WHAT IS THE REAL BENEFIT OF USING CHILD DIRECTED LANGUAGE FOR LANGUAGE MODELING?

Francesca Padovani, Jaap Jumelet, Yevgen Matusevych, Arianna Bisazza

Center for Language and Cognition (CLCG)
University of Groningen

VIDI PROJECT
Polyglot Machines

DATE
12/06/2025

TABLE OF CONTENT

01

CHILD DIRECTED LANGUAGE

The way caregivers speak to children, characterized by:

- simplified vocabulary
- exaggerated intonation
- repetition
- clear articulation

"Where's your shoe? Your shoe! Let's find your shoe."

"Uh-oh! What happened?"

"Look at the dooooggy!"

Long-standing view → *CDL supports and facilitates early language acquisition*

Foundational work by Ferguson (1964), showing similar speech adaptations and lexical use in CDL across languages (English, Spanish, Arabic, Marathi, Comanche, and Gilyak)

01

CHILD DIRECTED LANGUAGE

The way caregivers speak to children, characterized by:

- simplified vocabulary
- exaggerated intonation
- repetition
- clear articulation

"Where's your shoe? Your shoe! Let's find your shoe."

"Uh-oh! What happened?"

"Look at the dooooggy!"

Long-standing view → *CDL supports and facilitates early language acquisition*

Foundational work by Ferguson (1964), showing similar speech adaptations and lexical use in CDL across languages (English, Spanish, Arabic, Marathi, Comanche, and Gilyak)

Infeasable Experiment

What will happen to a child who grows up immersed only in **encyclopedic language?**

01

CHILD DIRECTED LANGUAGE

The way caregivers speak to children, characterized by:

- simplified vocabulary
- exaggerated intonation
- repetition
- clear articulation

“Where’s your shoe? Your shoe! Let’s find your shoe.”

“Uh-oh! What happened?”

“Look at the dooooggy!”

Infeasable Experiment

What will happen to a child who grows up immersed only in **encyclopedic language?**

Long-standing view → *CDL supports and facilitates early language acquisition*

Foundational work by Ferguson (1964), showing similar speech adaptations and lexical use in CDL across languages (English, Spanish, Arabic, Marathi, Comanche, and Gilyak)

Overextension of this view → Kempe et al. (2024)

Demonstrating that the **evidence** for the facilitatory role of CDL in language acquisition is **scarce** and specific to narrow domains, such as **prosody** and **register discrimination**, raising concerns about its generalizability

02

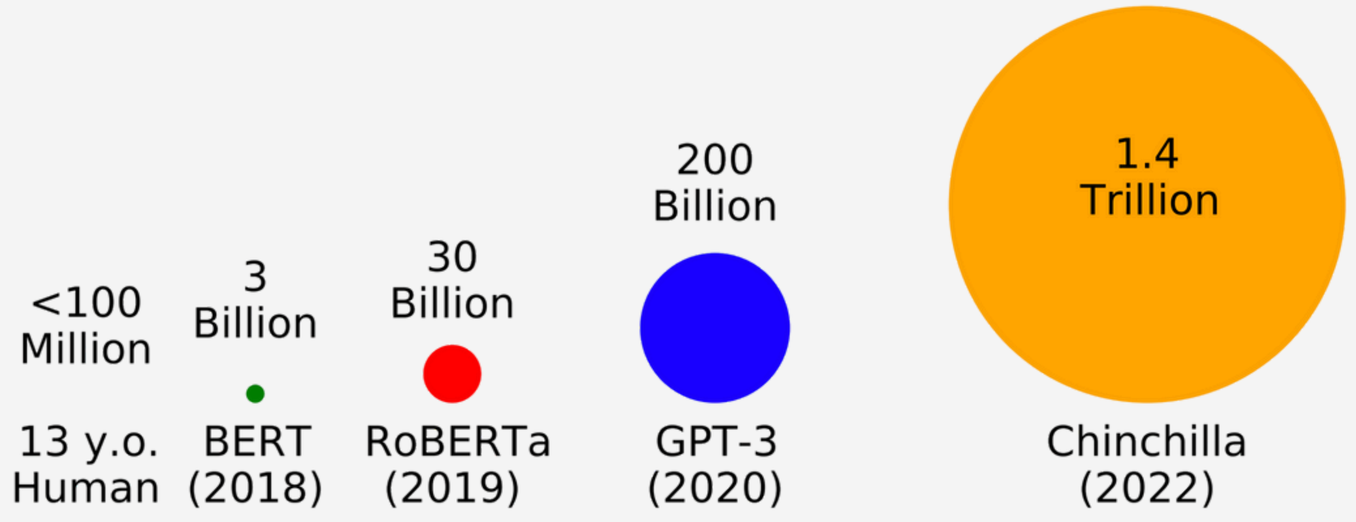
COMPUTATIONAL MODELING RESEARCH

Simulating Language Acquisition

Current paradigm → LLMs

LLMs are trained on cognitively implausible language input:

- size
- **type**



Groundbreaking study → ***BabyBERTa*** (Huebner et al., 2021), a masked LM trained on <u>5M tokens</u> of **CDL**, achieves syntactic ability similar to a much larger RoBERTa model trained on <u>30B tokens</u> of **ADL**.

Growing interest in investigating how training on CDL vs. ADL affects **syntactic learning** and linguistic competence generalization in language models (LMs)

02

COMPUTATIONAL MODELING RESEARCH

After BabyBERTa…conflicting results in the Literature

some findings highlight CDL’s **benefits** for *grammatical learning* and *inductive bias*

- Huebner et al., 2021 (English)
- You et al., 2021 (English)
- Mueller and Linzen, 2023 (English)
- Salhan et al., 2024 (French, German, Japanese and Chinese)

others find little or **no advantage**

- Gelderloos et al., 2020 (English)
- Yedetore et al., 2023 (English)
- Feng et al., 2024 (English)
- Bunzeck et al., 2025 (German)

03

REASSESSMENT OF CDL IMPACT

What complicates the comparison between these works?

- Variability in **training setups** (e.g. model use, hyperparameters settings etc)
- **Data selection** and **preprocessing** (e.g. filtering out children utterances, encoding of speaker role etc..)
- **Evaluation benchmarks** with different shortcomings

03

REASSESSMENT OF CDL IMPACT

What complicates the comparison between these works?

- Variability in **training setups** (e.g. model use, hyperparameters settings etc)
- **Data selection** and **preprocessing** (e.g. filtering out children utterances, encoding of speaker role etc..)
- **Evaluation benchmarks** with different shortcomings

Therefore, in this work we:

Systematically compare LMs trained on a **size-matched dataset** of **CHILDES** (CDL) **vs Wikipedia** (ADL)

03

REASSESSMENT OF CDL IMPACT

What complicates the comparison between these works?

- Variability in **training setups** (e.g. model use, hyperparameters settings etc)
- **Data selection** and **preprocessing** (e.g. filtering out children utterances, encoding of speaker role etc..)
- **Evaluation benchmarks** with different shortcomings

Therefore, in this work we:

→ Systematically compare LMs trained on a **size-matched dataset** of **CHILDES** (CDL) **vs Wikipedia** (ADL)

→ Across two architectures, **RoBERTa (MLM) and GPT-2 (CLM)**

03 REASSESSMENT OF CDL IMPACT

What complicates the comparison between these works?

- Variability in **training setups** (e.g. model use, hyperparameters settings etc)
- **Data selection** and **preprocessing** (e.g. filtering out children utterances, encoding of speaker role etc..)
- **Evaluation benchmarks** with different shortcomings

Therefore, in this work we:

→ Systematically compare LMs trained on a **size-matched dataset** of **CHILDES** (CDL) **vs Wikipedia** (ADL)

→ Across two architectures, **RoBERTa (MLM) and GPT-2 (CLM)**

→ Three languages → **English, French, German**

03 REASSESSMENT OF CDL IMPACT

What complicates the comparison between these works?

- Variability in **training setups** (e.g. model use, hyperparameters settings etc)
- **Data selection** and **preprocessing** (e.g. filtering out children utterances, encoding of speaker role etc..)
- **Evaluation benchmarks** with different shortcomings

Therefore, in this work we:

→ Systematically compare LMs trained on **CHILDES (CDL) vs Wikipedia (ADL)**

→ Across two architectures, **RoBERTa (MLM) and GPT-2 (CLM)**

→ Three languages → **English, French, German**

→ Using **four evaluation benchmarks of minimal pairs** to assess formal grammatical competence (focus on syntax)

03 REASSESSMENT OF CDL IMPACT

What complicates the comparison between these works?

- Variability in **training setups** (e.g. model use, hyperparameters settings etc)
- **Data selection** and **preprocessing** (e.g. filtering out children utterances, encoding of speaker role etc..)
- **Evaluation benchmarks** with different shortcomings

Therefore, in this work we:

→ Systematically compare LMs trained on **CHILDES (CDL) vs Wikipedia (ADL)**

→ Across two architectures, **RoBERTa (MLM) and GPT-2 (CLM)**

→ Three languages → **English, French, German**

→ Using **four evaluation benchmarks of minimal pairs** to assess formal grammatical competence (focus on syntax)

→ We use already available benchmarks **BLiMP, Zorro and CLAMS** (Warstadt et al., 2020; Huebner et al., 2021; Mueller et al., 2020)

→ We introduce a new **Frequency-Informed Testing (FIT) methodology** which we apply to the CLAMS benchmark

03

ALREADY EXISTING MINIMAL PAIR BENCHAMARKS

e.g. CLAMS (Warstadt et al., 2020)

Paradigm	Minimal Pair
Simple Agreement	*the pilot [smiles/*smile]*
Agreement in prepositional phrases	*the author next to the guard [laughs/*laugh]*
Agreement in subject relative clauses	*the surgeon that admires the guard [is/*are]* young
Agreement in object relative clauses	*the senator that the ministers admire [swims/*swim]*
Agreement in VP coordinates	*the farmer is short and [laughs/*laugh]*
Agreement in long VP coordinates	*the surgeon that admires the guard [is/*are]* young

03

RESULTS ON ALREADY EXISTING MINIMAL PAIR BENCHAMARKS

CHILDES ≈ Wiki

Model	Training Data	BLiMP	Zorro	CLAMS English	CLAMS French	CLAMS German
CLM	CHILDES	0.61 ± 0.02	$\mathbf{0.76 \pm 0.04}$	0.60 ± 0.01	0.64 ± 0.01	0.69 ± 0.03
	Wiki	0.61 ± 0.02	0.69 ± 0.04	$\mathbf{0.71 \pm 0.01}$	$\mathbf{0.80 \pm 0.01}$	$\mathbf{0.81 \pm 0.01}$
MLM	CHILDES	0.59 ± 0.03	0.66 ± 0.05	0.57 ± 0.02	0.59 ± 0.02	0.70 ± 0.01
	Wiki	0.59 ± 0.03	0.67 ± 0.03	$\mathbf{0.63 \pm 0.01}$	$\mathbf{0.69 \pm 0.01}$	$\mathbf{0.75 \pm 0.01}$

Table 2: Model accuracies on BLiMP, Zorro, and CLAMS, averaged across paradigms and model seeds.

RESULTS ON ALREADY EXISTING MINIMAL PAIR BENCHAMARKS

CHILDES ≈ Wiki CHILDES ≳ Wiki

Model	Training Data	BLiMP	Zorro	CLAMS English	CLAMS French	CLAMS German
CLM	CHILDES	0.61 ± 0.02	$\mathbf{0.76 \pm 0.04}$	0.60 ± 0.01	0.64 ± 0.01	0.69 ± 0.03
	Wiki	0.61 ± 0.02	0.69 ± 0.04	$\mathbf{0.71 \pm 0.01}$	$\mathbf{0.80 \pm 0.01}$	$\mathbf{0.81 \pm 0.01}$
MLM	CHILDES	0.59 ± 0.03	0.66 ± 0.05	0.57 ± 0.02	0.59 ± 0.02	0.70 ± 0.01
	Wiki	0.59 ± 0.03	0.67 ± 0.03	$\mathbf{0.63 \pm 0.01}$	$\mathbf{0.69 \pm 0.01}$	$\mathbf{0.75 \pm 0.01}$

Table 2: Model accuracies on BLiMP, Zorro, and CLAMS, averaged across paradigms and model seeds.

03

RESULTS ON ALREADY EXISTING MINIMAL PAIR BENCHAMARKS

CHILDES $\approx$ Wiki — CHILDES $\gtrsim$ Wiki — CHILDES $\ll$ Wiki

Model	Training Data	BLiMP	Zorro	CLAMS English	CLAMS French	CLAMS German
CLM	CHILDES	0.61 ± 0.02	**0.76 ± 0.04**	0.60 ± 0.01	0.64 ± 0.01	0.69 ± 0.03
	Wiki	0.61 ± 0.02	0.69 ± 0.04	**0.71 ± 0.01**	**0.80 ± 0.01**	**0.81 ± 0.01**
MLM	CHILDES	0.59 ± 0.03	0.66 ± 0.05	0.57 ± 0.02	0.59 ± 0.02	0.70 ± 0.01
	Wiki	0.59 ± 0.03	0.67 ± 0.03	**0.63 ± 0.01**	**0.69 ± 0.01**	**0.75 ± 0.01**

Table 2: Model accuracies on BLiMP, Zorro, and CLAMS, averaged across paradigms and model seeds.

Why do we observe these mixed results?

03

RESULTS ON ALREADY EXISTING MINIMAL PAIR BENCHAMARKS

CHILDES $\approx$ Wiki — CHILDES $\gtrsim$ Wiki — CHILDES $\ll$ Wiki

Model	Training Data	BLiMP	Zorro	CLAMS English	CLAMS French	CLAMS German
CLM	CHILDES	0.61 ± 0.02	**0.76 ± 0.04**	0.60 ± 0.01	0.64 ± 0.01	0.69 ± 0.03
	Wiki	0.61 ± 0.02	0.69 ± 0.04	**0.71 ± 0.01**	**0.80 ± 0.01**	**0.81 ± 0.01**
MLM	CHILDES	0.59 ± 0.03	0.66 ± 0.05	0.57 ± 0.02	0.59 ± 0.02	0.70 ± 0.01
	Wiki	0.59 ± 0.03	0.67 ± 0.03	**0.63 ± 0.01**	**0.69 ± 0.01**	**0.75 ± 0.01**

Table 2: Model accuracies on BLiMP, Zorro, and CLAMS, averaged across paradigms and model seeds.

Why do we observe these mixed results?

Potential problems with existing benchmarks

BLIMP *(67 paradigms - 12 linguistic phenomena)* → semi-automated generation process with lexical items systematically varied within manually crafted sentence templates. This approach still produces **semantically odd or implausible sentences** and **does not account for the vocabulary typical of CDL**.

ZORRO (23 paradigms - 13 linguistic phenomena) → **improves lexical alignment** between CDL and ADL, but **excludes subword-split items** limiting its fairness and generalizability for evaluating true syntactic competence in language models.

CLAMS (7 paradigms - 1 linguistic phenomena) → **lexical frequencies** may be an important **confounder** in the evaluation.

04

FIT-CLAMS FREQUENCY INFORMED TESTING METHODOLOGY APPLIED TO CLAMS

We generate **two sets of minimal pairs**, each *guided by* the **lexical distribution** of each training corpus (CHILDES vs Wikipedia), ensuring a spread of high- and low-frequency items.

Nouns	**Bin**	**Freq**	**Df**
roommate, roommates	*0*	*2*	*CHILDES*
resident, residents	*1*	*6*	*CHILDES*
librarian, librarians	*2*	*13*	*CHILDES*
officer, officers	*3*	*36*	*CHILDES*
toddler, toddlers	*4*	*90*	*CHILDES*
farmer, farmers	*5*	*264*	*CHILDES*
policeman, policemen	*6*	*380*	*CHILDES*
doctor, doctors	*7*	*656*	*CHILDES*
man, men	*8*	*2156*	*CHILDES*
daddy, daddies	*9*	*7027*	*CHILDES*

Verbs	**Bin**	**Freq**	**Long VP**	**Df**
await, awaits	*0*	*2*	*the guests*	*CHILDES*
complains, complain	*1*	*8*	*about the noise*	*CHILDES*
arrives, arrive	*2*	*17*	*at the station*	*CHILDES*
disappears, disappear	*2*	*42*	*from the scene*	*CHILDES*
bows, bow	*4*	*243*	*to the king*	*CHILDES*
hides, hide	*4*	*391*	*from the chicken*	*CHILDES*
leaves, leave	*6*	*1793*	*the room*	*CHILDES*
sits, sit	*7*	*4219*	*in the car*	*CHILDES*
thinks, think	*8*	*2156*	*about the trip*	*CHILDES*
goes, go	*9*	*27620*	*to the new store*	*CHILDES*

We do the same for Wikipedia and for all the three languages!

RESULTS ON FIT-CLAMS

Training	Eval. lex.	EN	FR	DE
CHILDES	CHILDES	0.63 ± 0.02	0.78 ± 0.04	0.73 ± 0.03
	Wiki	0.63 ± 0.03	0.67 ± 0.03	0.69 ± 0.04
Wiki	CHILDES	$\mathbf{0.72 \pm 0.03}$	$\mathbf{0.86 \pm 0.02}$	$\mathbf{0.83 \pm 0.02}$
	Wiki	$\mathbf{0.75 \pm 0.02}$	$\mathbf{0.88 \pm 0.06}$	$\mathbf{0.82 \pm 0.03}$

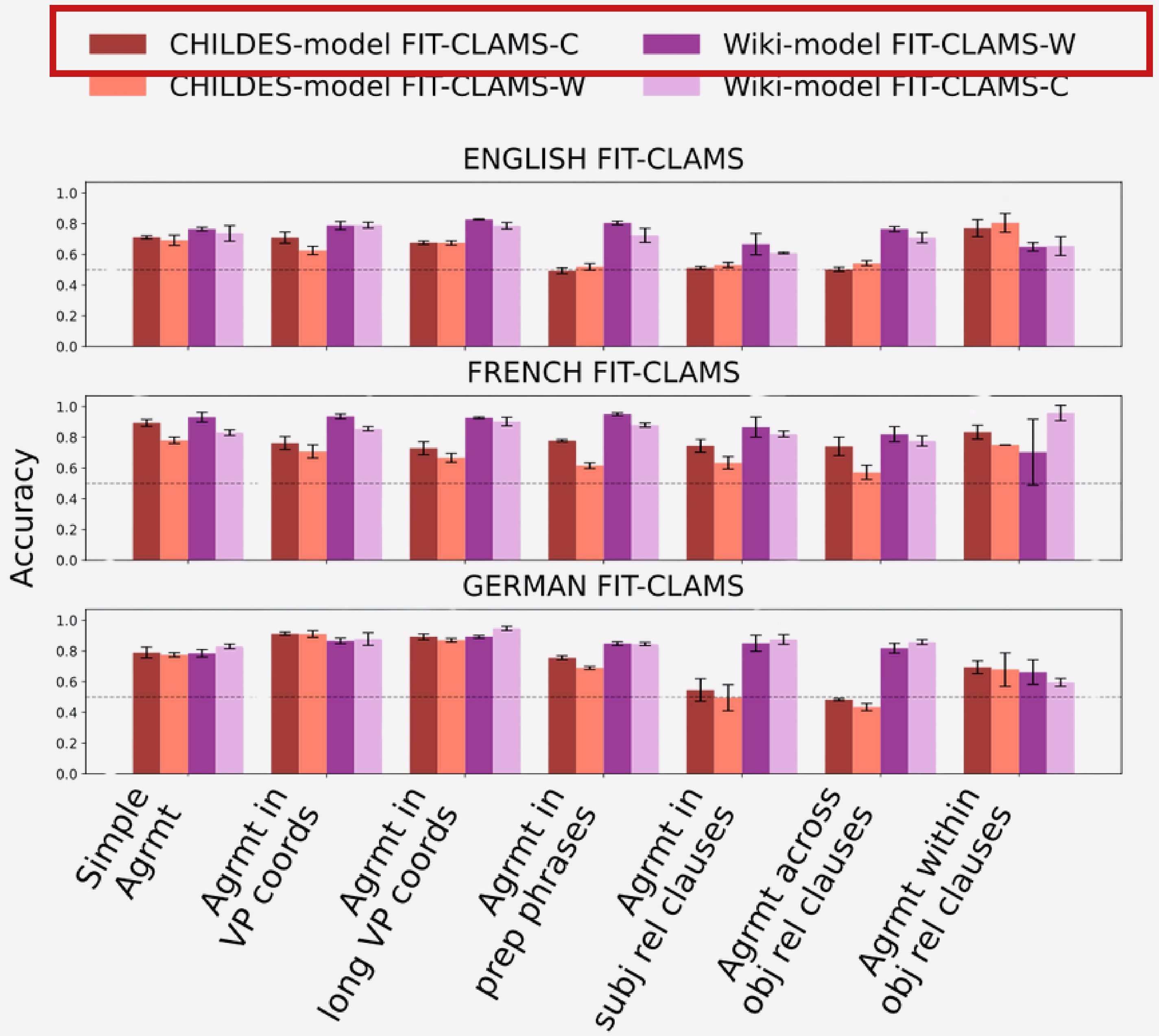
OBSERVATIONS:

1. Overall **better performance** than on CLAMS
2. Better performance on **in-distribution** than on **out-distribution** (except the German model trained on Wikipedia)
3. Importantly, the **most pronounced contrast** is still between **CHILDES vs Wikipedia**

04

RESULTS ON FIT-CLAMS

Training	Eval. lex.	EN	FR	DE
CHILDES	CHILDES	0.63 ± 0.02	0.78 ± 0.04	0.73 ± 0.03
	Wiki	0.63 ± 0.03	0.67 ± 0.03	0.69 ± 0.04
Wiki	CHILDES	**0.72 ± 0.03**	**0.86 ± 0.02**	**0.83 ± 0.02**
	Wiki	**0.75 ± 0.02**	**0.88 ± 0.06**	**0.82 ± 0.03**

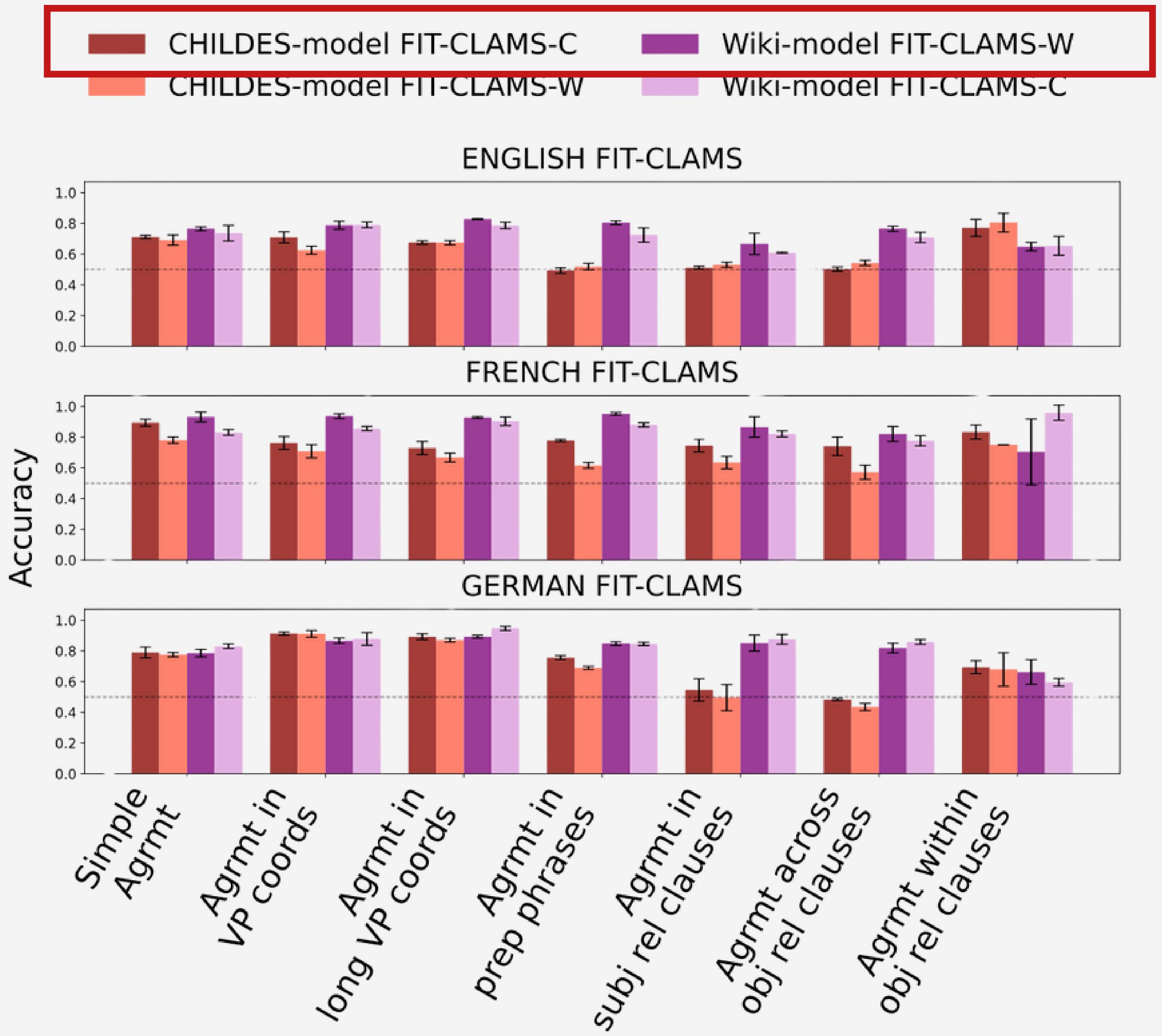


OBSERVATIONS:

1. Overall **better performance** than on CLAMS
2. Better performance on **in-distribution** than on **out-distribution** (except the German model trained on Wikipedia)
3. Importantly, the **most pronounced contrast** is still between **CHILDES vs Wikipedia**

04 RESULTS ON FIT-CLAMS

Training	Eval. lex.	EN	FR	DE
CHILDES	CHILDES	0.63 ± 0.02	0.78 ± 0.04	0.73 ± 0.03
	Wiki	0.63 ± 0.03	0.67 ± 0.03	0.69 ± 0.04
Wiki	CHILDES	**0.72 ± 0.03**	**0.86 ± 0.02**	**0.83 ± 0.02**
	Wiki	**0.75 ± 0.02**	**0.88 ± 0.06**	**0.82 ± 0.03**



OBSERVATIONS:

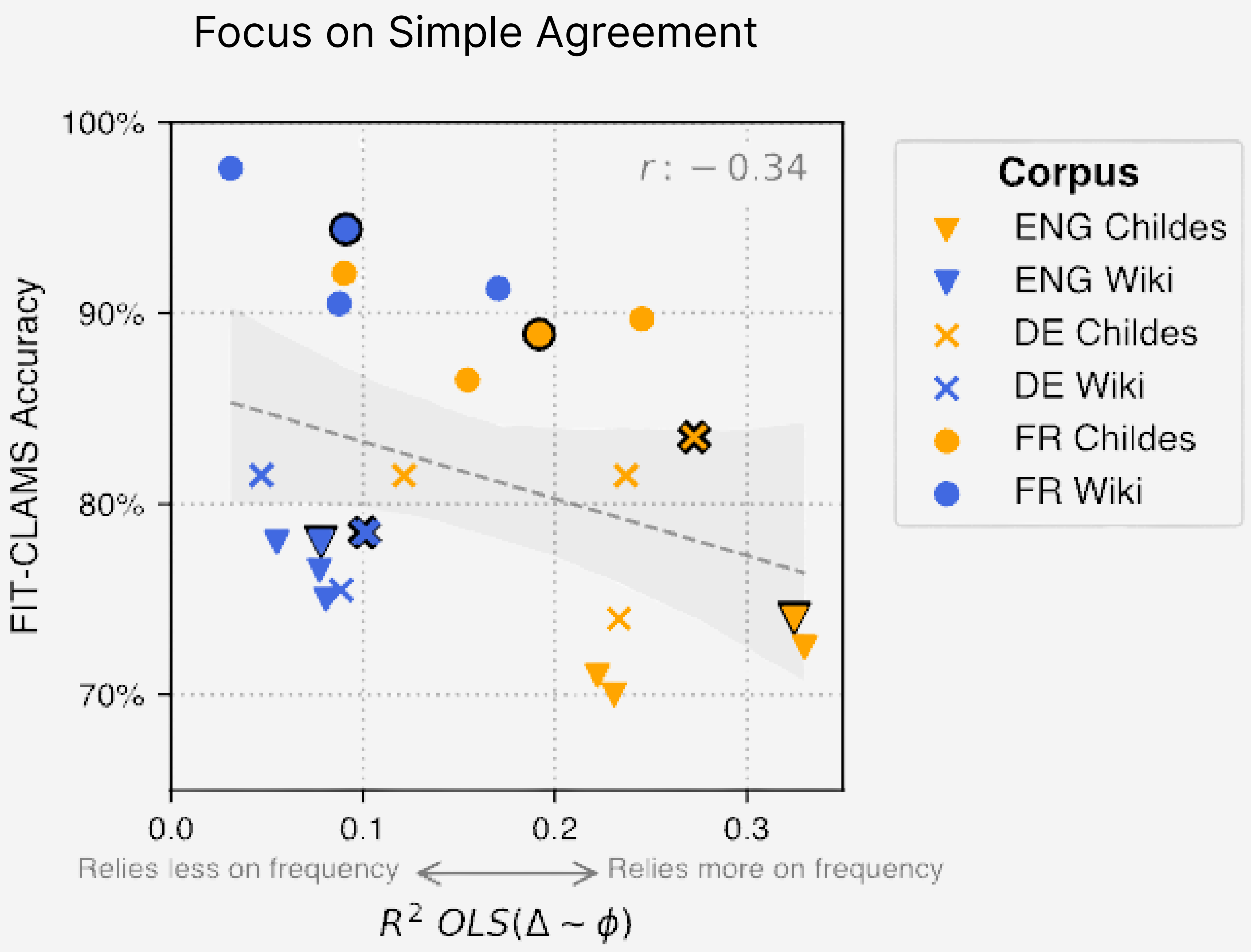
1. Overall **better performance** than on CLAMS
2. Better performance on **in-distribution** than on **out-distribution** (except the German model trained on Wikipedia)
3. Importantly, the **most pronounced contrast** is still between **CHILDES vs Wikipedia**

Even when strictly controlling for lexical frequency, models trained on Wikipedia continue to show a systematic advantage.

05

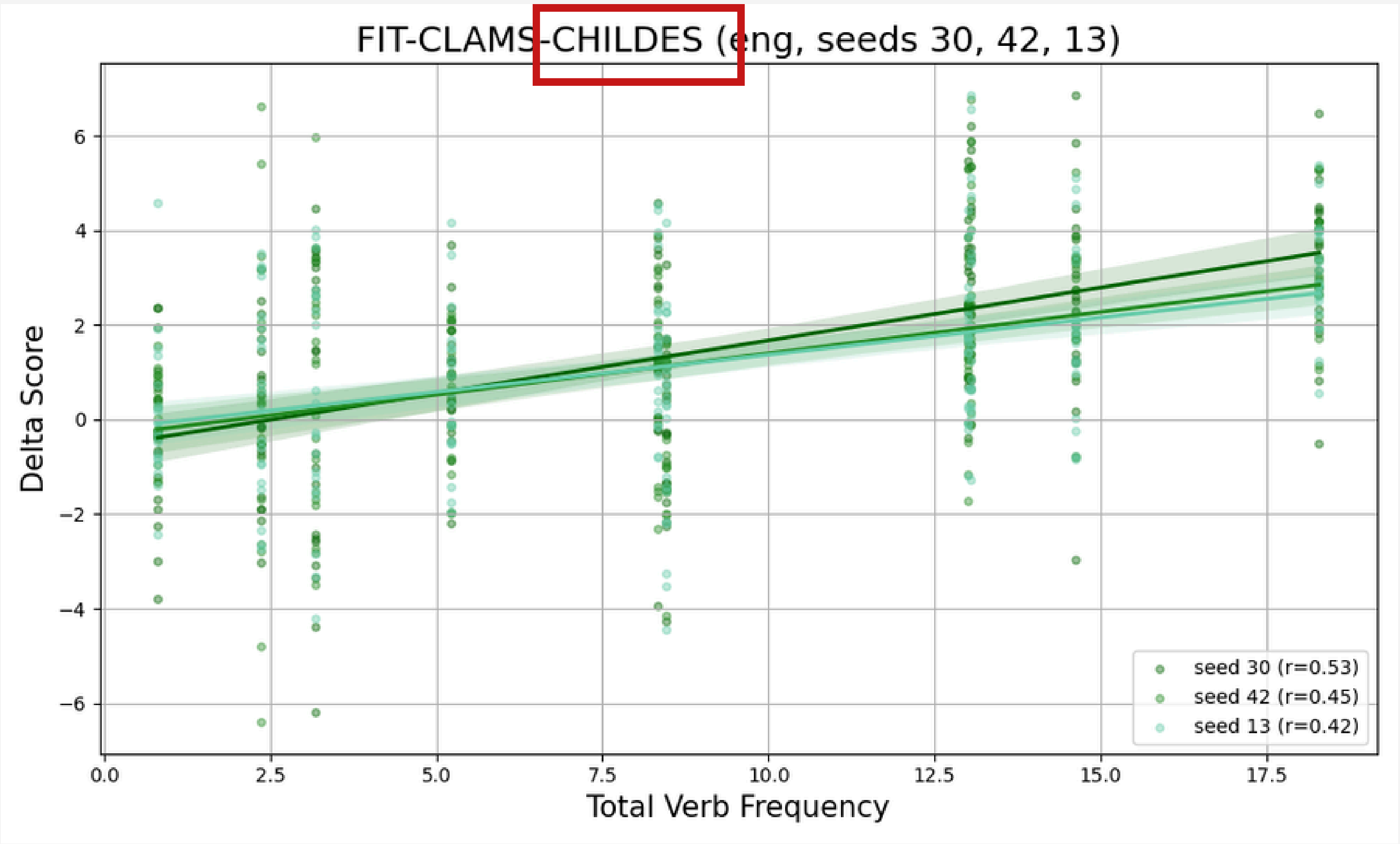
REGRESSION ANALYSIS IMPACT OF DISTRIBUTIONAL PROPERTIES

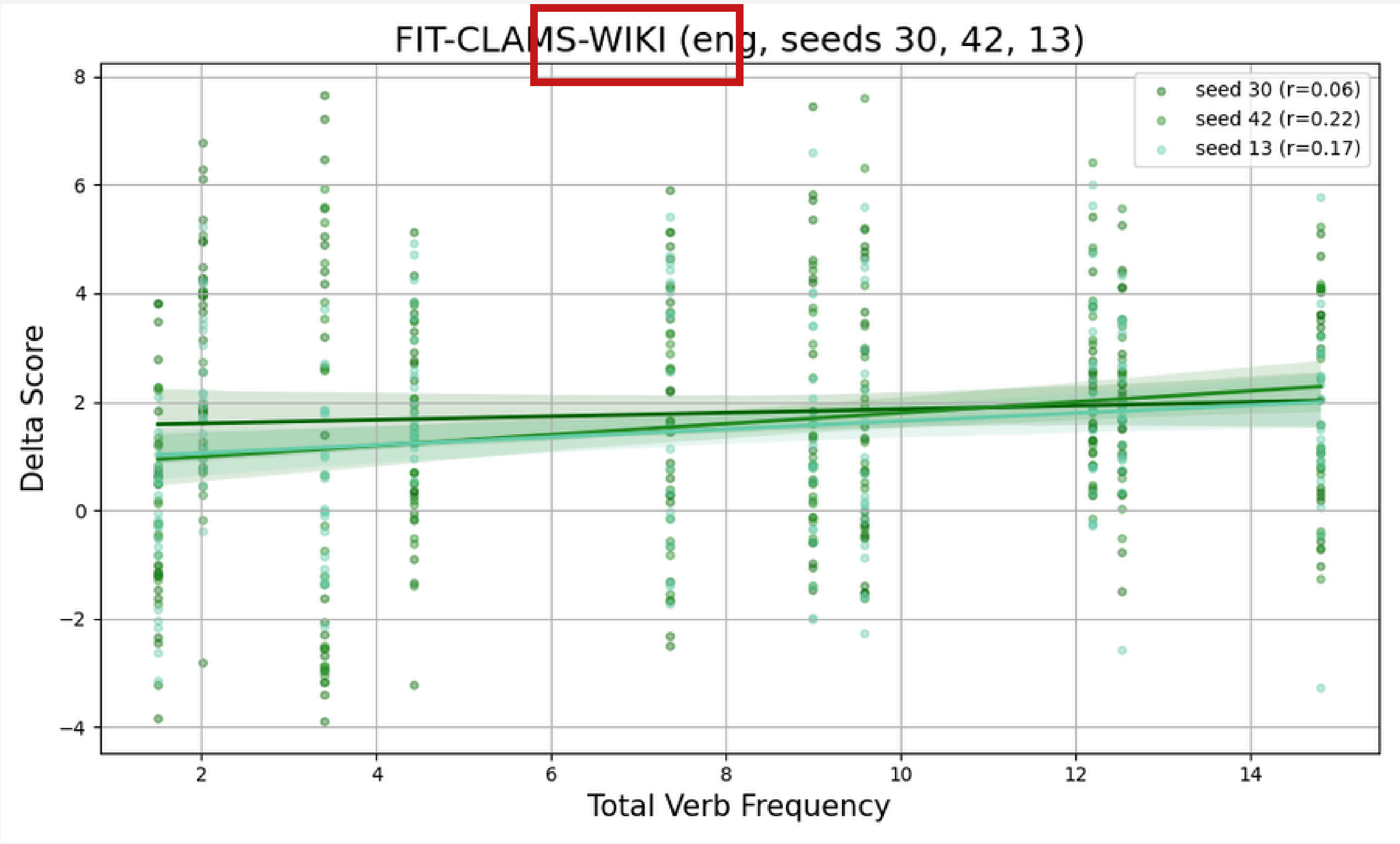
A model that builds up a robust representation of number agreement will be better able to generalize to infrequent constructions, without relying on memorization (Lakretz et al., 2019; Patil et al., 2024).



Relation between LM accuracy on FIT-CLAMS and proportion of variance (R^2) explained by the OLS regression fitted on lexical frequency factors. **The lower the R^2 is, the less the LM's behavior is driven by lexical frequency**.

05 REGRESSION ANALYSIS IMPACT OF DISTRIBUTIONAL PROPERTIES





- **Frequency** significantly **correlates** with **performance** for CHILDES-models
- This holds also for the other two languages with a less pronounced difference between CHILDES and Wikipedia

06 RECONTEXTUALIZATION OF CDL

🧠 Limitations of Current Modeling Approaches

- Language models are trained in ***static***, *non-interactive environments*, unlike human learners
- No **feedback**, **developmental grounding**, or **cognitive constraints** (e.g., working memory)

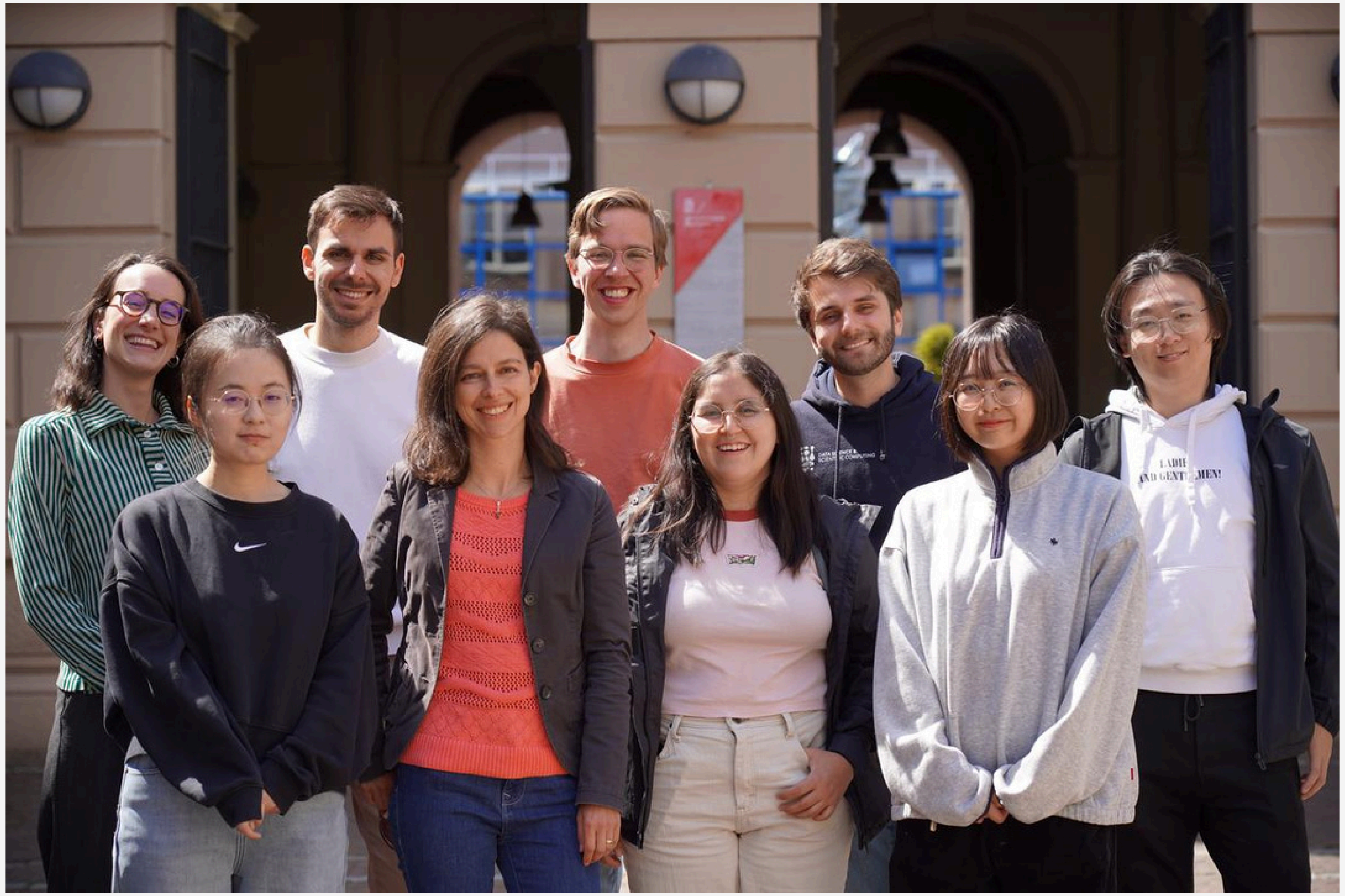
🔍 Rethinking CDL in Modeling

- CDL might hold particular promise when integrated into models that simulate interactive, situated communication (Beuls and Van Eecke, 2024; Stöpler et al., 2025), shifting the focus toward the **communicative and contextual factors essential to language acquisition**, which are absent in static text-based training regimes.

- LM experiments can still **contribute significantly to the study of human language acquisition**, where the benefits of CDL remain poorly understood (Kempe et al., 2024), by helping to uncover specific properties of CDL that make it particularly suitable for specific kinds of learning outcomes

07 THANK YOU ! : -))

Scan the QR code to access the paper, our group website and my personal website!



07 REFERENCES

Blasi, Dami.n, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic Inequalities in Language Technology Performance across the World’s Languages. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics.

Bunzeck, Bastian, and Sina Zarrieß. "Fifty shapes of BLiMP: syntactic learning curves in language models are not uniform, but sometimes unruly." Proceedings of the 2024 CLASP Conference on Multimodality and Interaction in Language Learning. 2024.

Huebner, Philip A., et al. "BabyBERTa: Learning more grammar with small-scale child-directed language." Proceedings of the 25th conference on computational natural language learning. 2021.

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 6282-6293.

Lester, Nicholas A., Steven Moran, Aylin C. Küntay, Shanley E.M. Allen, Barbara Pfeiler, Sabine Stoll. 2022. Detecting structured repetition in child-surrounding speech: Evidence from maximally diverse languages. Cognition, Volume 221.

Mueller, Aaron, et al. "Cross-linguistic syntactic evaluation of word prediction models." *arXiv preprint arXiv:2005.00187* (2020).

Onnis, Luca, Heidi R. Waterfall, and Shimon Edelman. "Learn locally, act globally: Learning language from variation set cues." *Cognition* 109.3 (2008): 423-430

Rüst, Olivier, et al. "The Acquisition of Case Systems in Typologically Diverse Languages: Children Gradually Generalize Grammatical Rules." (2021): 672-685

Stumper, Barbara, et al. "“Frequent frames” in German child-directed speech: A limited cue to grammatical categories." Cognitive science 35.6 (2011): 1190-1205.

You, Guanghao, et al. "Child-directed speech is optimized for syntax-free semantic inference." *Scientific Reports* 11.1 (2021): 16527.

Warstadt, Alex, et al. "BLiMP: The benchmark of linguistic minimal pairs for English." Transactions of the Association for Computational Linguistics 8 (2020): 377-392.