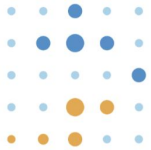


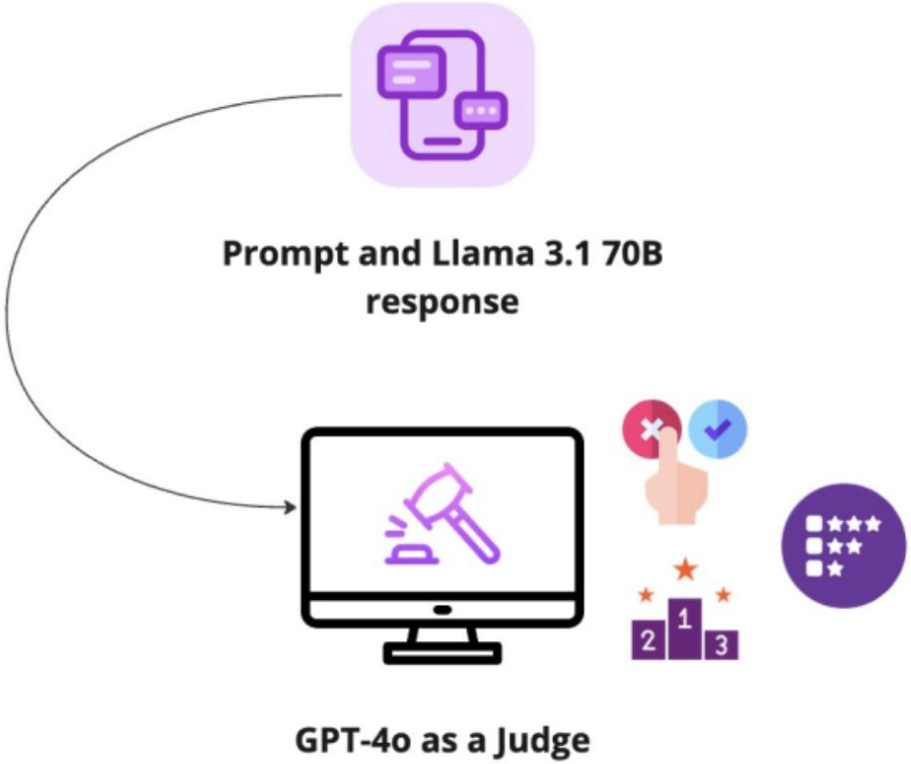
NATURAL LANGUAGE PROCESSING

UNIVERSITY OF CAMBRIDGE

# Human Validated Grammar Profiles

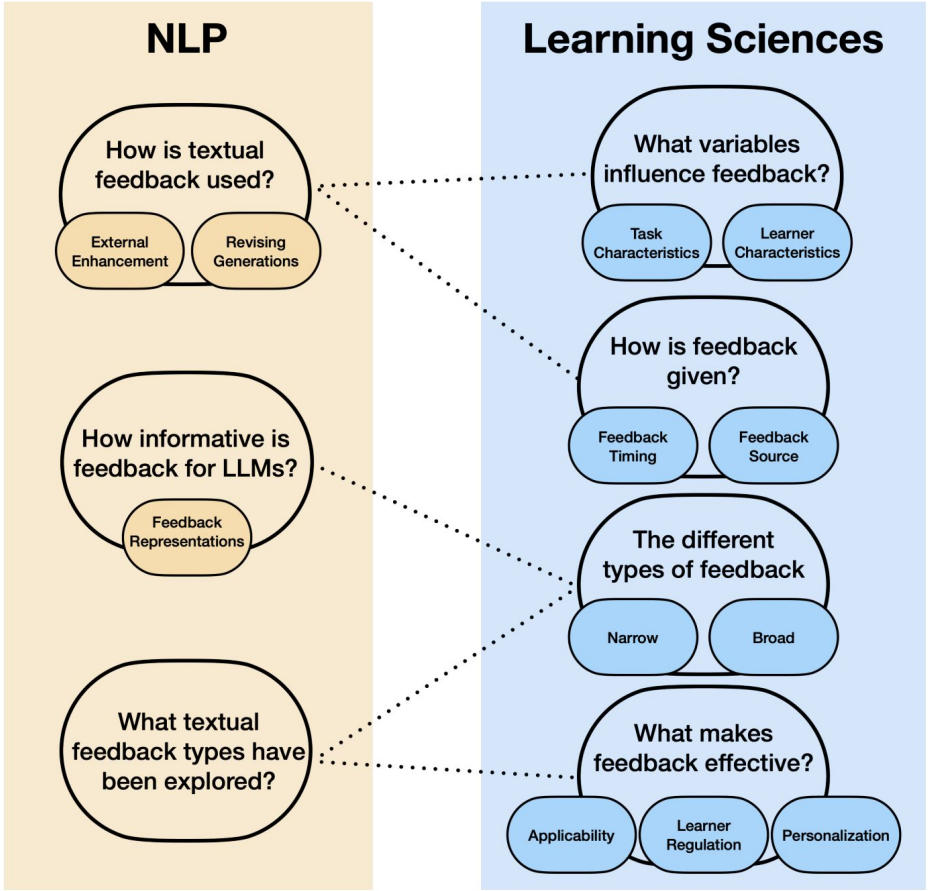
**Suchir Salhan**
sas245@cam.ac.uk

# A Starting Point: Connecting LLMs to Learning




LLMs are, in some sense, ***learners*** and not “*reliable judges*”. “LLM as Judge” is a naive and potentially overly optimistic perspective.

# A Starting Point: Connecting LLMs to Learning




LLMs are, in some sense, ***learners*** and not "*reliable judges*". "LLM as Judge" is a naive and potentially overly optimistic perspective.

# BabyLM Challenge

Instead of chasing trillion-parameter models, wouldn't we all be better off if

we built small and efficient alternatives to LLMs that can be pretrained from scratch to solve real-world ML/NLP problems?

we could design Language Models with a "cognitively-plausible" architecture?

# BabyLM Challenge

Instead of chasing trillion-parameter models, wouldn't we all be better off if

we built small and efficient alternatives to LLMs that can be pretrained from scratch to solve real-world ML/NLP problems?

we could design Language Models with a "cognitively-plausible" architecture?

🌱 Environmentally-Friendly
🤝 Open-Access Foundation Models

🧠 Cognitive Science
💬 First Language (L1) Acquisition

# Evaluation in the BabyLM Challenge

| Model | BLiMP | BLiMP Suppl. | EWoK | GLUE | Av. |
|---|---|---|---|---|---|
| BabyLlama | 69.8 | 59.5 | 50.7 | 63.3 | 60.8 |
| LTG-BERT | 60.6 | 60.8 | 48.9 | 60.3 | 57.7 |

**Table 1** Example of Language Model Evaluation from the BabyLM Shared Task 2024

# Evaluation in the BabyLM Challenge

| Model | BLiMP | BLiMP Suppl. | EWoK | GLUE | Av. |
|---|---|---|---|---|---|
| BabyLlama | 69.8 | 59.5 | 50.7 | 63.3 | 60.8 |
| LTG-BERT | 60.6 | 60.8 | 48.9 | 60.3 | 57.7 |

**Table 1** Example of Language Model Evaluation from the BabyLM Shared Task 2024

i. **Benchmark of Linguistic Minimal Pairs for English (BLiMP) (Warstadt, Parrish, Liu, Mohananey, Peng, Wang & Bowman 2020):** This is a metric for formal linguistic competence, comparing the predictions at a **critical word** in a syntactically acceptable and unacceptable minimal pair. The sentences only differ with respect to a single feature, and success is determined if $P(w_{\text{c, acceptable}}) > (P(w_{\text{c, unacceptable}})$ for a critical word $w_c$.

ii. **SuperGLUE (Wang, Pruksachatkun, Nangia, Singh, Michael, Hill, Levy & Bowman 2019):** A proxy for the "functional competence" of a language model (Steuer, Mosbach & Klakow 2023), SuperGLUE evaluates for a wide range of natural language understanding (NLU) problems, including question answering, natural language inference and linguistic acceptability judgements.

# A Closer Look at BLiMP

| Phenomenon | N | Acceptable Example | Unacceptable Example |
|---|---|---|---|
| ANAPHOR AGR. | 2 | *Many girls insulted themselves.* | *Many girls insulted herself.* |
| ARG. STRUCTURE | 9 | *Rose wasn't disturbing Mark.* | *Rose wasn't boasting Mark.* |
| BINDING | 7 | *Carlos said that Lori helped him.* | *Carlos said that Lori helped himself.* |
| CONTROL/RAISING | 5 | *There was bound to be a fish escaping.* | *There was unable to be a fish escaping.* |
| DET.-NOUN AGR. | 8 | *Rachelle had bought that chair.* | *Rachelle had bought that chairs.* |
| ELLIPSIS | 2 | *Anne's doctor cleans one important book and Stacey cleans a few.* | *Anne's doctor cleans one book and Stacey cleans a few important.* |
| FILLER-GAP | 7 | *Brett knew what many waiters find.* | *Brett knew that many waiters find.* |
| IRREGULAR FORMS | 2 | *Aaron broke the unicycle.* | *Aaron broken the unicycle.* |
| ISLAND EFFECTS | 8 | *Whose hat should Tonya wear?* | *Whose should Tonya wear hat?* |
| NPI LICENSING | 7 | *The truck has clearly tipped over.* | *The truck has ever tipped over.* |
| QUANTIFIERS | 4 | *No boy knew fewer than six guys.* | *No boy knew at most six guys.* |
| SUBJECT-VERB AGR. | 6 | *These casseroles disgust Kayla.* | *These casseroles disgusts Kayla.* |

# Evaluating Syntax

a. *Adjunct islands*
*I know what the patron got mad after the librarian placed ___ on the wrong shelf.
b. *Complex NP islands*
*I know what the actress bought the painting that depicted ___ yesterday.
c. *Coordinate structure islands*
*I know what the man bought and ___ at the antique shop.
d. *Left-branch islands*
*I know how expensive you bought ___ a car last week.
e. *Sentential subject islands*
*I know who for the seniors to defeat ___ will be trivial.
f. *Subject islands*
*I know who the painting by ___ fetched a high price.
g. Wh-*islands*
*I know who Alex said whether your friend insulted ___ yesterday.

Figure 3: Islands associated with syntactic constraints, based on Ross (1967) and Huang (1982)

# Evaluating Syntax

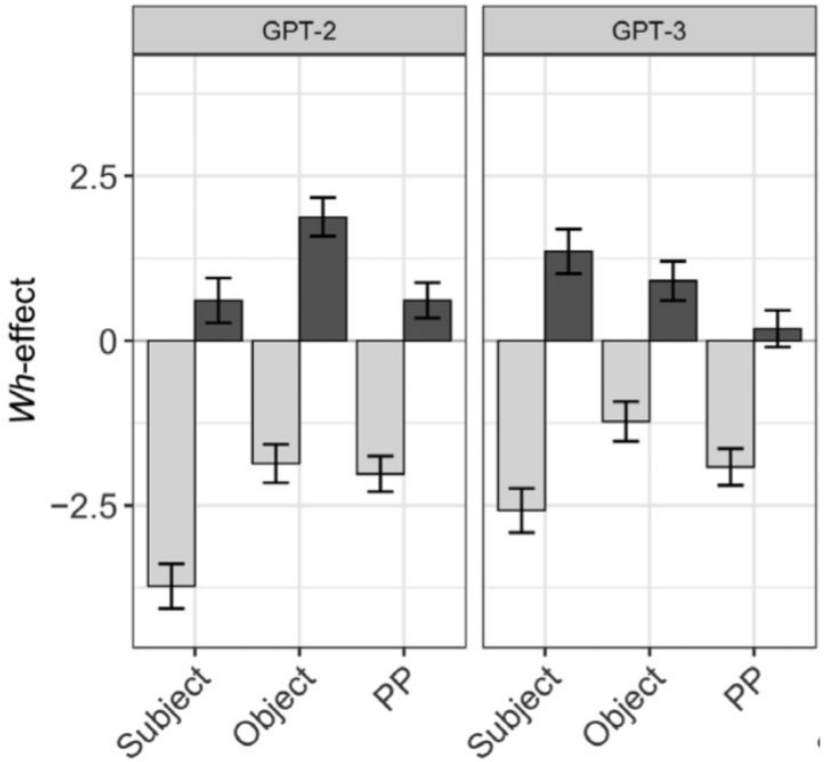



Figure 2: GPT-2 and GPT-3 show sensitivity to island conditions. *Figure from Wilcox, Futrell & Levy (2024) "Using Syntactic Models to Test Syntactic Learnability"*, available from: `https://www.colinphillips.net/wp-content/uploads/2024/05/wilcox2023.pdf`

# Evaluating Syntax

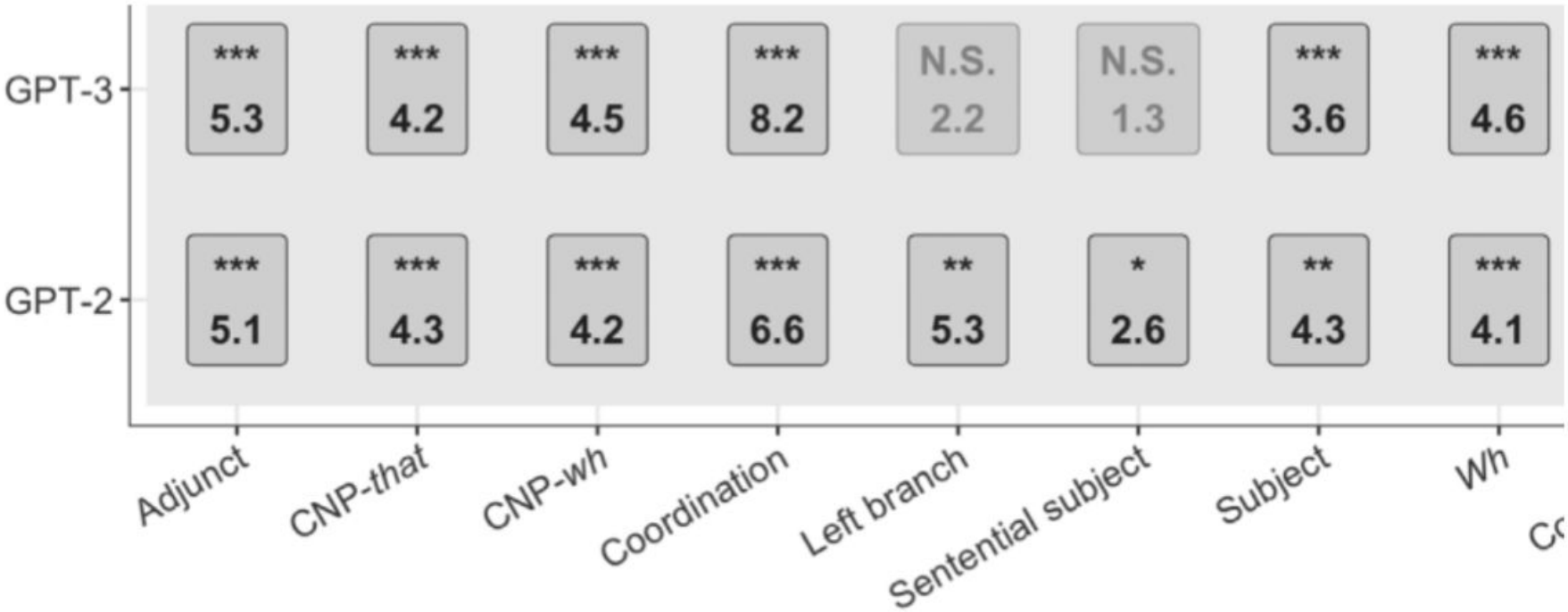



Figure 4: Summary of wh-effects across island test sets for GPT-2 and GPT-3

# Evaluating Syntax

| Category | Sentence |
|---|---|
| C-command | A lot of patients who can sell some couch didn't investigate **themselves/*itself**. |
| Principle A Case 1 | The teenagers explain that **they/*themselves** aren't breaking all glasses. |
| Principle A Case 2 | Eric imagines himself **taking/*took** every rug. |
| Domain 1 | Carla had explained that Samuel has discussed **her/*herself**. |
| Domain 2 | Donald can imagine those college campuses are boring **themselves/*himself**. |
| Domain 3 | Steven explains Kayla won't hurt herself v Kayla explains Steven won't hurt herself. |
| Reconstruction | It's himself that **this cashier attacked/*attacked this cashier**. |

# Limitations of BLiMP

BLiMP does not offer full coverage of ellipsis, since it only considers sentences of equal length. The ellipsis paradigms cover special cases of NP ellipsis (or more, precisely, in X-bar terms **N-bar Ellipsis**) that meet this practical constraint:
*Brad passed one big museum and Eva passed several.* v * *Brad passed one museum and Eva passed several big.*

It is worth mentioning that English has several forms of **predicate/VP ellipsis (VPE)**:

- Auxiliary VPE: Susan has read War and Peace, but Maria hasn't.
- Modifier VPE: Susan can speak French, and Maria can too.
- Pseudogapping: Susan doesn't eat pasta, but she does pizza.
- Antecedent Contained Deletion: Susan has read every book Maria has.

Typologically, we can note that many Romance and Germanic languages lack Auxiliary VPE, although they do have Auxiliary VPE, and pseudogapping is also more marginal here. Syntacticians typically attribute these differences to the nature of the English auxiliary system.

# A Real Example: What does BLiMP tell us?

| Phenomena | Untied | | Tied | |
|---|---|---|---|---|
| | CE Loss | Z Loss | CE Loss | Z Loss |
| anaphor agreement | 0.922 | 0.9185 | 0.915 | **0.96** |
| argument structure | 0.6821 | **0.7629** | 0.7492 | 0.7427 |
| binding | 0.7110 | 0.7751 | 0.764 | **0.7913** |
| control raising | 0.7250 | 0.7656 | **0.7822** | 0.7326 |
| determiner noun agreement | 0.8194 | 0.8785 | **0.8871** | 0.8804 |
| ellipsis | 0.6605 | 0.7515 | **0.7785** | 0.73 |
| filler gap dependency | 0.5183 | 0.5417 | 0.5187 | **0.5619** |
| irregular forms | 0.8835 | **0.9560** | 0.9510 | 0.9385 |
| island effects | 0.4928 | 0.4752 | 0.4579 | **0.5138** |
| npi licensing | **0.6949** | 0.6693 | 0.6550 | 0.6594 |
| quantifiers | 0.5963 | 0.6350 | 0.5973 | **0.6525** |
| subject verb agreement | 0.7567 | **0.8388** | 0.7850 | 0.7875 |
| Average | 0.7052 | **0.7473** | 0.7367 | 0.7459 |

**Table 4** Detailed BLiMP Accuracy Scores for 14M Model Series

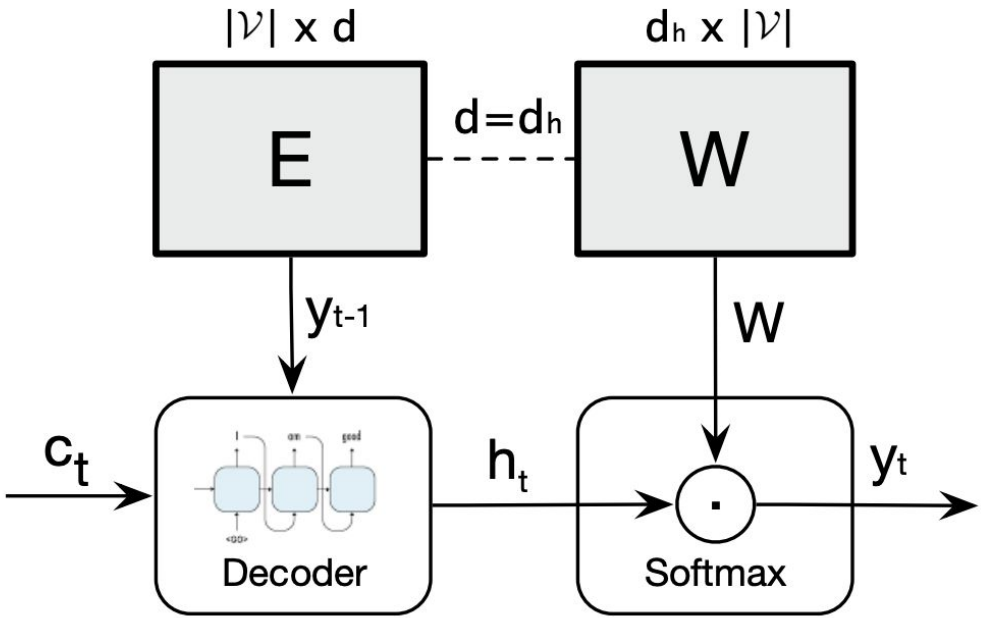
# A Real Example: What does BLiMP tell us?

| Phenomena | Untied | | Tied | |
|---|---|---|---|---|
| | CE Loss | Z Loss | CE Loss | Z Loss |
| anaphor agreement | 0.922 | 0.9185 | 0.915 | **0.96** |
| argument structure | 0.6821 | **0.7629** | 0.7492 | 0.7427 |
| binding | 0.7110 | 0.7751 | 0.764 | **0.7913** |
| control raising | 0.7250 | 0.7656 | **0.7822** | 0.7326 |
| determiner noun agreement | 0.8194 | 0.8785 | **0.8871** | 0.8804 |
| ellipsis | 0.6605 | 0.7515 | **0.7785** | 0.73 |
| filler gap dependency | 0.5183 | 0.5417 | 0.5187 | **0.5619** |
| irregular forms | 0.8835 | **0.9560** | 0.9510 | 0.9385 |
| island effects | 0.4928 | 0.4752 | 0.4579 | **0.5138** |
| npi licensing | **0.6949** | 0.6693 | 0.6550 | 0.6594 |
| quantifiers | 0.5963 | 0.6350 | 0.5973 | **0.6525** |
| subject verb agreement | 0.7567 | **0.8388** | 0.7850 | 0.7875 |
| Average | 0.7052 | **0.7473** | 0.7367 | 0.7459 |

**Table 4** Detailed BLiMP Accuracy Scores for 14M Model Series

**Cambridge-KAIST Collaboration**
*with Dr James Thorne's group*

# Example: Weight Tying




**Standard Output Layer (softmax linear unit) with or without weight tying.**

$$W = E^T$$

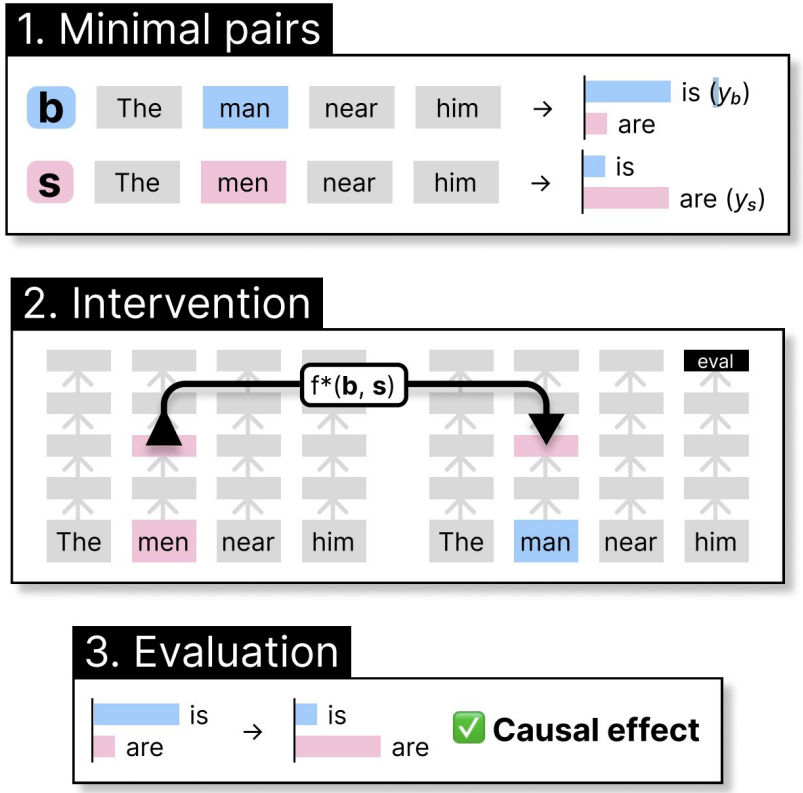
**Figure from Pappas et al (2018)**
https://aclanthology.org/W18-6308.pdf

# A Real Example: What does BLiMP tell us?

| Phenomena | Untied | | Tied | |
|---|---|---|---|---|
| | CE Loss | Z Loss | CE Loss | Z Loss |
| anaphor agreement | 0.922 | 0.9185 | 0.915 | **0.96** |
| argument structure | 0.6821 | **0.7629** | 0.7492 | 0.7427 |
| binding | 0.7110 | 0.7751 | 0.764 | **0.7913** |
| control raising | 0.7250 | 0.7656 | **0.7822** | 0.7326 |
| determiner noun agreement | 0.8194 | 0.8785 | **0.8871** | 0.8804 |
| ellipsis | 0.6605 | 0.7515 | **0.7785** | 0.73 |
| filler gap dependency | 0.5183 | 0.5417 | 0.5187 | **0.5619** |
| irregular forms | 0.8835 | **0.9560** | 0.9510 | 0.9385 |
| island effects | 0.4928 | 0.4752 | 0.4579 | **0.5138** |
| npi licensing | **0.6949** | 0.6693 | 0.6550 | 0.6594 |
| quantifiers | 0.5963 | 0.6350 | 0.5973 | **0.6525** |
| subject verb agreement | 0.7567 | **0.8388** | 0.7850 | 0.7875 |
| Average | 0.7052 | **0.7473** | 0.7367 | 0.7459 |

**Table 4** Detailed BLiMP Accuracy Scores for 14M Model Series

# We need *mechanisms*, not scores



# We need *mechanisms*, not scores

PICO

https://www.picolm.io/

**A Lightweight Framework for Studying Learning Dynamics**

Work in collaboration with Richard Diehl Martinez (Buttery Group).
With support of “Accelerate Programme” (Ryan Daniels)

# We need *mechanisms*, not scores

PICO https://www.picolm.io/

**A Lightweight Framework for Studying Learning Dynamics**

## PICO-Analyze

**Model Components**

- Weight matrices
- Activation values
- Gradient tensors

**Compound Components**

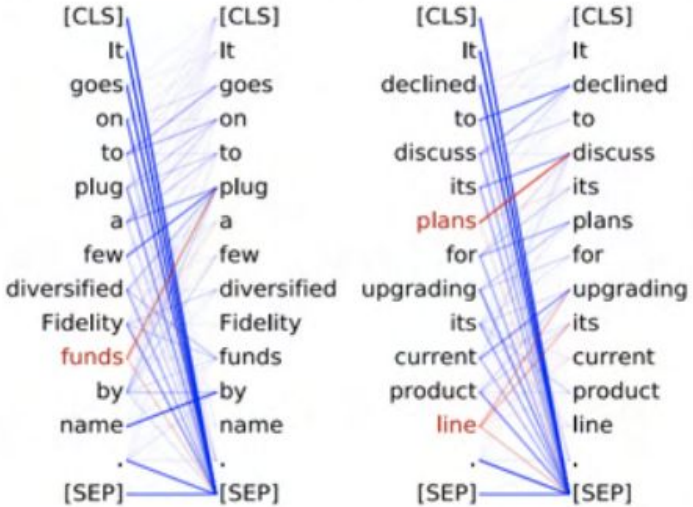
- OV-Circuits (combining value and output projections)
- Induction heads
- Attention heads
- Feed-forward blocks

Work in collaboration with Richard Diehl Martinez (Buttery Group).

# Side Note: *Mechanistic?*

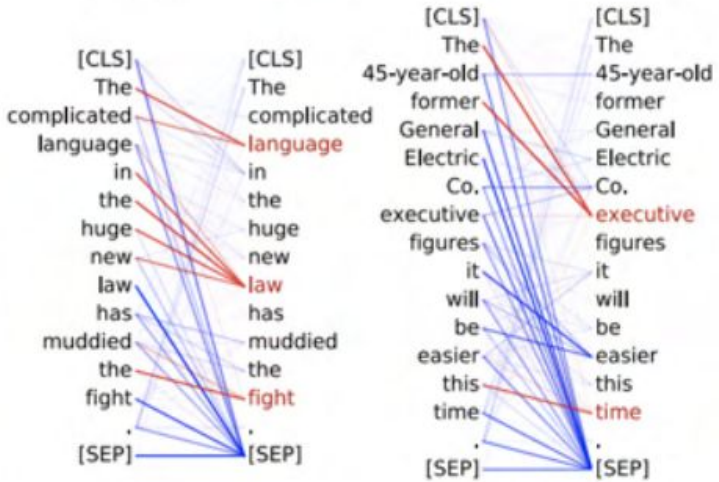
**Head 8-10**

- **Direct objects** attend to their verbs
- 86.8% accuracy at the **dobj** relation

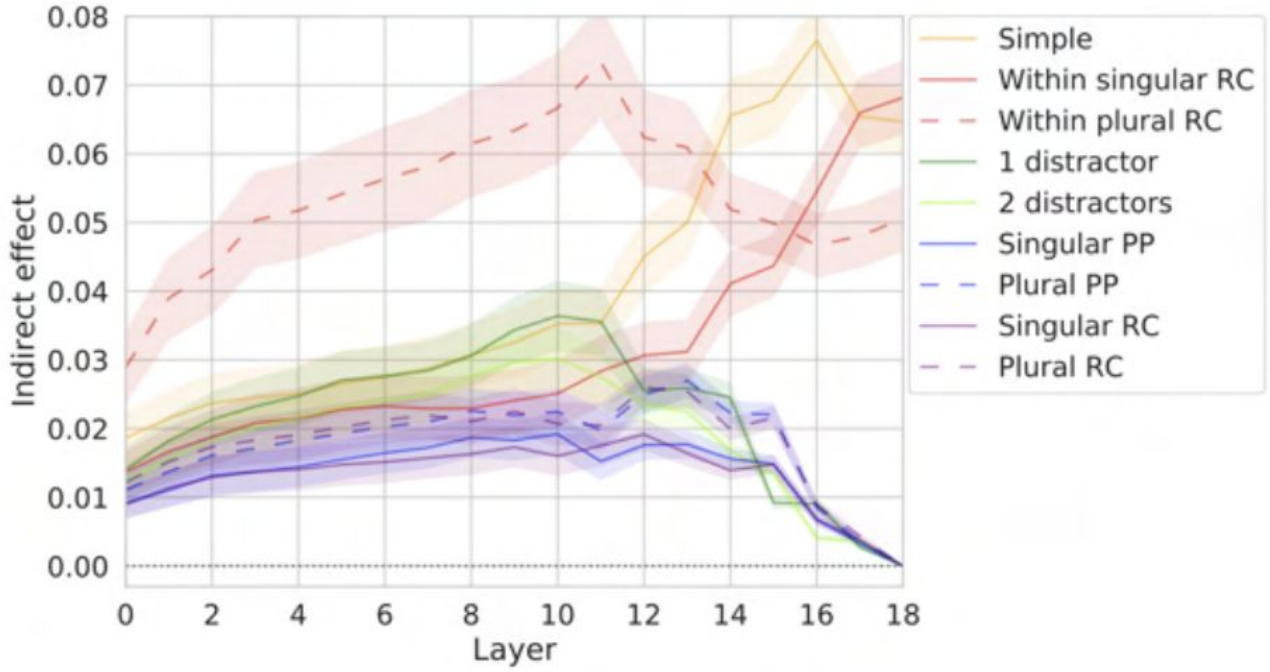



**Head 8-11**

- **Noun modifiers** (e.g., determiners) attend to their noun
- 94.3% accuracy at the **det** relation



# Sample Efficient Models rely on “Good Generalisation”

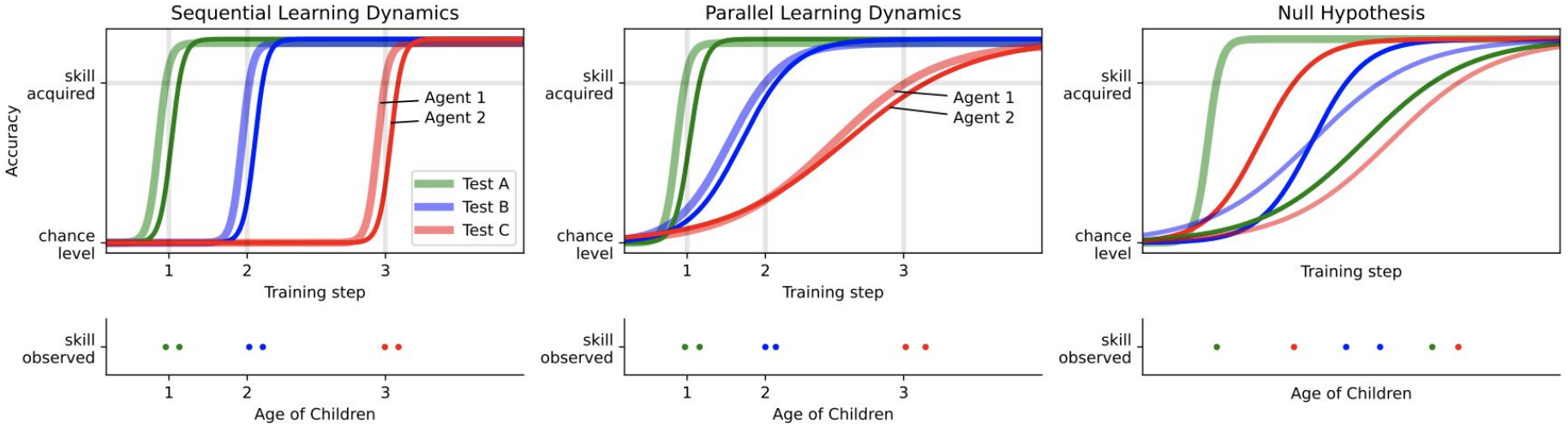



**Transformer-XL: dissociation between local (red/yellow) and non-local (purple/green) agreement.**

# Grammar Profiling: Enter Evanson et al (2023)

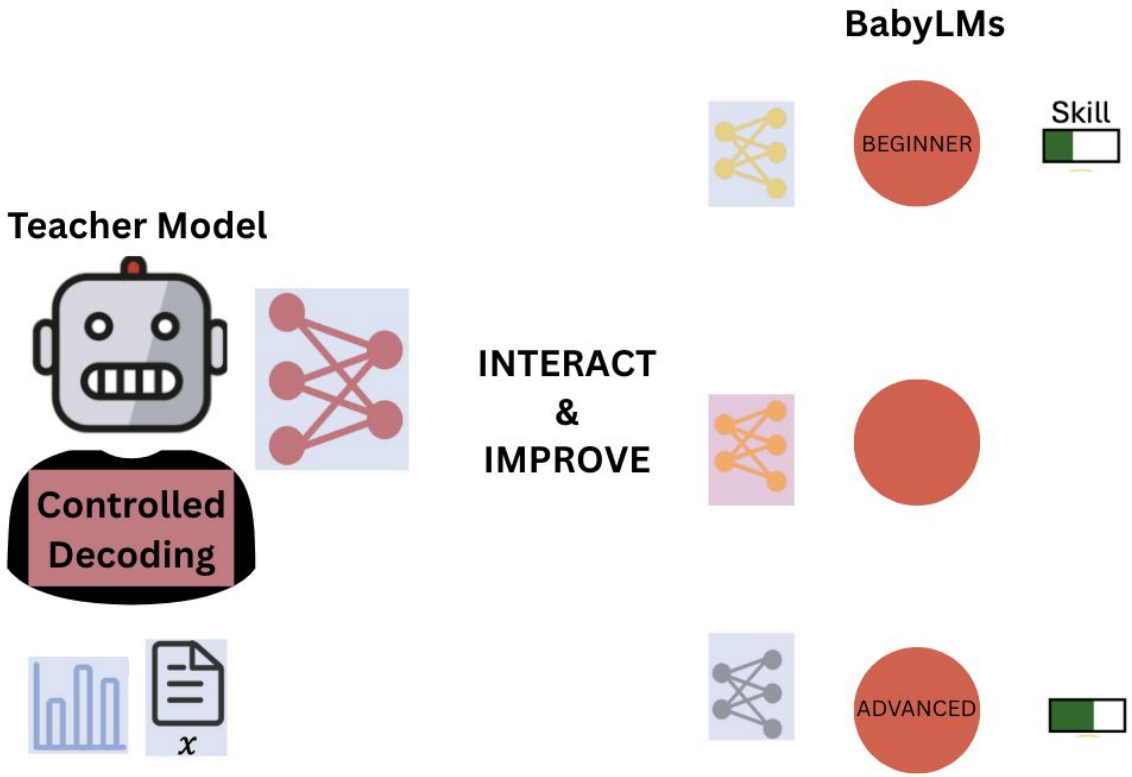
| Stage | Children | Language Model |
|---|---|---|
| 1 | Simple sentences in Subject-Verb (SV) order | SV agreement across simple sentences |
| 2 | Wh-questions | SV agreement in questions |
| 3 | Relative Clauses (RCs) | SV agreement across object RCs |

# Grammar Profiling: Enter Evanson et al (2023)



# An Interactive & Adaptive Language Model Playground

**Small Language Teacher Model with Controlled Decoding "interacts" with Cognitive Proxies (Student Models) with fine-grained rewards**




**Suchir Salhan**
sas245@cam.ac.uk

# Another (Non-Pedagogical) Application of the EGP

- Minimal pairs were **artificially generated** using from abstract grammars that exemplify syntactic phenomena – this easily yields a large number of sentences, which can help control for other possible sources of noise in test materials. Generation scripts use templates to sample lexical items with **selectional restrictions**, which annotate the morphological, syntactic, and semantic features of over 3000 items.

- **Human Evaluation:** Human benchmarking is important in several NLP tasks. It is a useful proxy for the difficulty of different tasks. For BLiMP, the authors used 20 validators who rated five pairs from each of the 67 paradigms for 6,700 judgments.

EGP: Naturally-occurring sentences

EGP: human-validated and graded (development; stages not ages)

# *Human-Validated* Grammar Profiles (Salhan et al *ongoing*)

*Figure from Nuria Bosch-Masip*

# *Human-Validated* Grammar Profiles (Salhan et al *ongoing*)

**Pronouns: possessives, reflexives, reciprocals**

**Determiners: Demonstratives, Possessives & Quantities**

**Conjunction: Coordinating, Subordinating**

**Comparatives, Imperatives, Exclamatives**

**TAM**

FORM:USE Distinction, Meta-Data

# Measuring Grammaticality

AP Comparative

I think that it is awful, because that means that in Spain it will be **even hotter/*even hot** than it is now
(Adjectives, Comparatives, B1)

Negative Declarative

I know you **couldn't come to my party** so I want to tell you about my presents and party
*** you comen't/ you not could come/ you could come not**
(A2 Waystage, 2004, Turkish, PASS)

# Measuring Grammaticality

A **timid, shy, self-conscious, over-sensitive and vulnerable** person can yearn to make friends with someone who is **very self-assured, confident, decisive, even bossy**

FORM: COMBINING MULTIPLE ADJECTIVES
Poland, C2 Mastery

I kept silent when I was introduced to **that new girlfriend of his**.

FORM/USE: WITH 'THAT ... OF'
Poland, C2 Mastery

# EGP + Syntactic Challenge Sets

Perhaps, we need to maintain some notion of descriptive linguistic coverage rather than **prescriptive targets** in grammar profiling?

→ measures of fluency, diversity and style may be important.

In a BabyLM context, we may care more about comparisons with L1 and L2 comparison and *varieties of English*.

# EGP + Syntactic Challenge Sets

**An Example from Sluicing:**
*Ellipsis that occurs in direct and indirect interrogatives introduced by [+wh]-expression.*

**Finiteness Mismatch**
The baseball player went public with his desire to be traded. He doesn't care where (he will be traded).

**Tense Mismatch**
Your favorite plant is alive, but you can never be sure how long (it will be alive).

**Modality Mismatch**
Sally knows that there is always the potential for awful things to happen, but she doesn't know when (awful things might happen).

**Polarity Mismatch**
Either the Board grants the license by December 15 or it explains why (the Board didn't grant the license by December 15).

# BLiMP causes a “typological bottleneck”

| Name | Size (k) | N | Language |
|---|---|---|---|
| BLiMP (Warstadt et al., 2020) | 67 | 67 | English |
| CLiMP (Xiang et al., 2021) | 16 | 16 | Chinese |
| SLING (Song et al., 2022) | 38 | 38 | Chinese |
| ZhoBLiMP (Liu et al., 2024) | 35 | 118 | Chinese |
| BLiMP-NL (Suijkerbuijk et al., 2024) | 8.4 | 22 | Dutch |
| JBLiMP (Someya and Oseki, 2023) | 0.33 | 39 | Japanese |
| RuBLiMP (Taktasheva et al., 2024) | 45 | 45 | Russian |
| NoCoLa (Jentoft and Samuel, 2023) | 99.1 | 11 | Norwegian |
| DaLAJ (Volodina et al., 2021) | 4.8 | 4 | Swedish |
| LINDSEA (Leong et al., 2023) | 0.38 | 38 | Indonesian |
| | 0.2 | 20 | Tamil |
| CLAMS (Mueller et al., 2020) | 331.5 | 7 | 5 Languages* |
| COMPS (Misra et al., 2023) | 49.3 | 4 | English |

Table 1: Summary of existing minimal pair datasets. Benchmarks in red represent *grammatical* tasks while benchmarks in blue denote *conceptual* minimal pairs. Size: # of minimal pairs in total, N: # of linguistic paradigms. *: English, French, German, Hebrew, Russian.

# BLiMP causes a “typological bottleneck”

| | Model | | English | Japanese | Chinese | French | German |
|---|---|---|---|---|---|---|---|
| Non-CL | SSLM (WIKI) | | 64.60% | 55.42% | 48.01% | 70.68% | 59.63% |
| | MAO-BABYBERTA | | 75.48% * | 61.21% | 51.32% | **80.00%** | 68.78% |
| CL | GROWING | | 71.13% | 79.30% | 56.22% | 76.21% | 71.13% |
| | INWARDS | | 71.05% | 81.32% | 54.26% | 79.01% | 69.34% |
| | MMM | (UPOS) | 74.22% | **87.31%** | **58.79%** | 75.93% | **73.25%** |
| | | (SEM) | **77.35%** | | 55.01% | | |

**Table 3** Evaluation of MAO-BABYBERTA (“vanilla” SSLM architecture without objective curricula) and the three Objective Curricula (GROWING, INWARDS, and MMM) on the following syntactic minimal pairs datasets: BLIMP (English), JBLIMP (Japanese), SLING (Chinese), CLAMS (French and German). Performance is compared to SSLM (WIKI). This is the same architecture trained on non-CDS training data. *This reports the performance of the best-performing “vanilla” model by Diehl Martinez et al. (2023) on the same architecture used to train our model. **Bolded** results indicate the highest accuracy of all the models.

# Grammar Profiling beyond English

**Complexity-graded, typologically-motivated evaluation benchmarks do not exist beyond English, but they should!**

**CLAMS**: Cross-Linguistic Syntactic Evaluation of Word Prediction Models

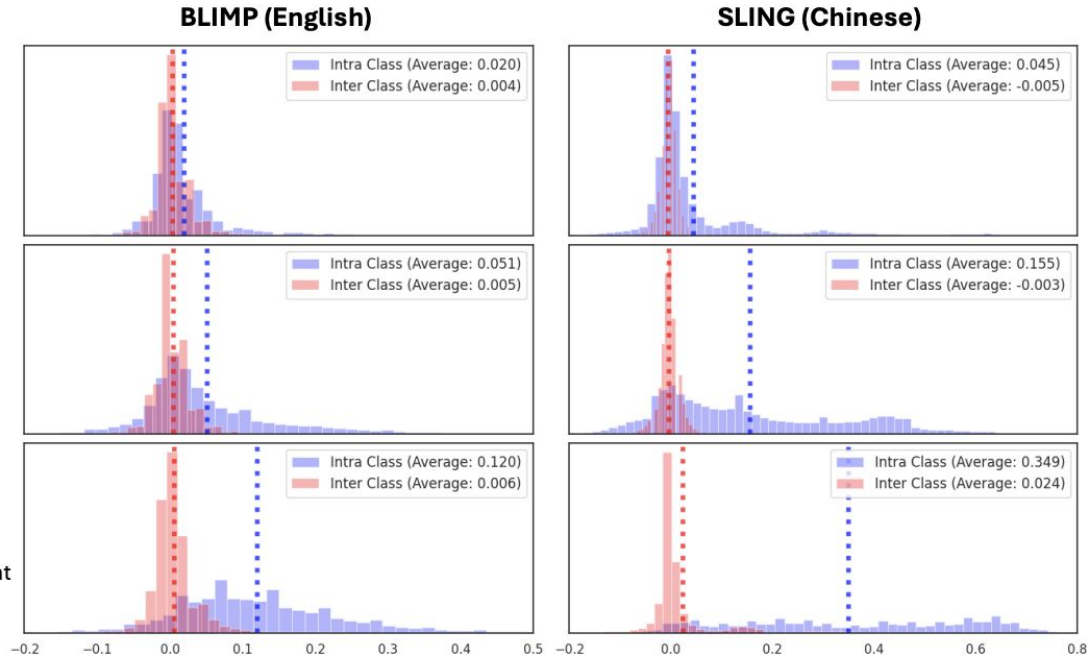
| | Model | | English | Japanese | Chinese | French | German |
|---|---|---|---|---|---|---|---|
| Non-CL | SSLM (WIKI) | | 64.60% | 55.42% | 48.01% | 70.68% | 59.63% |
| | MAO-BABYBERTA | | 75.48% * | 61.21% | 51.32% | **80.00%** | 68.78% |
| CL | GROWING | | 71.13% | 79.30% | 56.22% | 76.21% | 71.13% |
| | INWARDS | | 71.05% | 81.32% | 54.26% | 79.01% | 69.34% |
| | MMM | (UPOS) | 74.22% | **87.31%** | **58.79%** | 75.93% | **73.25%** |
| | | (SEM) | **77.35%** | | 55.01% | | |

# BLiMP causes a “typological bottleneck”

**3rd level Classification**
- Syntax
- Semantics
- Morphology
- Syntax-Semantics

**2nd level Classification**
- Island Effects
- NPI Licensing
- Anaphor Agreement
- Binding
- …

**1st level Classification**
- Adjunct Island
- Only NPI Licensor Present
- Anaphor Gender Agreement
- Principle A Case 1
- …




**3rd level Classification**
- Syntax
- Semantics

**2nd level Classification**
- Alternative Question
- Aspect
- Definiteness Effect
- Polarity Items
- …

**1st level Classification**
- 还是…吗 (haishi…ma)
- 在…过 (zai…guo)
- 定指_每(definiteness_every)
- 或多或少 (more or less)
- …

See Zhou et al (2025, COLING) for interesting discussion.
https://aclanthology.org/2025.coling-main.459.pdf

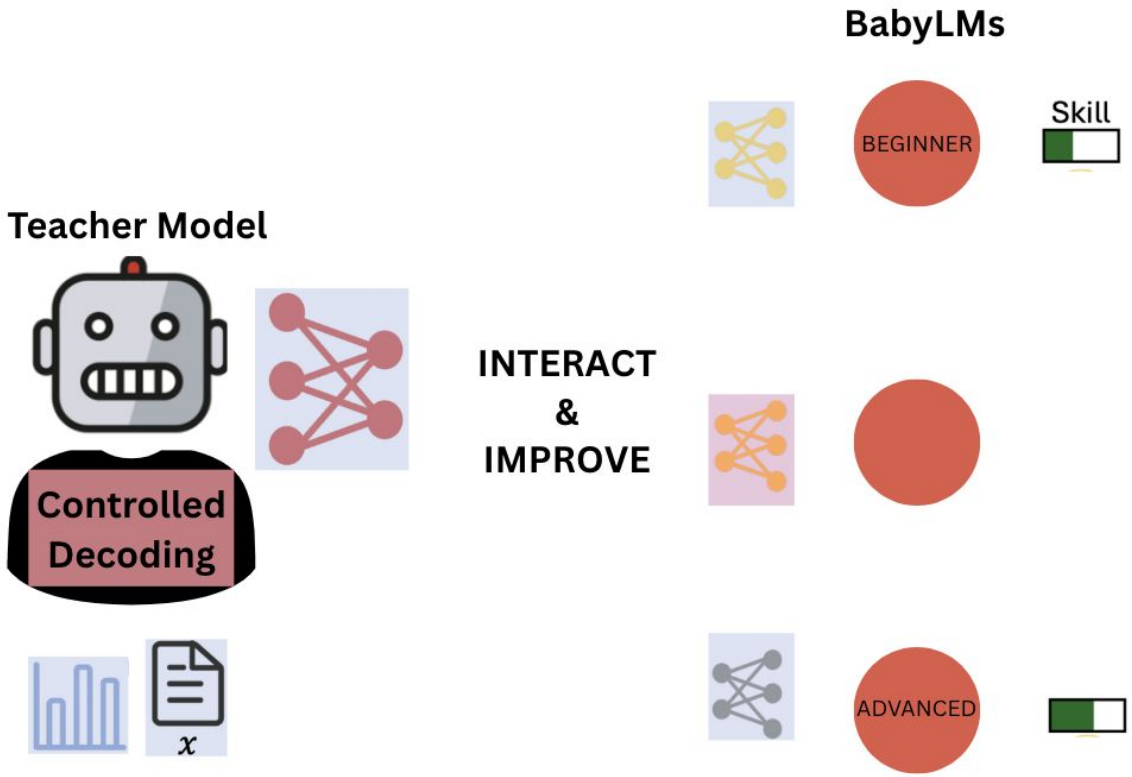
# BabyLMs & SLMs: Back to L2 Acquisition

BabyLMs and SLMs are novel AI methods that (attempt to) precisely leverage a rich literature from linguistics, cognitive science

BUT the big question: how does it help with SLA analysis and effective didactics in real-life education?

# An Interactive & Adaptive Language Model Playground

**Small Language Teacher Model with Controlled Decoding "interacts" with Cognitive Proxies (Student Models) with fine-grained rewards**




**Suchir Salhan**
sas245@cam.ac.uk