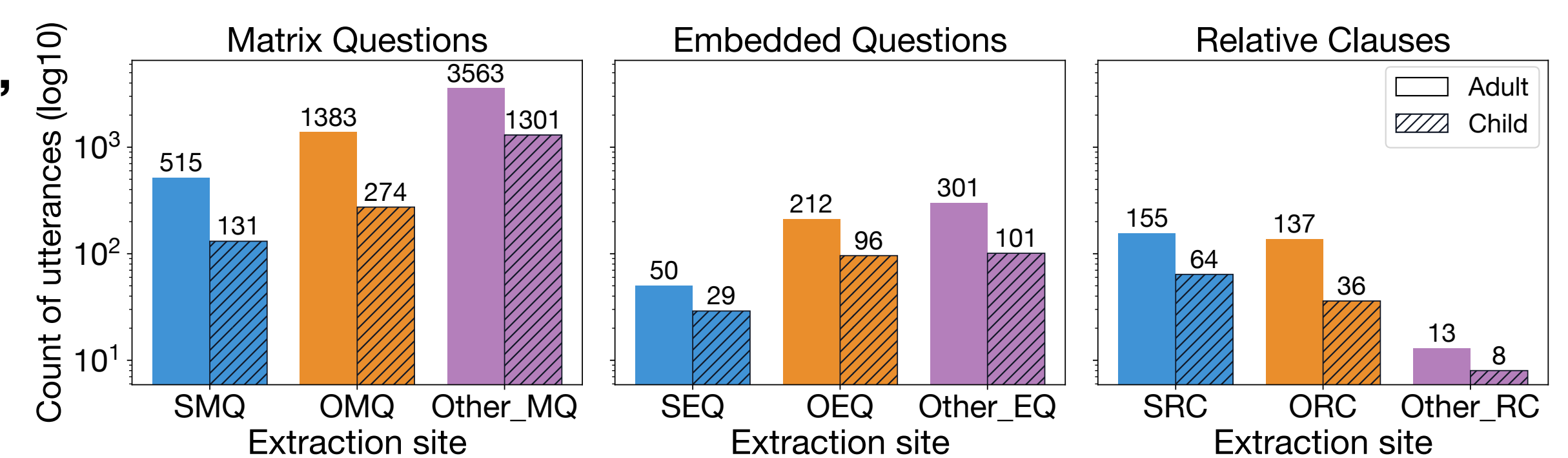
INPUT AND PRODUCTION OF AN INDIVIDUAL CHILD

We studied child Laura from the Braunwald corpus (75,740 utterances across 15-77 months):

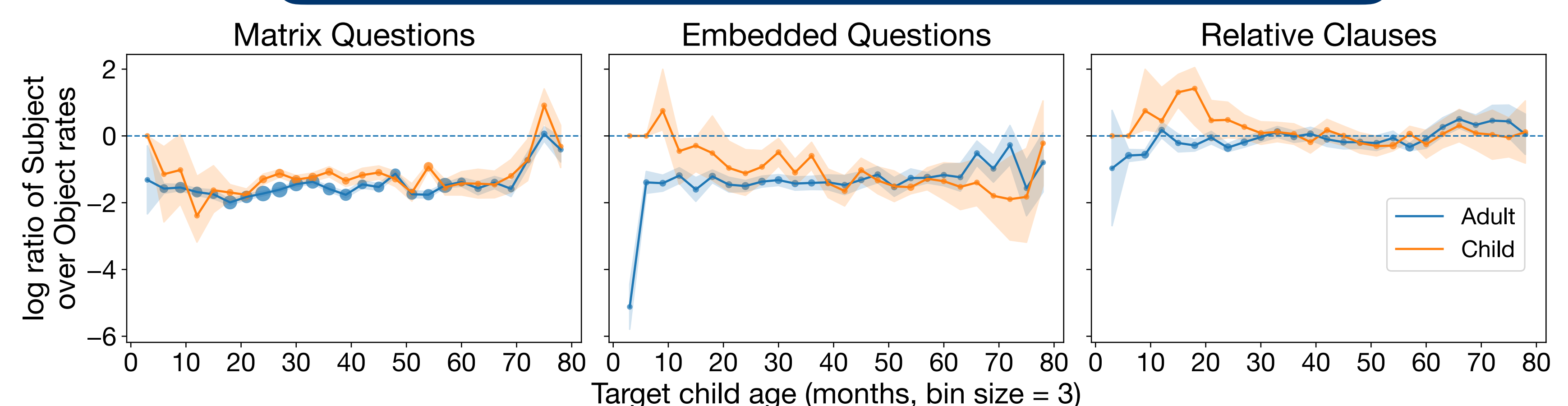


- FGD utterances remain a small but increasing share of child speech;
- Child production gradually approaching and sometimes exceeding adult input rates.

- At the individual-child level, matrix questions dominate both input and production, while embedded questions and relative clauses are much rarer.



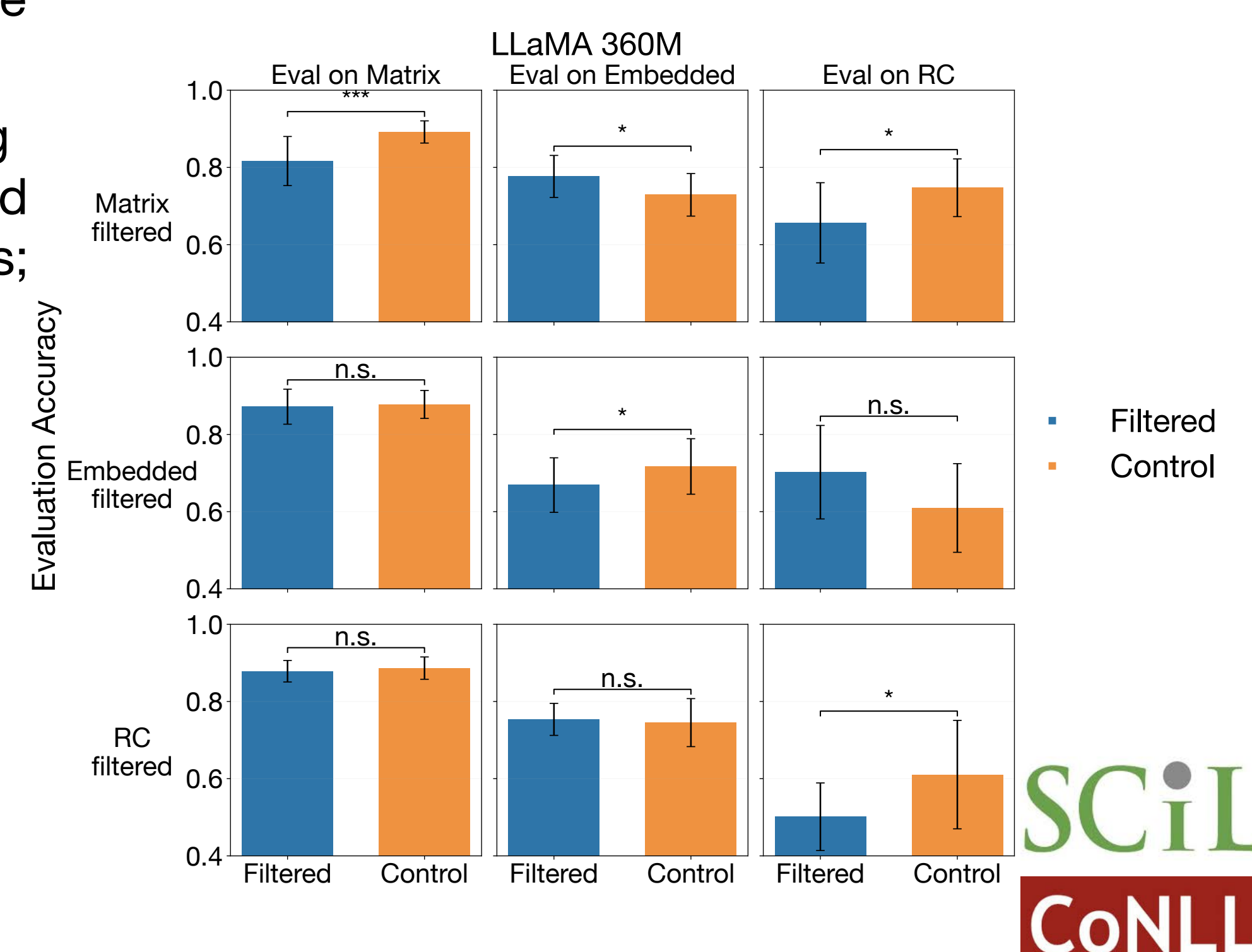
SUBJECT VS. OBJECT EXTRACTION ASYMMETRY



- Wh-question constructions are consistently object-biased, whereas relative clauses are much closer to balanced between subject and object extractions.

BABYLM FILTERED-CORPUS TRAINING

- **Question** = does FGD knowledge generalize across constructions?
- **Method** = filtered-corpus training: removing one FGD construction from training data and evaluate performance in other constructions;
- **Result:**
 - Removing a construction hurts performance on that construction → direct exposure matters for FGD learning.
 - Removing matrix questions (the most frequent) weakens performance on the others, but not vice versa → *asymmetric generalization from frequent to rarer FGD constructions*.



SCiL
CoNLL

