# From Exposure to Productivity: Investigating the Mechanisms of Syntactic Rule Learning

Daoxin Li, Department of Linguistics, UNC-Chapel Hill

BUCLD, November 8, 2026

Language Learning & Computation Lab



# What does it mean to acquire a language?



# How do learners turn linguistic experience into productive rules?

# Roadmap

Three lines of work that address complementary aspects of the same scientific problem:

- Computational modeling: What rules can emerge from realistic input?
- Artificial language learning experiments: How do human learners make use of input data to acquire rules?
- Human-LLM comparisons: What can the similarities and differences tell us about  rule learning mechanisms?

# Today

- Syntax: today's focus, but not limited to it
- Distributional learning: Learning by tracking and utilizing distributional cues in the input (e.g., Maratsos & Chalkley, 1980; Braine, 1987; Saffran, Aslin & Newport, 1996).
  - a view of language acquisition that prioritizes the formal properties of grammatical structures (Chomsky, 1955).



# Computational modeling: Learning verb argument structure (Li, 2024, in prep)

# Verb argument structure

The specification of features of arguments required by a verb for the structure to be well-formed, e.g., the number and type of arguments.

- *She fell.* vs. **I fell her.*
- **She touched.* vs. *She touched me.*
- **I put the book.* vs. *I put the book on the desk.*

# Syntax-semantics mapping rules

Linguists agree that there are systematic syntax-semantics mapping rules (e.g., Gruber, 1965; Jackendoff, 1978; Ladusaw & Dowty, 1988) and there is extensive evidence that children know these rules from a young age (e.g., Pinker, 1989; Naigles, 1990; Yuan & Fisher, 2009; Arunachalam & Waxman, 2010).

- Causation – transitive (e.g., *open, eat*)
- Motion – PP (e.g., *put, walk*)
- Transfer – dative (e.g., *give, send*)

# The learnability problem

How does a child discover the mapping rules?

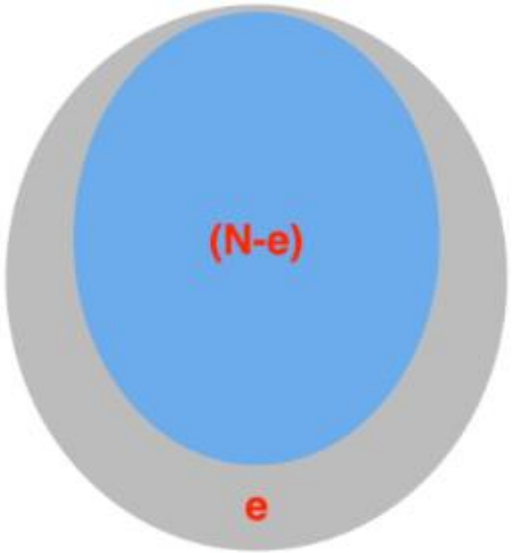
# The learnability problem

How does a child discover the mapping rules?

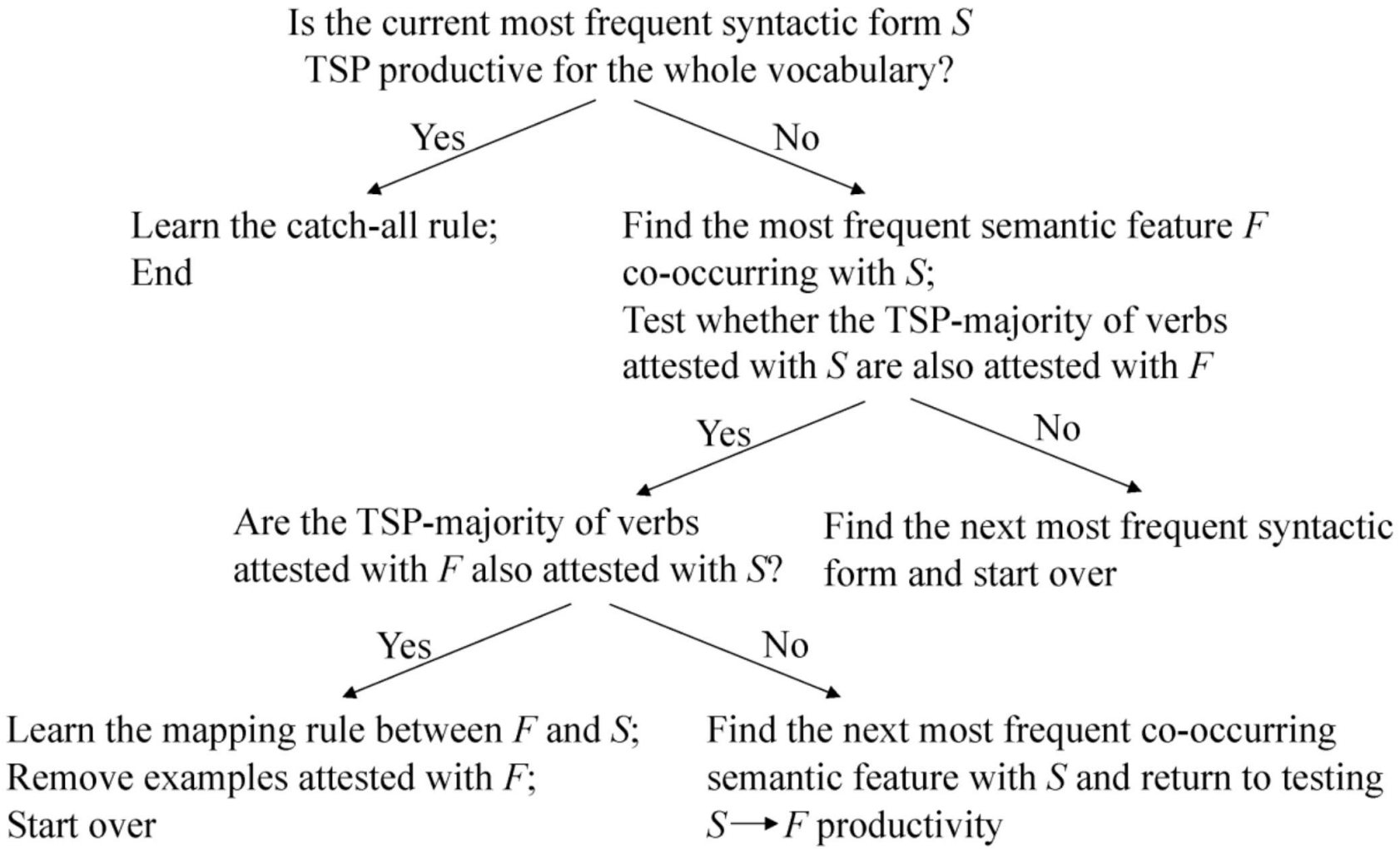
- Unlikely entirely universal or innate given the considerable variabilities across languages and idiosyncrasies within (e.g., Fisher, Gleitman & Gleitman, 1991; Bowerman & Brown, 2008; Bavin & Stoll, 2013).
  - English: many transfer verbs cannot take double objects (Levin, 1993)
  - Yucatec Mayan: motion paths not encoded by PP's (Bohnemeyer, 2008)
  - Chamorro: no double-object construction (Chung, 1998)

# A computational model for learning verb argument structure

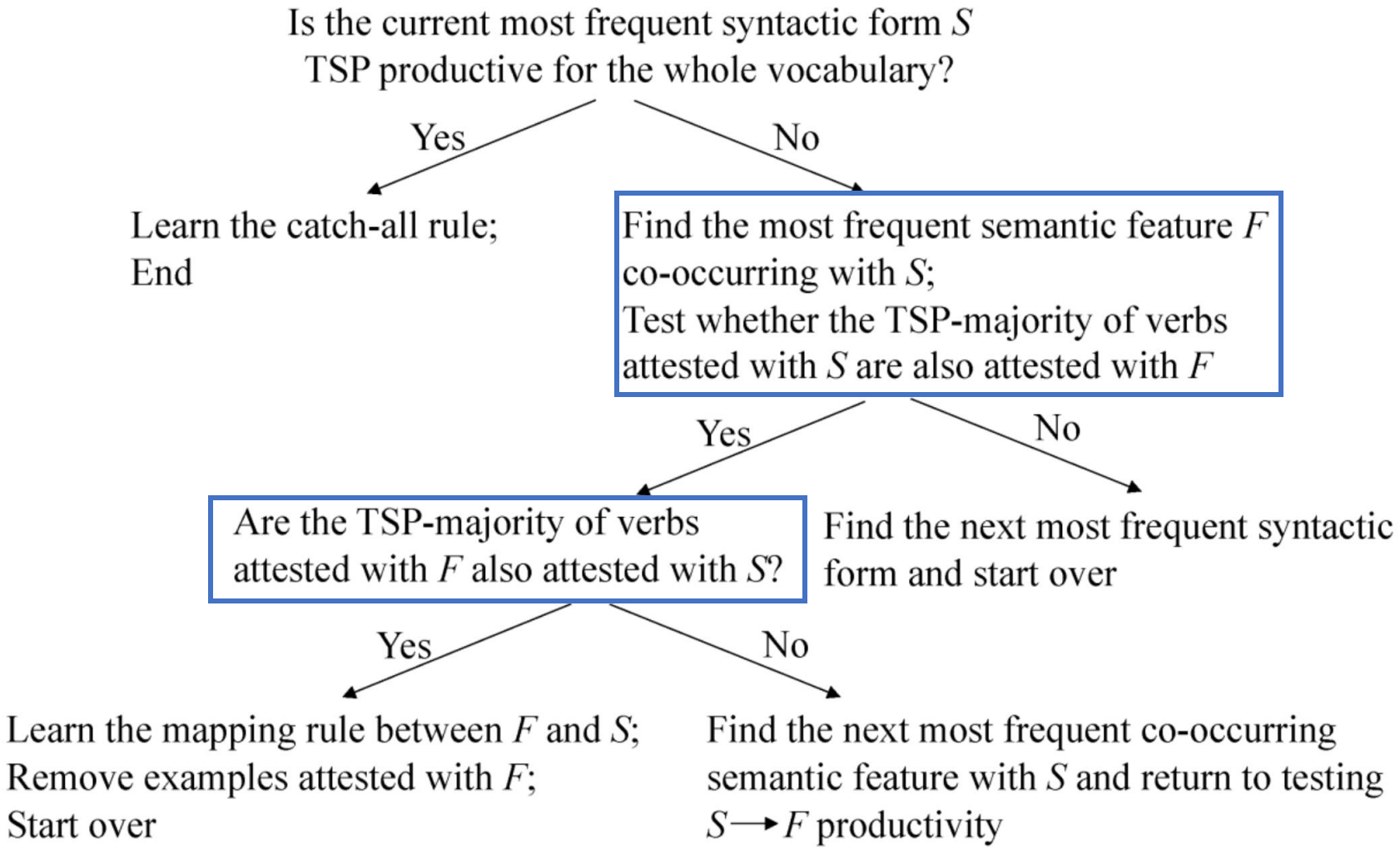
- Model input: (verb, syntactic frame, semantic features)
  - e.g., ("bite", "V NP", ("act", "causation"))
- Based on the Tolerance/Sufficiency Principle (Yang, 2016)
  - Let a generalization *R* be defined over a set of *N* items in the input. *R* is productive if and only if *M*, the number of items attested to follow *R*, exceeds *N - N/lnN*.



# Major procedures



# Major procedures

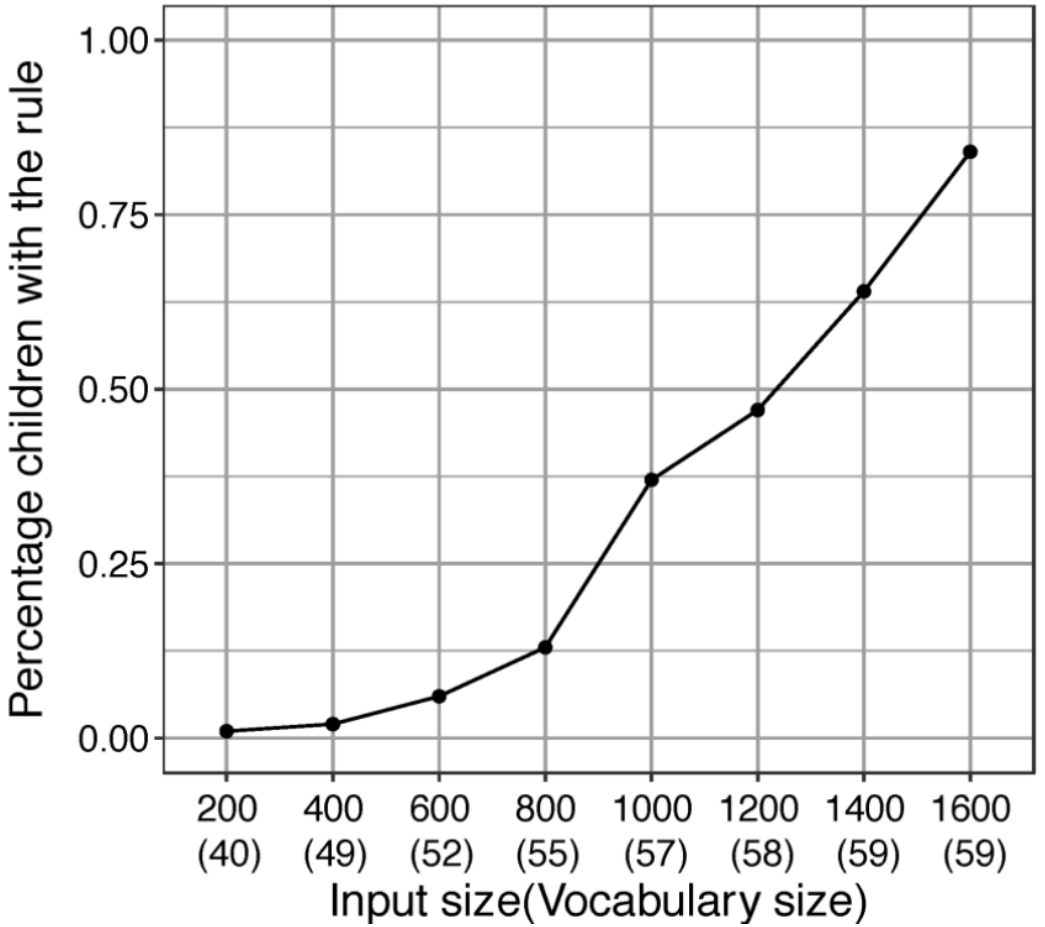


# Simulation

- Data: Alex's (1;4 to 3;5) input from Providence corpus (Demuth et al. 2006) containing 60 most frequent verbs in child English (Rowe & Goldin-Meadow, 2009; Carlson et al., 2014)
  - manually coded the verbs' syntactic frames and the semantic features observable in the accompanying video
- Mapping rules learned:
  - causation - transitive
  - act - intransitive
  - change of state - intransitive
  - creation - double object
  - motion - PP
  - transfer - dative

# Simulation

- The model can learn the well-documented causative overgeneralizations: intransitive & change of state → transitive (e.g., *Kendall fall that toy*.) (Bowerman, 1982)
- Model comparisons: The current model yields learning outcomes that are more accurate and more consistent with human behavior than a Bayesian model (Alishali & Stevenson, 2008).



# Study 1 takeaway

- Syntax-semantics mapping rules for verb argument structure are learnable from realistic language-specific input data without pre-specified form-meaning associations.
- Computational modeling force us to turn verbal theories of learning into explicit, testable hypotheses about the mechanisms of acquisition.



# But how do human learners actually form linguistic generalizations?



# Artificial language learning (ALL) experiments: Learning and revising a generalization (Li, under review)

# Questions

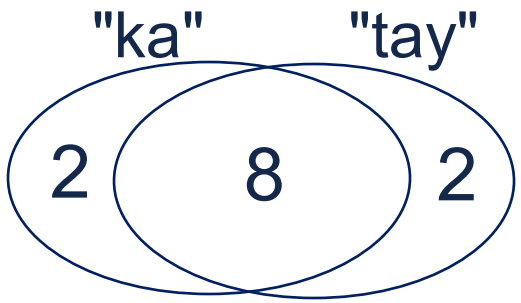
- What evidence leads to linguistic generalizations?
  - More frequent patterns are more likely to be generalized (e.g., Koulaguina & Shi, 2019; Schuler et al., 2016; Wonnacott et al., 2008; Jarosz et al., 2025).
  - The relative contributions of *type frequency* (number of unique forms) vs. *token frequency* (total occurrences) remain debated (e.g., Baayen, 1993; Bybee, 1995; Nosofsky, 1988; Rumelhart & McClelland, 1986; Yang, 2016).
- What if the relevant frequency distribution changes over time?
  - Learners continually encounter new evidence.
  - Common assumption: Learners update their grammar or linguistic representations as they receive new input (e.g., Chang et al., 2006; Gibson & Wexler, 1994; Perfors et al., 2011; Yang, 2003).
  - Relatively little direct experimental evidence.

# Experiment 1: Establishing the baseline

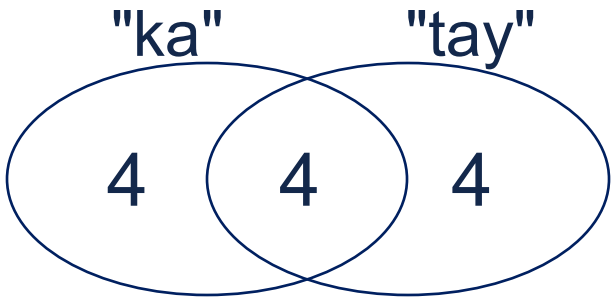
- An artificial language consisting of
  - 12 pseudowords (all conforming to English phonotactics, adapted from Wonnacott et al., 2008)
  - 2 markers ("ka" and "tay")
- Sentences: a marker + a word (e.g., "ka norg")

# Experiment 1: Establishing the baseline

- Productive condition:




- Unproductive condition:

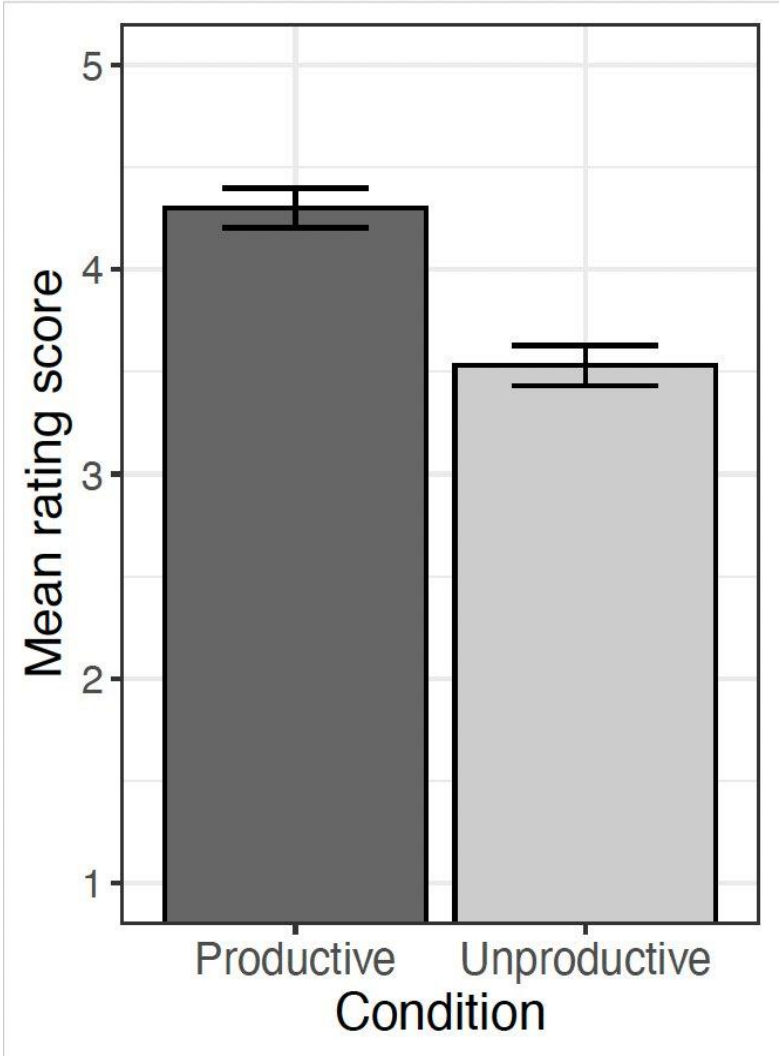



- Token frequency was uniform across words and balanced across markers.

# Experiment 1: Establishing the baseline

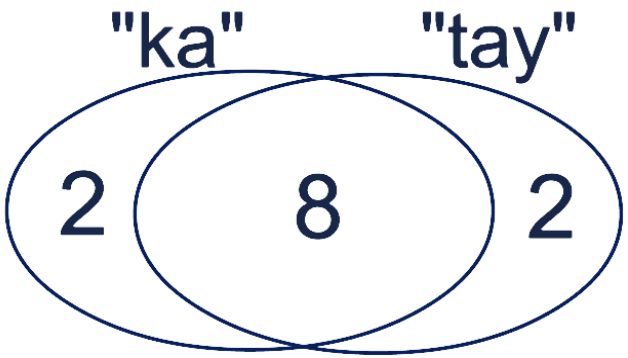
- Test: rating task on a scale 1-5
- Critical items: unattested combinations, i.e., "ka"-only words used with "tay", and "tay"-only words used with "ka"
  - Assessing participants' learning of the generalization that all words can be productively used with both markers

# Experiment 1: Establishing the baseline



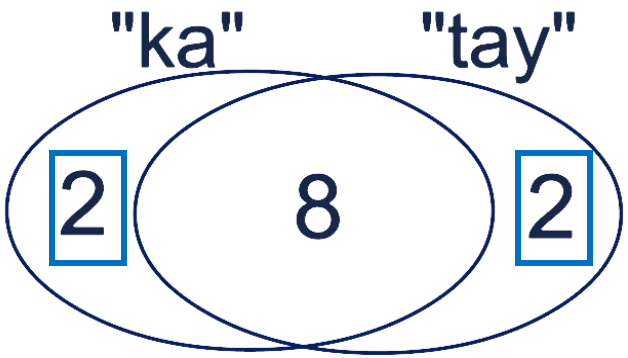
# Experiment 2: Type vs. token frequency

- Pitting type frequency against token frequency:
  - Same type frequency distribution as Experiment 1 Productive condition



# Experiment 2: Type vs. token frequency

- Pitting type frequency against token frequency:
  - Same type frequency distribution as Experiment 1 Productive condition



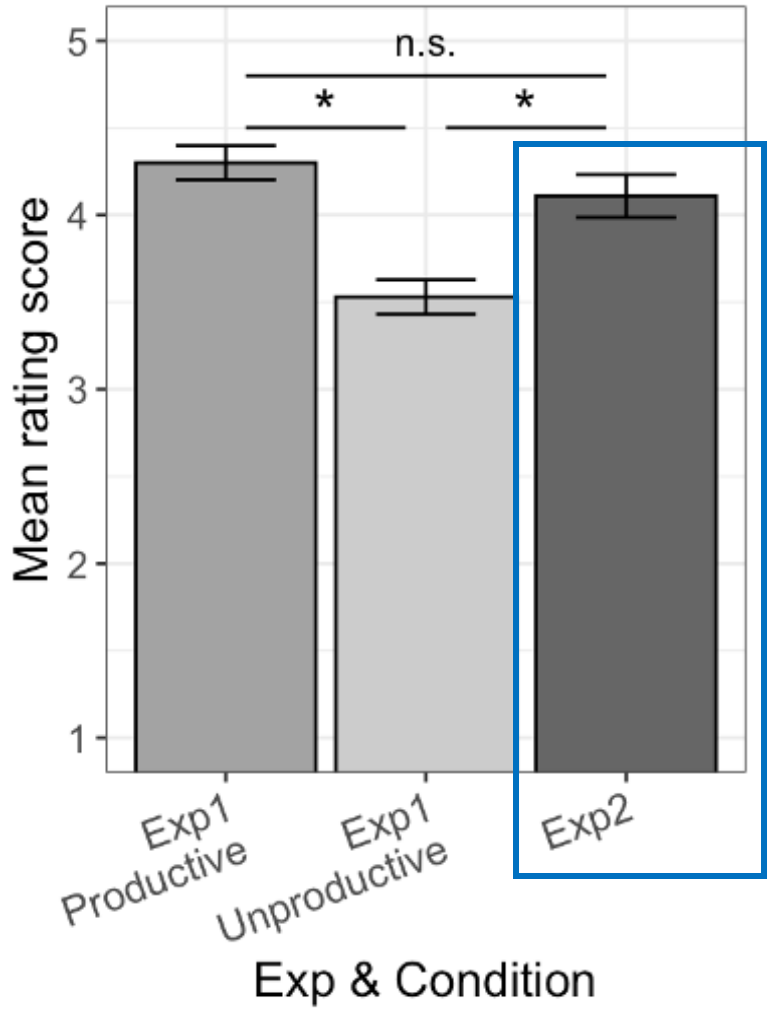

  - Same token frequency distribution as Experiment 1 Unproductive condition: Token frequency of non-alternating words was four times that of alternating words

# Experiment 2: Type vs. token frequency

- If type frequency is the primary determinant of productivity: Experiment 2 should pattern with Experiment 1 Productive condition.
- If token frequency is primary: Experiment 2 should pattern with Experiment 1 Unproductive condition.
- If both cues exert substantial independent effects: Performance should fall between the two.

# Experiment 2: Type vs. token frequency

- Type frequency is the primary driver of generalization in the current paradigm!
- This doesn't mean that token frequency is unimportant to language learning.



# Experiment 3: Can learners revise a generalization?



# Experiment 3: Can learners revise a generalization?

| | Alternating words | "ka"-only words | "tay"-only words | Proportion of alternating words |
|---|---|---|---|---|
| Stage 1 | 4 | 1 | 1 | 2/3 |
| Stage 2 | 0 | 3 | 3 | 0 |
| Overall | 4 | 4 | 4 | 1/3 |

# Experiment 3: Can learners revise a generalization?

| | Alternating words | "ka"-only words | "tay"-only words | Proportion of alternating words |
|---|---|---|---|---|
| Stage 1 | 4 | 1 | 1 | 2/3 |
| Stage 2 | 0 | 3 | 3 | 0 |
| Overall | 4 | 4 | 4 | 1/3 |

# Experiment 3: Can learners revise a generalization?

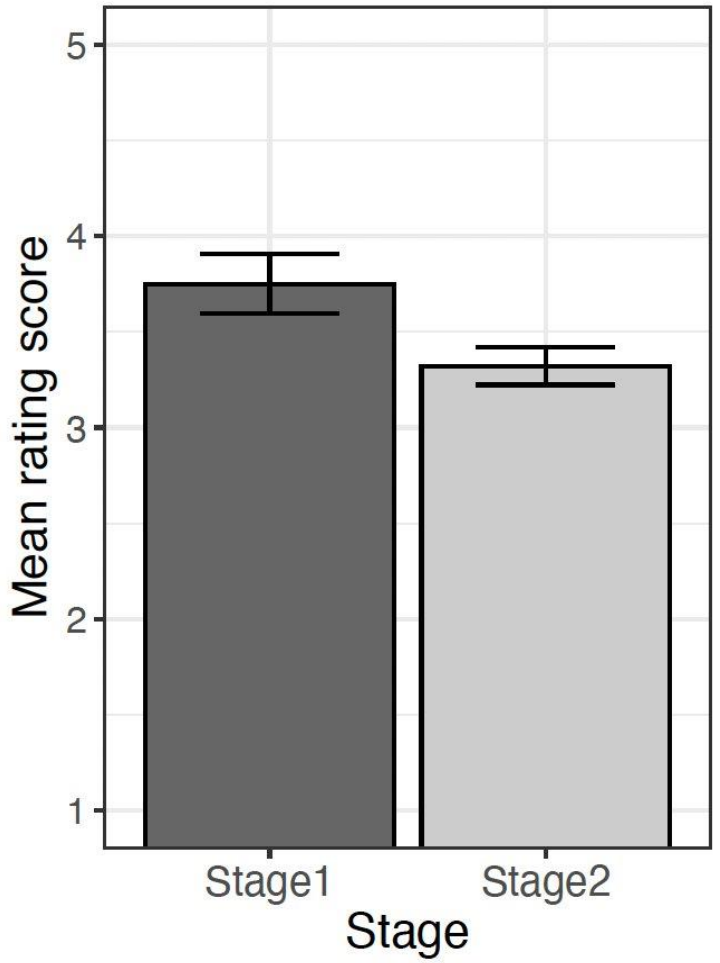
| | Alternating words | "ka"-only words | "tay"-only words | Proportion of alternating words |
|---|---|---|---|---|
| Stage 1 | 4 | 1 | 1 | 2/3 |
| Stage 2 | 0 | 3 | 3 | 0 |
| Overall | 4 | 4 | 4 | 1/3 |

# Experiment 3: Can learners revise a generalization?

| | Alternating words | "ka"-only words | "tay"-only words | Proportion of alternating words |
|---|---|---|---|---|
| Stage 1 | 4 | 1 | 1 | 2/3 |
| Stage 2 | 0 | 3 | 3 | 0 |
| Overall | 4 | 4 | 4 | 1/3 |

- Uniform token frequency
- If learners continue to revise their generalization as new evidence arrives, ratings should decrease from Stage 1 to Stage 2.

# Experiment 3: Can learners revise a generalization? - Yes!



# Experiment 3

- Were participants updating based on accumulated evidence or simply following the most recent input?

| | Alternating words | "ka"-only words | "tay"-only words | Proportion of alternating words |
|---|---|---|---|---|
| Stage 1 | 4 | 1 | 1 | 2/3 |
| Stage 2 | 0 | 3 | 3 | 0 |
| Overall | 4 | 4 | 4 | 1/3 |

# Experiment 3

- Were participants updating based on accumulated evidence or simply following the most recent input?

| | Alternating words | "ka"-only words | "tay"-only words | Proportion of alternating words |
|---|---|---|---|---|
| Stage 1 | 4 | 1 | 1 | 2/3 |
| Stage 2 | 0 | 3 | 3 | 0 |
| Overall | 4 | 4 | 4 | 1/3 |

# Experiment 3

- Were participants updating based on accumulated evidence or simply following the most recent input?

| | Alternating words | "ka"-only words | "tay"-only words | Proportion of alternating words |
|---|---|---|---|---|
| Stage 1 | 4 | 1 | 1 | 2/3 |
| Stage 2 | 0 | 3 | 3 | 0 |
| Overall | 4 | 4 | 4 | 1/3 |

# Experiment 4: Cumulative evidence or most recent input?

| | Alternating words | "ka"-only words | "tay"-only words | Proportion of alternating words |
|---|---|---|---|---|
| Stage 1 | 6 | 0 | 0 | 1 |
| Stage 2 | 2 | 2 | 2 | 1/3 |
| Overall | 8 | 2 | 2 | 2/3 |

# Experiment 4: Cumulative evidence or most recent input?

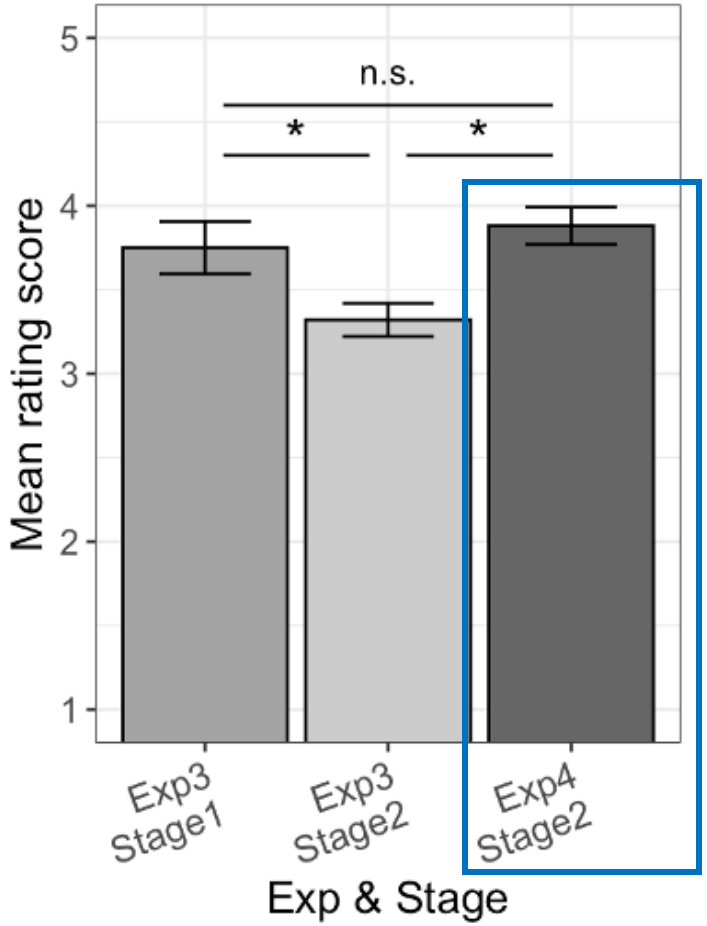
| | Alternating words | "ka"-only words | "tay"-only words | Proportion of alternating words |
|---|---|---|---|---|
| Stage 1 | 6 | 0 | 0 | 1 |
| Stage 2 | 2 | 2 | 2 | 1/3 |
| Overall | 8 | 2 | 2 | 2/3 |

# Experiment 4: Cumulative evidence or most recent input?

| | Alternating words | "ka"-only words | "tay"-only words | Proportion of alternating words |
|---|---|---|---|---|
| Stage 1 | 6 | 0 | 0 | 1 |
| Stage 2 | 2 | 2 | 2 | 1/3 |
| Overall | 8 | 2 | 2 | 2/3 |

- If participants relied primarily on the most recent input, Experiment 4 Stage 2 ratings should resemble Experiment 3 Stage 2.
- If they integrated evidence across stages, their ratings should instead remain high.

# Experiment 4: Cumulative evidence or most recent input? - Cumulative evidence!



# Study 2 takeaway

- Productivity is supported by type frequency.
- Generalizations are dynamically updated from cumulative evidence.
- Generalization formation and revision may reflect a common computation.
- Developing mechanistic accounts of language acquisition requires integrating approaches that combine *computational precision*, *interpretability*, and *empirical plausibility*.



# LLMs vs. humans in learning generalizations

## (Li & Sun, in prep)

# LLMs vs. humans

- Recent advances in Large Language Models (LLMs) have intensified debates:
  - Can these models acquire and represent knowledge in ways comparable to humans?
  - Can the study of artificial intelligence and human intelligence inform each other?

(e.g., Binz & Schulz, 2023; Goldstein et al., 2022; Linzen & Baroni, 2021; Ma & Xu, 2025; Warstadt & Bowman, 2022; Yasunaga et al., 2024)

# Putting LLMs and humans to the same test

We adapt published ALL experiments to examine:

- Can LLMs can acquire abstract linguistic generalizations from controlled distributional input?
- Do they show the same patterns of generalization as human learners?

# Setup



- Models:
  - GPT-6 Astra
  - GPT-4o
  - o4-mini
- Learning paradigm: API-based in-context learning with artificial language stimuli
- Matched the original experiments' sample size
- Parameters: *temperature=0, top_p=1*, or *reasoning effort=low*

# Case study: syntactic categories

- Words are organized in syntactic categories (e.g., nouns, verbs, adjectives).
- How can learners discover which words belong to which category?

(e.g., Mintz, Newport & Bever, 2002; Mintz, 2003; LouAnn Gerken, Wilson, & Lewis, 2005; Reeder, Newport & Aslin, 2013, 2017; Schuler et al., 2017; Liang, Marsala & Yang, 2022)

# Reeder, Newport & Aslin (2013)

- Hypothesis: Shared distributional contexts provide evidence for shared category membership.
  - "I remembered to **eat** yesterday"
  - "We saw the cat **eat** earlier"
  - ...
  - "They remember to **nerk** yesterday"
  - ✅ "saw it **nerk** earlier"

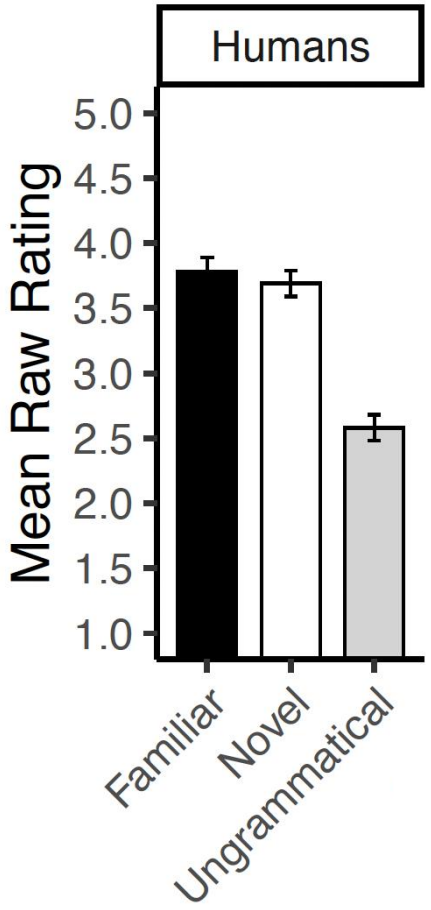
# Reeder, Newport & Aslin (2013)

- Training: Exposure to a subset of sentences generated by the artificial grammar (Q)A**X**B(R)
- Test: Rate three types of AXB sentences on a scale of 1 – 5
  - Familiar (grammatical + previously heard)
  - Novel (grammatical + never heard)
    - Critical: Can learners generalize beyond exposure?
  - Ungrammatical (violates the grammar, e.g., AXA, BXB)

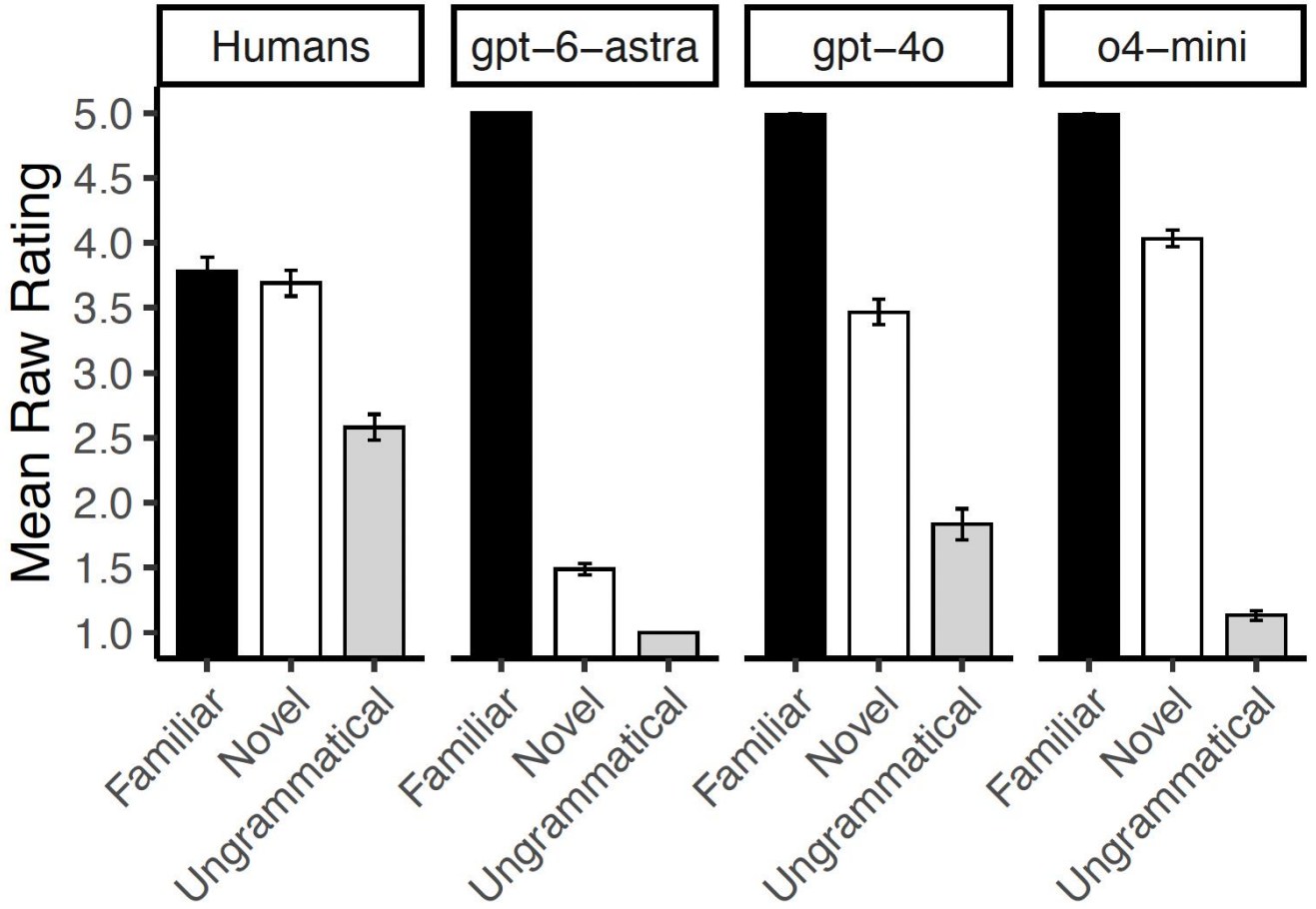
# Reeder, Newport & Aslin (2013)

- Experiment 1:
  - complete overlap (every X occurs with every A and every B)
  - dense (2/3 of all AXB combinations heard during training)
- Experiment 2:
  - complete overlap
  - sparse (1/3 of all AXB combinations heard during training)

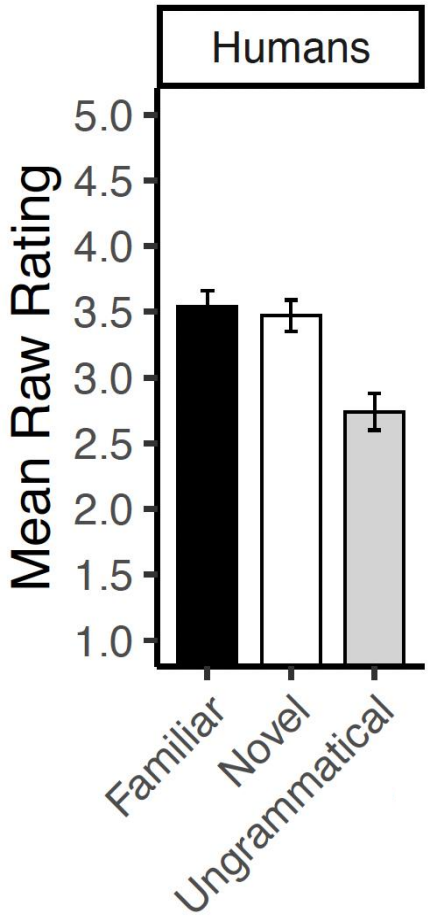
# Humans vs. models: complete & dense



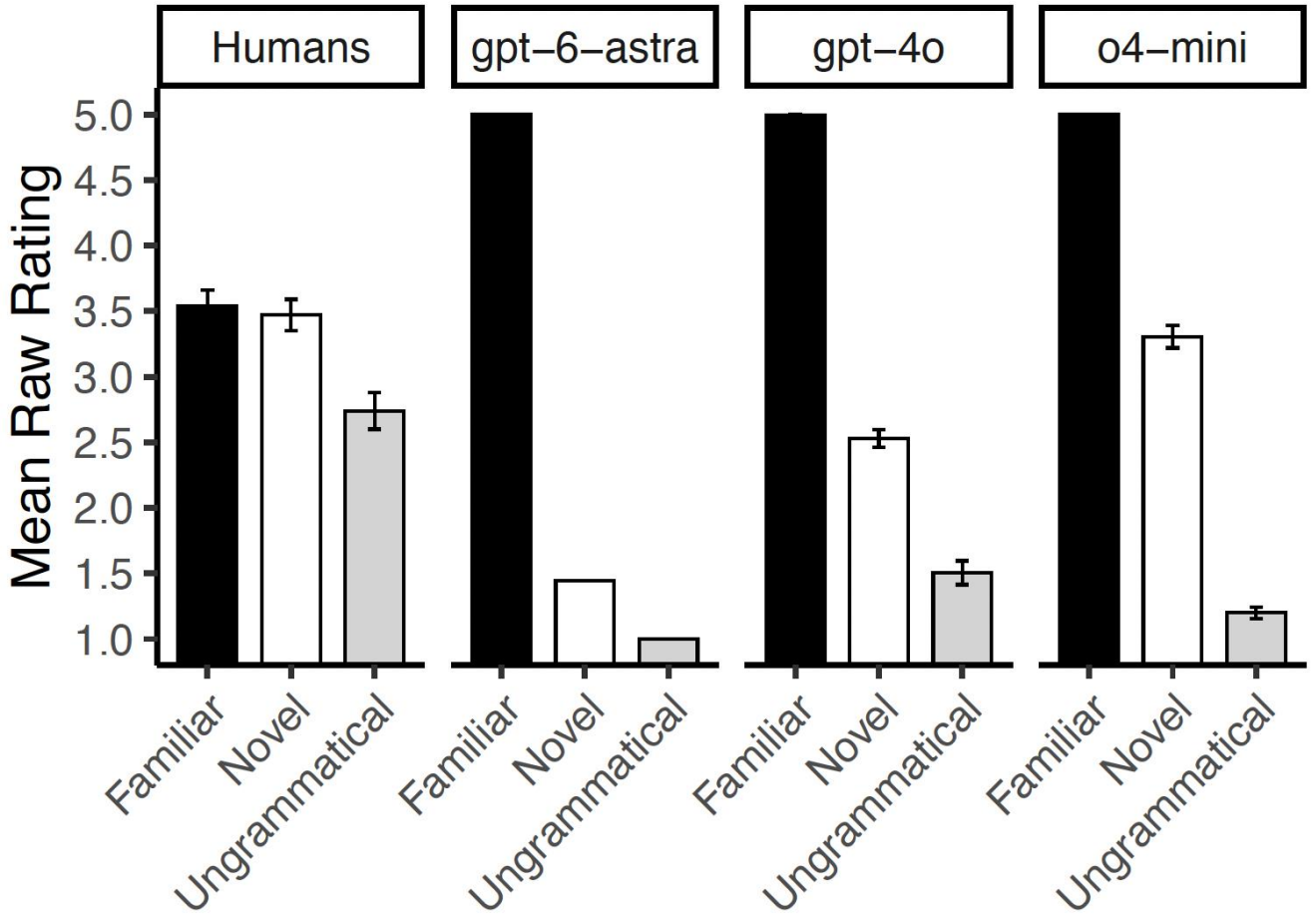
# Humans vs. models: complete & dense



# Humans vs. models: complete & sparse



# Humans vs. models: complete & sparse



# Study 3 takeaway

- LLMs can successfully extract distributional regularities from input.
- But they often struggle with generalizing to novel sentences.
  - may rely heavily on local similarity to exposure sentences rather than abstracting productive rules as human learners.
- Empirical paradigms from language acquisition can provide a powerful framework for directly comparing how humans and artificial systems learn and generalize.



# Conclusion

# Toward a mechanistic account of syntactic rule learning

- Take learnability seriously: Before assuming that grammatical knowledge must be innate, we need to investigate what can be learned from realistic input, and how.
- Rule learning is more than tracking statistics: Learners constantly seek evidence for productive generalizations from imperfect data.
- Mechanisms matter: Computational models, behavioral experiments, and comparisons across learning systems help us identify what makes this kind of generalization possible.

# Thank you!


- Collaborator: Jingwen Sun
- All research participants
- Penn Language Science community and UNC Linguistics Dept, especially Charles Yang, Kathryn Schuler, John Trueswell, Misha Becker
- Xiaomeng Ma
- Members of UNC Language Learning & Computation Lab
- Funding: UNC-Chapel Hill College of Arts and Sciences, UNC AI Acceleration Program