# Small Language Models as a tool to investigate the acquisition of the Null-Subject constraint

Thomas Morton

August 21, 2025

I want to investigate **how humans learn syntactic generalizations** from linguistic evidence.

I seek to use **large language models as candidate models** of a language learner that we can **intervene on and investigate** the representations of.

I chose to use the **acquisition and representation of null subjects** as a case study in this investigation.

# The Null-Subject Constraint in Language Acquisition

## Cross-Linguistic Variation

Languages differ in whether they allow phonologically null subjects

**English (Non-pro-drop):**

- *She/*∅ runs*
- *It/*∅ rains*
- Overt subjects required

**Italian (Pro-drop):**

- *(Lei) corre* '(She) runs'
- *∅/*Ci piove* '(It) rains'
- Rich verbal agreement

**Acquisition Challenge (Hyams, 1986; Rizzi, 1994) :**

- Children show early null subjects across languages
- Must learn when overt subjects are required vs. optional

# The Poverty of Stimulus Problem

## Learnability Challenge

How do children learn from positive evidence alone? (Hyams & Wexler, 1993)

**The Problem:**

- Children rarely receive negative evidence (*"$\emptyset$ runs" is ungrammatical)

# The Poverty of Stimulus Problem

**Learnability Challenge**

How do children learn from positive evidence alone? (Hyams & Wexler, 1993)

**The Problem:**

- Children rarely receive negative evidence (*"∅ runs" is ungrammatical)
- Yet they successfully acquire language-specific constraints

# The Poverty of Stimulus Problem

**Learnability Challenge**

How do children learn from positive evidence alone? (Hyams & Wexler, 1993)

**The Problem:**

- Children rarely receive negative evidence (*"∅ runs" is ungrammatical)
- Yet they successfully acquire language-specific constraints
- **Poverty of Stimulus:** Input seems insufficient for learning (Chomsky, 1959)

# The Poverty of Stimulus Problem

**Learnability Challenge**

How do children learn from positive evidence alone? (Hyams & Wexler, 1993)

**The Problem:**

- Children rarely receive negative evidence (*"∅ runs" is ungrammatical)
- Yet they successfully acquire language-specific constraints
- **Poverty of Stimulus:** Input seems insufficient for learning (Chomsky, 1959)

**Competing Accounts:**

- **Indirect distributional evidence:** Orthogonal grammatical information guides linguistic generalizations (Yang, 2004)

# The Poverty of Stimulus Problem

**Learnability Challenge**

How do children learn from positive evidence alone? (Hyams & Wexler, 1993)

**The Problem:**

- Children rarely receive negative evidence (*"∅ runs" is ungrammatical)
- Yet they successfully acquire language-specific constraints
- **Poverty of Stimulus:** Input seems insufficient for learning (Chomsky, 1959)

**Competing Accounts:**

- **Indirect distributional evidence:** Orthogonal grammatical information guides linguistic generalizations (Yang, 2004)
- **Direct evidence:** Overt forms signal constraints (Hyams & Wexler, 1993)

# Large Language Models as Candidate Learners

**Key Advantages:**

- **Controlled input:** Manipulate training data systematically

# Large Language Models as Candidate Learners

**Key Advantages:**

- **Controlled input:** Manipulate training data systematically
- **Developmental tracking:** Observe learning over time

# Large Language Models as Candidate Learners

**Key Advantages:**

- **Controlled input:** Manipulate training data systematically
- **Developmental tracking:** Observe learning over time
- **Multiple architectures:** Test different learning mechanisms

# Large Language Models as Candidate Learners

**Key Advantages:**

- **Controlled input:** Manipulate training data systematically
- **Developmental tracking:** Observe learning over time
- **Multiple architectures:** Test different learning mechanisms

**Recent Evidence (Hu et al., 2020; Warstadt & Bowman, 2020):**

- LLMs acquire many syntactic generalizations

# Large Language Models as Candidate Learners

**Key Advantages:**

- **Controlled input:** Manipulate training data systematically
- **Developmental tracking:** Observe learning over time
- **Multiple architectures:** Test different learning mechanisms

**Recent Evidence (Hu et al., 2020; Warstadt & Bowman, 2020):**

- LLMs acquire many syntactic generalizations
- Show human-like developmental trajectories

# Large Language Models as Candidate Learners

**Key Advantages:**

- **Controlled input:** Manipulate training data systematically
- **Developmental tracking:** Observe learning over time
- **Multiple architectures:** Test different learning mechanisms

**Recent Evidence (Hu et al., 2020; Warstadt & Bowman, 2020):**

- LLMs acquire many syntactic generalizations
- Show human-like developmental trajectories
- BUT: Trained on massive datasets (>>child input)

# Large Language Models as Candidate Learners

**Key Advantages:**

- **Controlled input:** Manipulate training data systematically
- **Developmental tracking:** Observe learning over time
- **Multiple architectures:** Test different learning mechanisms

**Recent Evidence (Hu et al., 2020; Warstadt & Bowman, 2020):**

- LLMs acquire many syntactic generalizations
- Show human-like developmental trajectories
- BUT: Trained on massive datasets (>>child input)

**This Work:** Train models on developmentally-plausible corpora to test questions of sufficiency of the language input.

# The Contravariance Principle

## Why Do Models and Humans Converge? (Cao & Yamins, 2021)

Hard computational problems constrain possible solutions

**Core Insight:**

- **Hard problems** require satisfying multiple competing constraints

# The Contravariance Principle

## Why Do Models and Humans Converge? (Cao & Yamins, 2021)

Hard computational problems constrain possible solutions

**Core Insight:**

- **Hard problems** require satisfying multiple competing constraints
- **Few viable solutions** exist under such constraints

# The Contravariance Principle

## Why Do Models and Humans Converge? (Cao & Yamins, 2021)

Hard computational problems constrain possible solutions

**Core Insight:**

- **Hard problems** require satisfying multiple competing constraints
- **Few viable solutions** exist under such constraints
- **Different systems** solving the same hard problem converge

# The Contravariance Principle

## Why Do Models and Humans Converge? (Cao & Yamins, 2021)

Hard computational problems constrain possible solutions

**Core Insight:**

- **Hard problems** require satisfying multiple competing constraints
- **Few viable solutions** exist under such constraints
- **Different systems** solving the same hard problem converge

**Applied to Language:**

- Null-subject acquisition involves multiple competing constraints

# The Contravariance Principle

## Why Do Models and Humans Converge? (Cao & Yamins, 2021)

Hard computational problems constrain possible solutions

**Core Insight:**

- **Hard problems** require satisfying multiple competing constraints
- **Few viable solutions** exist under such constraints
- **Different systems** solving the same hard problem converge

**Applied to Language:**

- Null-subject acquisition involves multiple competing constraints
- Flexible learners (models and humans) should converge to similar solutions

# The Contravariance Principle

## Why Do Models and Humans Converge? (Cao & Yamins, 2021)

Hard computational problems constrain possible solutions

**Core Insight:**

- **Hard problems** require satisfying multiple competing constraints
- **Few viable solutions** exist under such constraints
- **Different systems** solving the same hard problem converge

**Applied to Language:**

- Null-subject acquisition involves multiple competing constraints
- Flexible learners (models and humans) should converge to similar solutions
- Convergence toward the simplest viable explanation

# The Planonic Representation Hypothesis

**Linking Theory: Internal Representations $\leftrightarrow$ Linguistic Competence**

Abstract syntactic knowledge should be observable in model representations

**Key Claims:**

- **Structural priming** reveals abstract grammatical representations (Bock, 1986)

# The Planonic Representation Hypothesis

## Linking Theory: Internal Representations ↔ Linguistic Competence

Abstract syntactic knowledge should be observable in model representations

**Key Claims:**

- **Structural priming** reveals abstract grammatical representations (Bock, 1986)
- **Cross-linguistic effects** suggest language-independent structure (Arnett et al., 2025; Michaelov et al., 2023)

# The Planonic Representation Hypothesis

## Linking Theory: Internal Representations $\leftrightarrow$ Linguistic Competence

Abstract syntactic knowledge should be observable in model representations

**Key Claims:**

- **Structural priming** reveals abstract grammatical representations (Bock, 1986)
- **Cross-linguistic effects** suggest language-independent structure (Arnett et al., 2025; Michaelov et al., 2023)
- **Null elements** can be positively represented (Momma et al., 2018)

# The Planonic Representation Hypothesis

**Linking Theory: Internal Representations $\leftrightarrow$ Linguistic Competence**

Abstract syntactic knowledge should be observable in model representations

**Key Claims:**

- **Structural priming** reveals abstract grammatical representations (Bock, 1986)
- **Cross-linguistic effects** suggest language-independent structure (Arnett et al., 2025; Michaelov et al., 2023)
- **Null elements** can be positively represented (Momma et al., 2018)

**Methodological Innovation:**

- Use structural priming to probe model competence

# The Planonic Representation Hypothesis

## Linking Theory: Internal Representations $\leftrightarrow$ Linguistic Competence

Abstract syntactic knowledge should be observable in model representations

**Key Claims:**

- **Structural priming** reveals abstract grammatical representations (Bock, 1986)
- **Cross-linguistic effects** suggest language-independent structure (Arnett et al., 2025; Michaelov et al., 2023)
- **Null elements** can be positively represented (Momma et al., 2018)

**Methodological Innovation:**

- Use structural priming to probe model competence
- Test representations beyond surface performance

# The Planonic Representation Hypothesis

**Linking Theory: Internal Representations $\leftrightarrow$ Linguistic Competence**

Abstract syntactic knowledge should be observable in model representations

**Key Claims:**

- **Structural priming** reveals abstract grammatical representations (Bock, 1986)
- **Cross-linguistic effects** suggest language-independent structure (Arnett et al., 2025; Michaelov et al., 2023)
- **Null elements** can be positively represented (Momma et al., 2018)

**Methodological Innovation:**

- Use structural priming to probe model competence
- Test representations beyond surface performance
- Bridge psycholinguistics and computational modeling

# Predictions for Null-Subject Learning

**Theoretical Prediction:**

- Models solving the hard null-subject problem should converge to human-like representations

# Predictions for Null-Subject Learning

**Theoretical Prediction:**

- Models solving the hard null-subject problem should converge to human-like representations
- The Model's generalizations should effect language learning in the same way that it effects human multilingual learning

# Predictions for Null-Subject Learning

**Theoretical Prediction:**

- Models solving the hard null-subject problem should converge to human-like representations
- The Model's generalizations should effect language learning in the same way that it effects human multilingual learning
- Abstract representations should transcend surface linguistic differences

# Predictions for Null-Subject Learning

**Theoretical Prediction:**

- Models solving the hard null-subject problem should converge to human-like representations
- The Model's generalizations should effect language learning in the same way that it effects human multilingual learning
- Abstract representations should transcend surface linguistic differences

**Empirical Tests:**

- **Chapter 1:** Do models learn null-subject constraints like humans?
- **Chapter 2:** Do bilingual models show human-like transfer effects?
- **Chapter 3:** Do models exhibit cross-linguistic structural priming?

# Contents

# Chapter 1: A controlled rearing study of the null-subject constraint in English

**Research Questions:**

- Do English Small Language Models learn the null-subject constraint in a human-like way?

# Chapter 1: A controlled rearing study of the null-subject constraint in English

**Research Questions:**

- Do English Small Language Models learn the null-subject constraint in a human-like way?
- What kind of linguistic information contributes to the learning of that constraint?

# Chapter 1: A controlled rearing study of the null-subject constraint in English

**Research Questions:**

- Do English Small Language Models learn the null-subject constraint in a human-like way?
- What kind of linguistic information contributes to the learning of that constraint?
- Do model's show human-like behavior in contexts with higher processing demands?

# Chapter 1: A controlled rearing study of the null-subject constraint in English

**Research Questions:**

- Do English Small Language Models learn the null-subject constraint in a human-like way?
- What kind of linguistic information contributes to the learning of that constraint?
- Do model's show human-like behavior in contexts with higher processing demands?

**Approach:**

- Controlled rearing experiments with ablated training data

# Chapter 1: A controlled rearing study of the null-subject constraint in English

**Research Questions:**

- Do English Small Language Models learn the null-subject constraint in a human-like way?
- What kind of linguistic information contributes to the learning of that constraint?
- Do model's show human-like behavior in contexts with higher processing demands?

**Approach:**

- Controlled rearing experiments with ablated training data
- Systematic removal of specific linguistic evidence types

# Chapter 1: A controlled rearing study of the null-subject constraint in English

**Research Questions:**

- Do English Small Language Models learn the null-subject constraint in a human-like way?
- What kind of linguistic information contributes to the learning of that constraint?
- Do model's show human-like behavior in contexts with higher processing demands?

**Approach:**

- Controlled rearing experiments with ablated training data
- Systematic removal of specific linguistic evidence types
- Manipulation of evaluation stimuli to examine contextual processing effects

# The Null-Subject Constraint in English

## What is the null-subject constraint?

English requires overt subjects in finite clauses (unlike Spanish, Italian, etc.)

**Adult English Constraint:**

- * ∅ *Finished the book* (ungrammatical)
- ✓ *She finished the book* (grammatical)

**Child Null-Subject Use Examples:**

- *Shake hands.*
- *Turn light off.*
- *Want go get it.*
- *Show mommy that.*
- *Now making muffins.*

# Performance vs. Competence Accounts

## Why Do Children Drop Subjects?

Two competing explanations for child null-subject patterns

**Performance Account (L. Bloom, 1970; P. Bloom, 1990):**

- Children drop subjects under processing load
- Cognitive resource limitations
- More drops with negation, longer sentences
- Subject/object asymmetry due to planning

**Competence Account (Hyams, 1986; Hyams & Wexler, 1993):**

- Children initially set null-subject parameter
- Must learn overt subject requirement
- Grammatical learning, not performance
- Direct evidence from overt pronouns

**Key Debate:** Processing limitation vs. grammatical parameter setting

# Specific Theoretical Accounts

**Yang's Variational Learning (Yang, 2003, 2004):**

- **Expletives** are crucial unambiguous evidence

# Specific Theoretical Accounts

**Yang's Variational Learning (Yang, 2003, 2004):**

- **Expletives** are crucial unambiguous evidence
- Probabilistic grammar competition

# Specific Theoretical Accounts

**Yang's Variational Learning (Yang, 2003, 2004):**

- **Expletives** are crucial unambiguous evidence
- Probabilistic grammar competition
- Rarity of expletives explains slow English acquisition

# Specific Theoretical Accounts

**Yang's Variational Learning (Yang, 2003, 2004):**

- **Expletives** are crucial unambiguous evidence
- Probabilistic grammar competition
- Rarity of expletives explains slow English acquisition

**Hyams' Triggering Theory (Hyams & Wexler, 1993):**

- **Non-uniform verbal morphology** triggers parameter reset

# Specific Theoretical Accounts

**Yang's Variational Learning (Yang, 2003, 2004):**

- **Expletives** are crucial unambiguous evidence
- Probabilistic grammar competition
- Rarity of expletives explains slow English acquisition

**Hyams' Triggering Theory (Hyams & Wexler, 1993):**

- **Non-uniform verbal morphology** triggers parameter reset
- Italian: uniformly rich, Mandarin: uniformly poor $\rightarrow$ null subjects

# Specific Theoretical Accounts

**Yang's Variational Learning (Yang, 2003, 2004):**

- **Expletives** are crucial unambiguous evidence
- Probabilistic grammar competition
- Rarity of expletives explains slow English acquisition

**Hyams' Triggering Theory (Hyams & Wexler, 1993):**

- **Non-uniform verbal morphology** triggers parameter reset
- Italian: uniformly rich, Mandarin: uniformly poor → null subjects
- English: inconsistent system → overt subjects required

# Specific Theoretical Accounts

**Yang's Variational Learning (Yang, 2003, 2004):**

- **Expletives** are crucial unambiguous evidence
- Probabilistic grammar competition
- Rarity of expletives explains slow English acquisition

**Hyams' Triggering Theory (Hyams & Wexler, 1993):**

- **Non-uniform verbal morphology** triggers parameter reset
- Italian: uniformly rich, Mandarin: uniformly poor → null subjects
- English: inconsistent system → overt subjects required

**Duguine's Inverse Approach (Duguine, 2017):**

- **Rich determiners + weak agreement** → overt subjects

‘

# Specific Theoretical Accounts

**Yang's Variational Learning (Yang, 2003, 2004):**

- **Expletives** are crucial unambiguous evidence
- Probabilistic grammar competition
- Rarity of expletives explains slow English acquisition

**Hyams' Triggering Theory (Hyams & Wexler, 1993):**

- **Non-uniform verbal morphology** triggers parameter reset
- Italian: uniformly rich, Mandarin: uniformly poor $\rightarrow$ null subjects
- English: inconsistent system $\rightarrow$ overt subjects required

**Duguine's Inverse Approach (Duguine, 2017):**

- **Rich determiners + weak agreement** $\rightarrow$ overt subjects
- Nominal domain provides indirect evidence

‘

# Specific Theoretical Accounts

**Yang's Variational Learning (Yang, 2003, 2004):**

- **Expletives** are crucial unambiguous evidence
- Probabilistic grammar competition
- Rarity of expletives explains slow English acquisition

**Hyams' Triggering Theory (Hyams & Wexler, 1993):**

- **Non-uniform verbal morphology** triggers parameter reset
- Italian: uniformly rich, Mandarin: uniformly poor → null subjects
- English: inconsistent system → overt subjects required

**Duguine's Inverse Approach (Duguine, 2017):**

- **Rich determiners + weak agreement** → overt subjects
- Nominal domain provides indirect evidence
- Focus shifts from verbal to determiner system

'

# Experimental Design Rationale

**Testing Causal Contributions**

Each experiment targets specific theoretical predictions

**Theory-Experiment Mapping:**

- **Exp 1 - Remove Expletives:** Tests Yang's prediction about expletive evidence

# Experimental Design Rationale

## Testing Causal Contributions

Each experiment targets specific theoretical predictions

**Theory-Experiment Mapping:**

- **Exp 1 - Remove Expletives:** Tests Yang's prediction about expletive evidence
- **Exp 2 - Impoverish Determiners:** Tests Duguine's rich determiner hypothesis

# Experimental Design Rationale

## Testing Causal Contributions

Each experiment targets specific theoretical predictions

**Theory-Experiment Mapping:**

- **Exp 1 - Remove Expletives:** Tests Yang's prediction about expletive evidence
- **Exp 2 - Impoverish Determiners:** Tests Duguine's rich determiner hypothesis
- **Exp 3 - Remove Articles:** Further tests determiner system importance

# Experimental Design Rationale

## Testing Causal Contributions

Each experiment targets specific theoretical predictions

**Theory-Experiment Mapping:**

- **Exp 1 - Remove Expletives:** Tests Yang's prediction about expletive evidence
- **Exp 2 - Impoverish Determiners:** Tests Duguine's rich determiner hypothesis
- **Exp 3 - Remove Articles:** Further tests determiner system importance
- **Exp 4 - Lemmatize Verbs:** Tests Hyams' verbal morphology prediction

# Experimental Design Rationale

## Testing Causal Contributions

Each experiment targets specific theoretical predictions

**Theory-Experiment Mapping:**

- **Exp 1 - Remove Expletives:** Tests Yang's prediction about expletive evidence
- **Exp 2 - Impoverish Determiners:** Tests Duguine's rich determiner hypothesis
- **Exp 3 - Remove Articles:** Further tests determiner system importance
- **Exp 4 - Lemmatize Verbs:** Tests Hyams' verbal morphology prediction
- **Exp 5 - Remove Pronouns:** Tests direct vs. indirect evidence accounts

# Experimental Design: Controlled Rearing

## Controlled Rearing Paradigm

Train models on systematically modified datasets to isolate evidence contributions

**Ablation Experiments:**

0. **Baseline:** Full training corpus
1. **Remove Expletives:** No *it/there* expletive constructions
2. **Impoverish Determiners:** Reduce *a/the* to *DET*
3. **Remove Articles:** No *a/the* entirely
4. **Lemmatize Verbs:** Remove *-s/-ed/-ing* morphology
5. **Remove Subject Pronominals:** No *I/you/he/she/it/we/they*

**Evaluation:** Null vs. overt subject preferences in controlled contexts

# Measures and Analysis: Overview

## Data and Coding

- Binary outcome: `over preference` $= 1$ when *overt* $<$ *null* surprisal
- Factors: `model`, `form_type`, `item_group`, `form`
- Training progress: $\log_{10}(\texttt{checkpoint} + 1)$
- Baseline condition as reference level

## Outcome Definition

- Minimal pairs: *null* vs. *overt* subject realization
- Binary response $Y \in \{0, 1\}$ encodes preference
- End-state: *overt* preference (probability scale)
- Acquisition-time: *null* preference

# Logistic Models: Learning Curves and Splines

## GLMMs

$$\operatorname{logit} \Pr(Y = 1) = \beta_0 + \operatorname{ns}\big(\log_{10}(t+1),\ k\big) + u_i$$

- Natural spline over log-checkpoint, complexity $k$
- Random intercept: $u_{\texttt{item}} \sim \mathcal{N}(0, \sigma^2)$
- Spline selection: AIC over $K \in \{3, \ldots, 7\}$

## Training Progress

- $\text{Log}_{10}$ scale: $\{0, 10, 100, 1\text{K}, 10\text{K}\}$
- Reflects neural network log-learning dynamics
- Uniform checkpointing across conditions

# Age of Acquisition (AoA) Analysis

## $t_{50}$ (Chance-Level Acquisition)

- Last crossing of 0.50 after burn-in ($\geq$ 100 checkpoints)
- Linear interpolation between fitted points
- Right-censored if no crossing
- Bootstrap 95% CIs ($n = 500$)

## $\mathrm{AoA}_{1/2}$ (Halfway-to-Asymptote)

- End-state $p_\infty$ from last 10% of training
- Threshold: $\theta = (p_\infty + 0.5)/2$
- First post-burn-in crossing of $\theta$
- Between-model $\Delta\mathrm{AoA}_{1/2}$ via paired bootstrap

# Materials: BabyLM Dataset

**Training Corpus**

- 90M word corpus designed for human-sized models
- Linguistically diverse with child-directed speech
- Models linguistic input of 10-14 year old child
- 10M word held-out test set
- 10M word ablation replacement set

**Dataset Composition**

- CHILDES (child-directed speech): 29M words
- Project Gutenberg (children's stories): 26M words
- OpenSubtitles (movie subtitles): 20M words
- Simple English Wikipedia: 15M words
- BNC dialogue + Switchboard: 9M words

# Evaluation Stimuli: Null vs. Overt Subjects

### Core Contrasts (English non-pro-drop)

- **Person/Number**: Anna finished. She/*∅ thinks...
- **Control**: Maria convinced her brother ∅/*him to leave
- **Expletives**: *∅/It seems that students passed
- **Topic shift**: Anna called Mark and *∅/he refused

### Minimal Pairs Design

- Sentences differ only in subject realization
- Lexical and contextual content held constant
- Testsfamilies
- Evaluates grammatical vs. processing accounts

# Processing Manipulations

**Context Complexity (Bloom 1990)**

1. **Simple**: The dog barked. He/*∅ scared...
2. **Long NPs**: The large brown dog with red collar barked...
3. **Embedded**: The dog that lived in the house...

**Negation Effects (Bloom 1970)**

1. **Target negation**: She/*∅ doesn't think...
2. **Context negation**: Anna didn't finish. She/*∅ thinks...
3. **Double negation**: Both context and target negated

*Tests processing load effects on subject drop preferences*

# Baseline Model – Training Curves

Figure 1: Model preference for null and overt evaluation stimuli over training, training steps transformed to log-scale to reflect model log-learning dynamics for Experiment 0 - Baseline

# Baseline Model – Training Curves

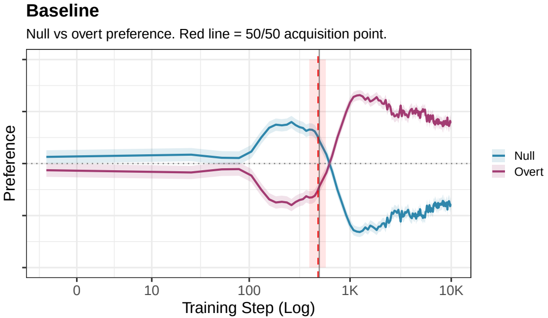



Figure 1: Model preference for null and overt evaluation stimuli over training, training steps transformed to log-scale to reflect model log-learning dynamics for Experiment 0 - Baseline

# Baseline Model – Training Curves

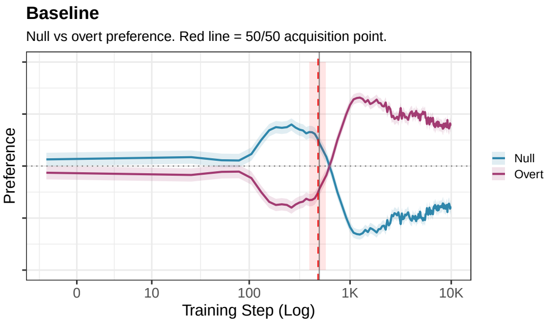



Figure 1: Model preference for null and overt evaluation stimuli over training, training steps transformed to log-scale to reflect model log-learning dynamics for Experiment 0 - Baseline

- Age of Acquisition analysis revealed that baseline achieved **AoA at checkpoint 727** (95% CI [664, 791]).

# Baseline Model – Training Curves

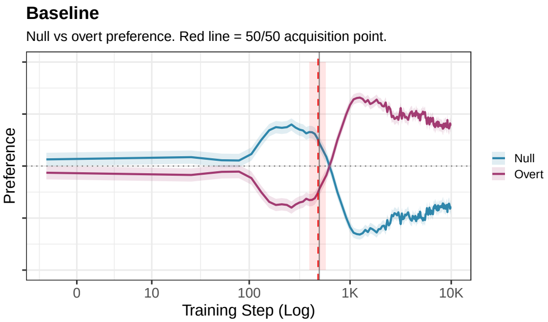



Figure 1: Model preference for null and overt evaluation stimuli over training, training steps transformed to log-scale to reflect model log-learning dynamics for Experiment 0 - Baseline

- Age of Acquisition analysis revealed that baseline achieved **AoA at checkpoint 727** (95% CI [664, 791]).
- A **63.4% preference for null subjects over first epoch** (95% CI [62.7, 64.1], $p<.001$)

# Baseline Model – Training Curves

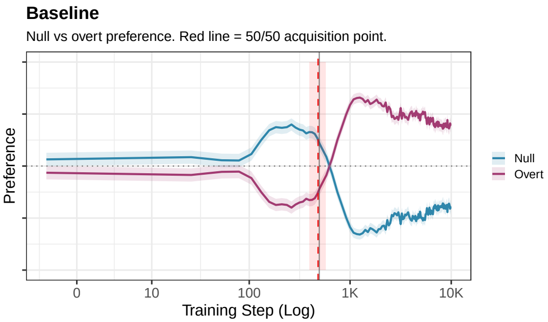



Figure 1: Model preference for null and overt evaluation stimuli over training, training steps transformed to log-scale to reflect model log-learning dynamics for Experiment 0 - Baseline

- Age of Acquisition analysis revealed that baseline achieved **AoA at checkpoint 727** (95% CI [664, 791]).
- A **63.4% preference for null subjects over first epoch** (95% CI [62.7, 64.1], $p<.001$)
- a **69.6% preference for overt subjects in the last two epochs of training** (95% CI [66.5%, 72.5%], $p < .001$)

# Exp 1: 'Remove Expletives' – Training Curves

Figure 2: Model overt preference over training, training steps transformed to log-scale to reflect model log-learning dynamics comparing Experiment 0 and Experiment 1.

# Exp 1: 'Remove Expletives' – Training Curves

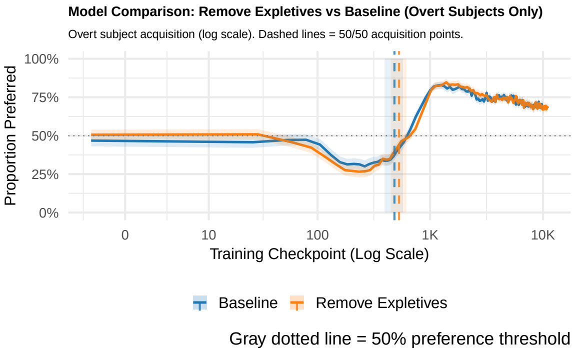



Figure 2: Model overt preference over training, training steps transformed to log-scale to reflect model log-learning dynamics comparing Experiment 0 and Experiment 1.

# Exp 1: ‘Remove Expletives’ – Training Curves

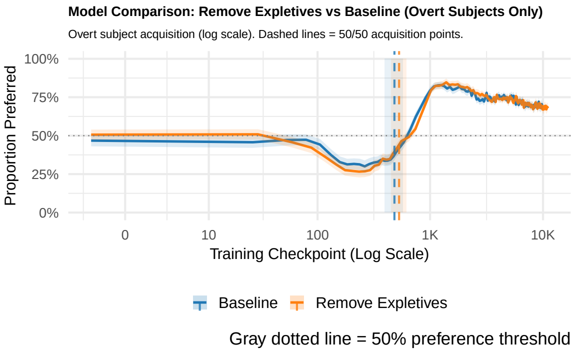



Figure 2: Model overt preference over training, training steps transformed to log-scale to reflect model log-learning dynamics comparing Experiment 0 and Experiment 1.

- Age of Acquisition analysis revealed that baseline achieved **AoA at checkpoint 767** (95% CI [709, 821]).
  - Which is significantly later than the baseline model ($\Delta$AoA = 39 epochs, 95% CI [24, 55], p<.001)

# Exp 1: ‘Remove Expletives’ – Training Curves

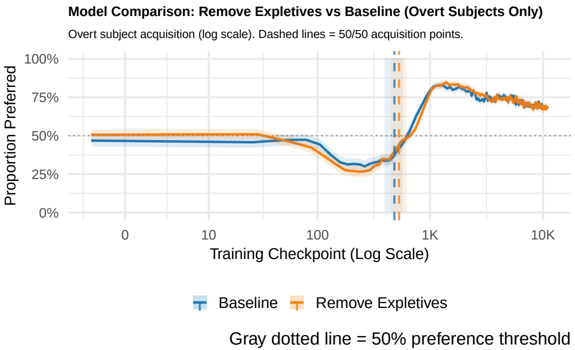



Figure 2: Model overt preference over training, training steps transformed to log-scale to reflect model log-learning dynamics comparing Experiment 0 and Experiment 1.

- Age of Acquisition analysis revealed that baseline achieved **AoA at checkpoint 767** (95% CI [709, 821]).
    - Which is significantly later than the baseline model ($\Delta$AoA $=$ 39 epochs, 95% CI [24, 55], p$<$.001)
- Start-performance and end-performance did not significantly differ from base model.

# Exp 2: 'Impoverish Determiners' – Training Curves

Figure 3: Model overt preference over training, training steps transformed to log-scale to reflect model log-learning dynamics comparing Experiment 0 and Experiment 1.

# Exp 2: 'Impoverish Determiners' – Training Curves

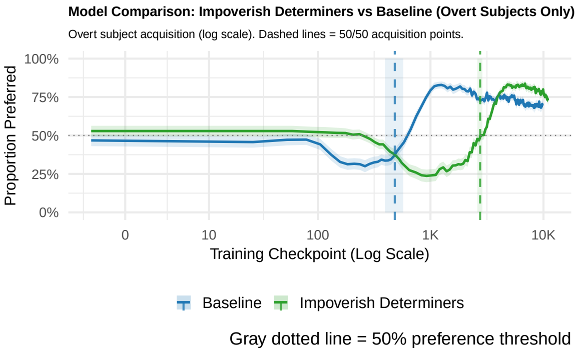



Figure 3: Model overt preference over training, training steps transformed to log-scale to reflect model log-learning dynamics comparing Experiment 0 and Experiment 1.

# Exp 2: ‘Impoverish Determiners’ – Training Curves

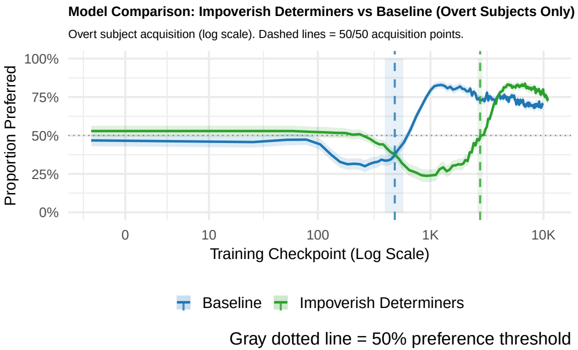



Figure 3: Model overt preference over training, training steps transformed to log-scale to reflect model log-learning dynamics comparing Experiment 0 and Experiment 1.

- Age of Acquisition analysis revealed that baseline achieved **AoA at checkpoint 3400** (95% CI [3307, 3499]).
    - Which is significantly later than baseline ($\Delta$AoA = 2672 epochs, 95% CI [2620, 2724], p<.001)

# Exp 2: 'Impoverish Determiners' – Training Curves

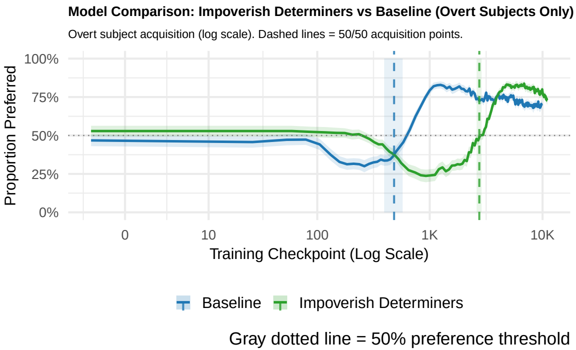



Figure 3: Model overt preference over training, training steps transformed to log-scale to reflect model log-learning dynamics comparing Experiment 0 and Experiment 1.

- Age of Acquisition analysis revealed that baseline achieved **AoA at checkpoint 3400** (95% CI [3307, 3499]).
  - Which is significantly later than baseline ($\Delta$AoA $=$ 2672 epochs, 95% CI [2620, 2724], p$<$.001)
- The model had a significant, but smaller preference for null subjects by the end of the first epoch.

# Exp 2: 'Impoverish Determiners' – Training Curves

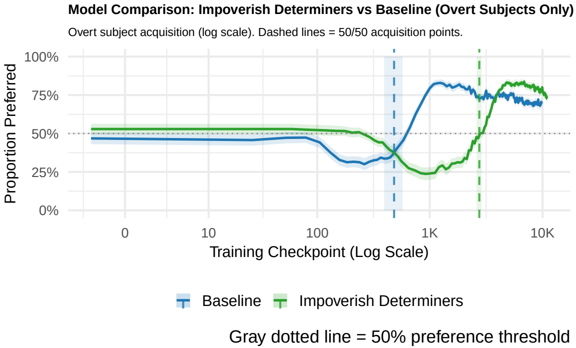



Figure 3: Model overt preference over training, training steps transformed to log-scale to reflect model log-learning dynamics comparing Experiment 0 and Experiment 1.

- Age of Acquisition analysis revealed that baseline achieved **AoA at checkpoint 3400** (95% CI [3307, 3499]).
  - Which is significantly later than baseline ($\Delta$AoA $=$ 2672 epochs, 95% CI [2620, 2724], p$<$.001)
- The model had a significant, but smaller preference for null subjects by the end of the first epoch.
- By the end of the final two epochs, it has the strongest preference of all models for overt subjects

# Exp 3: 'Remove Articles' – Training Curves

Figure 4: Model overt preference over training, training steps transformed to log-scale to reflect model log-learning dynamics comparing Experiment 0 and Experiment 3.

# Exp 3: 'Remove Articles' – Training Curves

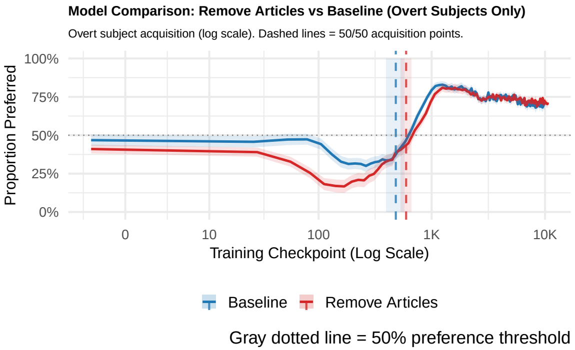



Figure 4: Model overt preference over training, training steps transformed to log-scale to reflect model log-learning dynamics comparing Experiment 0 and Experiment 3.

# Exp 3: ‘Remove Articles’ – Training Curves

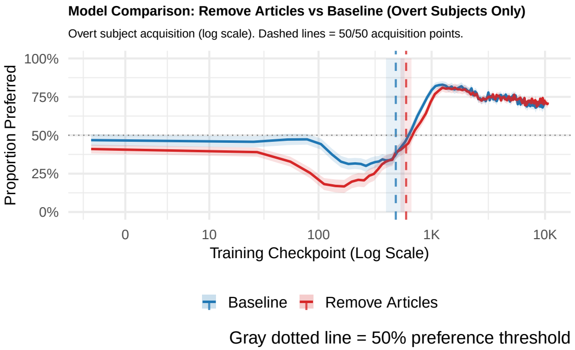



Figure 4: Model overt preference over training, training steps transformed to log-scale to reflect model log-learning dynamics comparing Experiment 0 and Experiment 3.

- Age of Acquisition analysis revealed that this model achieved **AoA at checkpoint 807** (95% CI [758, 861]).
  - Which is significantly later than baseline ($\Delta$AoA $=$ 80 epochs, 95% CI [81, 108], p$<$.001)

# Exp 3: 'Remove Articles' – Training Curves

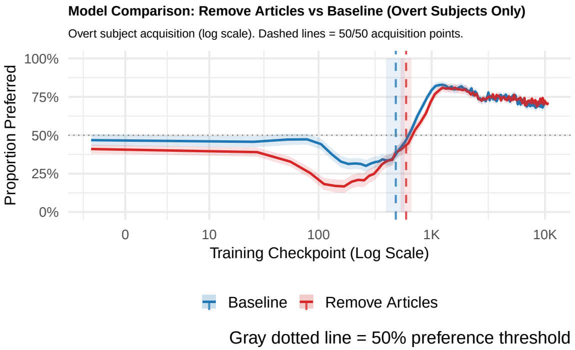



Figure 4: Model overt preference over training, training steps transformed to log-scale to reflect model log-learning dynamics comparing Experiment 0 and Experiment 3.

- Age of Acquisition analysis revealed that this model achieved **AoA at checkpoint 807** (95% CI [758, 861]).
  - Which is significantly later than baseline ($\Delta$AoA $=$ 80 epochs, 95% CI [81, 108], p$<$.001)
- Shows significantly stronger null preference in first epoch (71.7%) compared to baseline.

# Exp 3: 'Remove Articles' – Training Curves

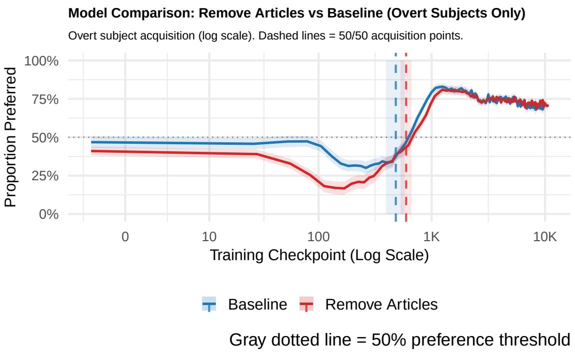



Figure 4: Model overt preference over training, training steps transformed to log-scale to reflect model log-learning dynamics comparing Experiment 0 and Experiment 3.

- Age of Acquisition analysis revealed that this model achieved **AoA at checkpoint 807** (95% CI [758, 861]).
    - Which is significantly later than baseline ($\Delta$AoA $=$ 80 epochs, 95% CI [81, 108], p$<$.001)
- Shows significantly stronger null preference in first epoch (71.7%) compared to baseline.
- End-state overt preference (68.2%) is significantly lower than baseline model.

# Exp 4: ‘Lemmatize Verbs’ – Training Curves

Figure 5: Model overt preference over training, training steps transformed to log-scale to reflect model log-learning dynamics comparing Experiment 0 and Experiment 4.

# Exp 4: ‘Lemmatize Verbs’ – Training Curves

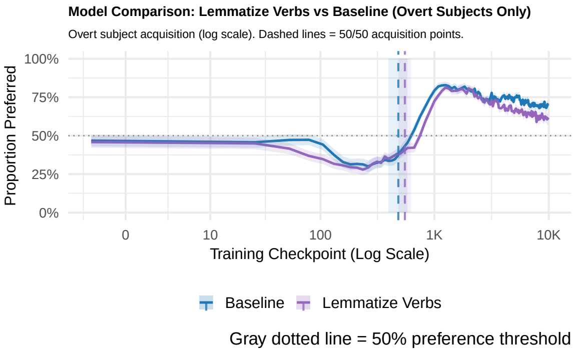



Figure 5: Model overt preference over training, training steps transformed to log-scale to reflect model log-learning dynamics comparing Experiment 0 and Experiment 4.

# Exp 4: ‘Lemmatize Verbs’ – Training Curves

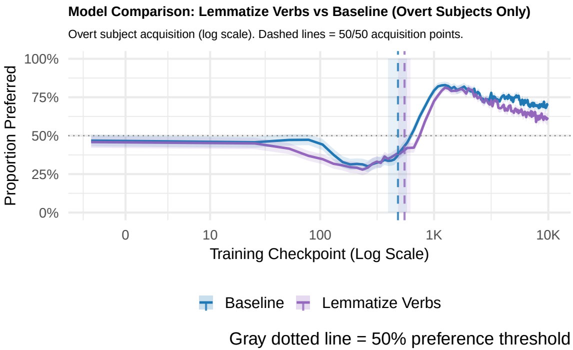



Figure 5: Model overt preference over training, training steps transformed to log-scale to reflect model log-learning dynamics comparing Experiment 0 and Experiment 4.

- Age of Acquisition analysis revealed that this model achieved **AoA at checkpoint 705** (95% CI [660, 748]).
    - Which is significantly **earlier** than baseline ($\Delta$AoA = -22 epochs, 95% CI [-43, -1.65], p = .034)

# Exp 4: ‘Lemmatize Verbs’ – Training Curves

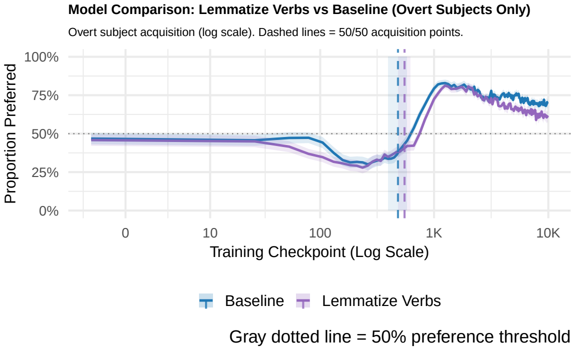



Figure 5: Model overt preference over training, training steps transformed to log-scale to reflect model log-learning dynamics comparing Experiment 0 and Experiment 4.

- Age of Acquisition analysis revealed that this model achieved **AoA at checkpoint 705** (95% CI [660, 748]).
    - Which is significantly **earlier** than baseline ($\Delta$AoA = -22 epochs, 95% CI [-43, -1.65], p = .034)
- Fastest acquisition among all interventions.

# Exp 5: ‘Remove Subject Pronominals’ – Training Curves

Figure 6: Model overt preference over training, training steps transformed to log-scale to reflect model log-learning dynamics comparing Experiment 0 and Experiment 5.

# Exp 5: ‘Remove Subject Pronominals’ – Training Curves

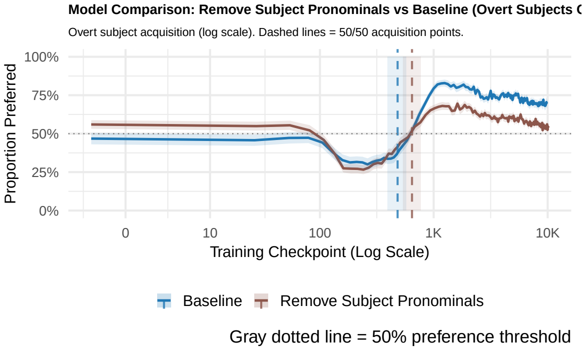



Figure 6: Model overt preference over training, training steps transformed to log-scale to reflect model log-learning dynamics comparing Experiment 0 and Experiment 5.

# Exp 5: ‘Remove Subject Pronominals’ – Training Curves

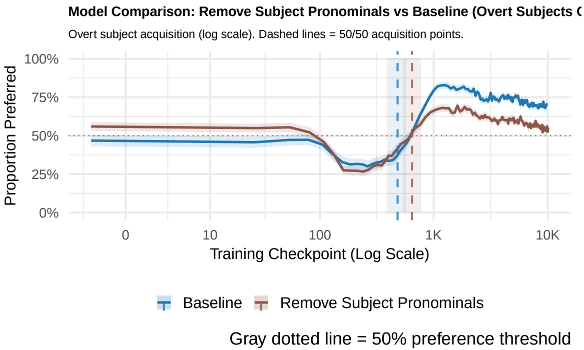



Figure 6: Model overt preference over training, training steps transformed to log-scale to reflect model log-learning dynamics comparing Experiment 0 and Experiment 5.

- Age of Acquisition analysis revealed that this model achieved **AoA at checkpoint 774** (95% CI [706, >5000]).
  - Slightly later than baseline ($\Delta$AoA = 47 epochs, $p<.05$)

# Exp 5: 'Remove Subject Pronominals' – Training Curves

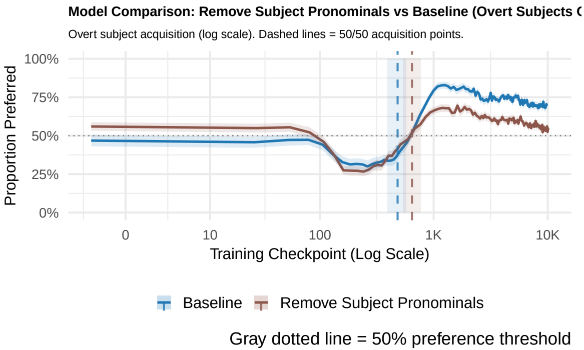
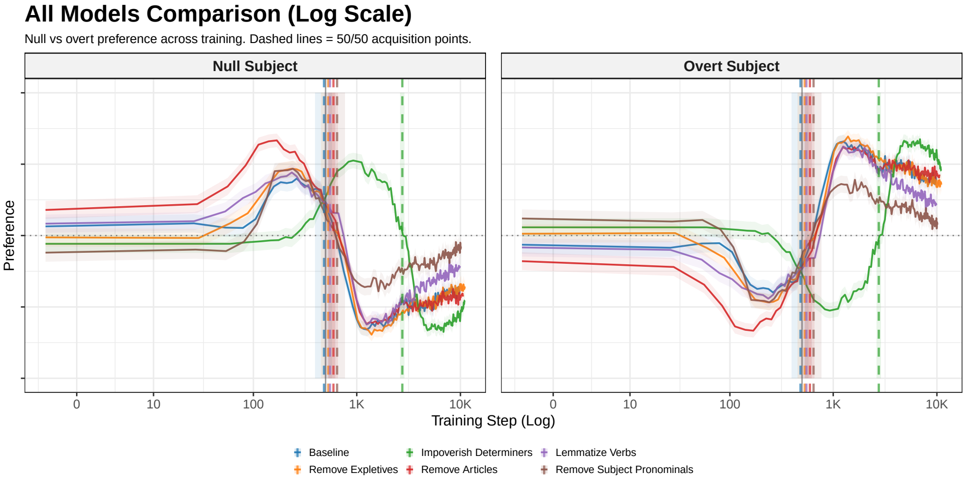



Figure 6: Model overt preference over training, training steps transformed to log-scale to reflect model log-learning dynamics comparing Experiment 0 and Experiment 5.

- Age of Acquisition analysis revealed that this model achieved **AoA at checkpoint 774** (95% CI [706, >5000]).
  - Slightly later than baseline ($\Delta$AoA = 47 epochs, p<.05)
- Weakest overall overt preference (54.4%) among all models.

# Cross-Model Comparison

Figure 7: Cross-model comparison of null subject acquisition trajectories (log scale)



# Processing Account: Predicted vs. Observed

**Bloom's Processing Account Prediction**

Under increased processing load, children should drop subjects MORE frequently

**Processing Manipulations:**

- Long noun phrases
- Embedded relative clauses
- Negation contexts
- Target vs. context negation

**Implication:** Do LLMs omit more subjects in contexts with heavy processing load?

# Exp 5: ‘Remove Subject Pronominals’ – Training Curves

Figure 8: Model overt preference over training, training steps transformed to log-scale to reflect model log-learning dynamics comparing Experiment 0 and Experiment 5.

# Exp 5: ‘Remove Subject Pronominals’ – Training Curves

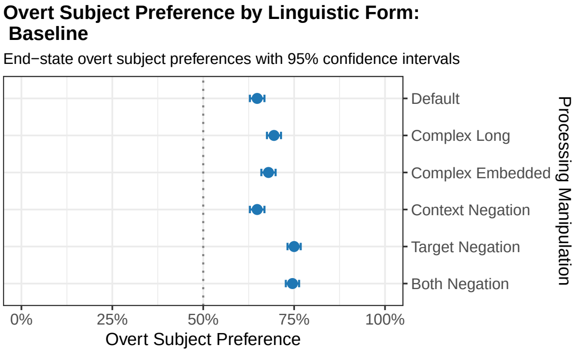



Figure 8: Model overt preference over training, training steps transformed to log-scale to reflect model log-learning dynamics comparing Experiment 0 and Experiment 5.

# Exp 5: ‘Remove Subject Pronominals’ – Training Curves

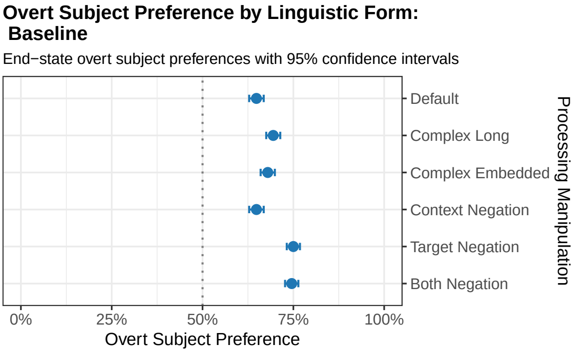



Figure 8: Model overt preference over training, training steps transformed to log-scale to reflect model log-learning dynamics comparing Experiment 0 and Experiment 5.

- Negation shows the strongest influence on models’ choice for overt subjects.

# Exp 5: 'Remove Subject Pronominals' – Training Curves

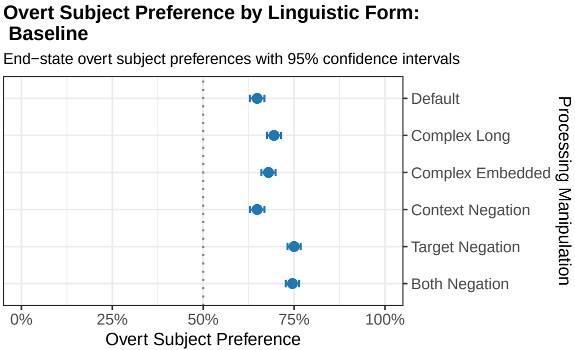



Figure 8: Model overt preference over training, training steps transformed to log-scale to reflect model log-learning dynamics comparing Experiment 0 and Experiment 5.

- Negation shows the strongest influence on models' choice for overt subjects.
- The model under these contexts show increased overt preference

# Exp 5: 'Remove Subject Pronominals' – Training Curves

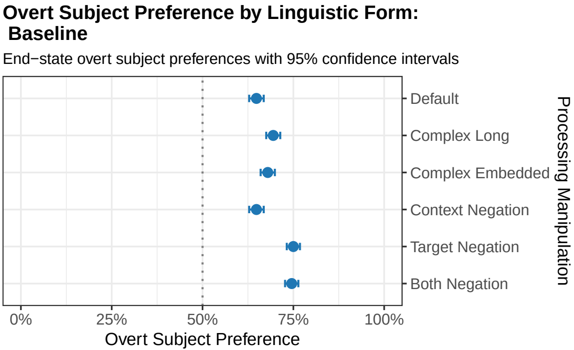



Figure 8: Model overt preference over training, training steps transformed to log-scale to reflect model log-learning dynamics comparing Experiment 0 and Experiment 5.

- Negation shows the strongest influence on models' choice for overt subjects.
- The model under these contexts show increased overt preference
- This is counter to the reported human pattern of higher null subject use in negation contexts.

# Processing Effects Across All Models

| Form | Baseline | Rmv. Expletives | Impvr. Detrmn. | Rmv. Articles | Lemmatize Verbs | Rmv. Subject Pronominals |
|---|---|---|---|---|---|---|
| Complex Long | ✓ | | ✓ | | | |
| Complex Emb | | | ✓ | | | |
| Context Negation | | | | | | |
| Target Negation | ✓ | ✓ | | ✓ | ✓ | ✓ |
| Both Negation | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

**Table 1:** Syntactic forms showing significant deviation from default performance by experimental model

# Processing Effects Across All Models

| Form | Baseline | Rmv. Expletives | Impvr. Detrmn. | Rmv. Articles | Lemmatize Verbs | Rmv. Subject Pronominals |
|---|---|---|---|---|---|---|
| Complex Long | ✓ | | ✓ | | | |
| Complex Emb | | | ✓ | | | |
| Context Negation | | | | | | |
| Target Negation | ✓ | ✓ | | ✓ | ✓ | ✓ |
| Both Negation | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

**Table 1:** Syntactic forms showing significant deviation from default performance by experimental model

- **Target/Both negation:** Universally increases overt preference

# Processing Effects Across All Models

| Form | Baseline | Rmv. Expletives | Impvr. Detrmn. | Rmv. Articles | Lemmatize Verbs | Rmv. Subject Pronominals |
|---|---|---|---|---|---|---|
| Complex Long | ✓ | | ✓ | | | |
| Complex Emb | | | ✓ | | | |
| Context Negation | | | | | | |
| Target Negation | ✓ | ✓ | | ✓ | ✓ | ✓ |
| Both Negation | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

**Table 1:** Syntactic forms showing significant deviation from default performance by experimental model

- **Target/Both negation:** Universally increases overt preference
- **Complex syntax:** Largely does not increase overt preference

# Processing Effects Across All Models

| Form | Baseline | Rmv. Expletives | Impvr. Detrmn. | Rmv. Articles | Lemmatize Verbs | Rmv. Subject Pronominals |
|---|---|---|---|---|---|---|
| Complex Long | ✓ | | ✓ | | | |
| Complex Emb | | | ✓ | | | |
| Context Negation | | | | | | |
| Target Negation | ✓ | ✓ | | ✓ | ✓ | ✓ |
| Both Negation | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

**Table 1:** Syntactic forms showing significant deviation from default performance by experimental model

- **Target/Both negation:** Universally increases overt preference
- **Complex syntax:** Largely does not increase overt preference
- **Context negation:** No effect across models

# Implications for Processing Theories

**Traditional View**

- Processing load $\rightarrow$ omit subjects
- "Good enough" processing
- Resource limitation effects

**Model Behavior**

- Processing load $\rightarrow$ *insert* subjects
- More explicit under complexity
- Robust to context effects

**Possible Explanations:**

- Models and children process complexity fundamentally differently
- Processing accounts may not fully explain child null subject errors
- Need empirical validation of processing effects in human production

# Universal Early Null Subject Stage

**Surprising Finding: All Models Show Initial Null Subject Preference**

Despite English being overt-subject, ALL models prefer null subjects early in training

**Theoretical Implications:**

# Universal Early Null Subject Stage

**Surprising Finding: All Models Show Initial Null Subject Preference**

Despite English being overt-subject, ALL models prefer null subjects early in training

**Theoretical Implications:**

- Consistent with some child acquisition patterns (null-first accounts)

# Universal Early Null Subject Stage

**Surprising Finding: All Models Show Initial Null Subject Preference**

Despite English being overt-subject, ALL models prefer null subjects early in training

**Theoretical Implications:**

- Consistent with some child acquisition patterns (null-first accounts)
- Contradicts Bloom's prediction that English children default to overt subjects

# Universal Early Null Subject Stage

**Surprising Finding: All Models Show Initial Null Subject Preference**

Despite English being overt-subject, ALL models prefer null subjects early in training

**Theoretical Implications:**

- Consistent with some child acquisition patterns (null-first accounts)
- Contradicts Bloom's prediction that English children default to overt subjects
- Could reflect model architecture bias OR environmental evidence

# Universal Early Null Subject Stage

**Surprising Finding: All Models Show Initial Null Subject Preference**

Despite English being overt-subject, ALL models prefer null subjects early in training

**Theoretical Implications:**

- Consistent with some child acquisition patterns (null-first accounts)
- Contradicts Bloom's prediction that English children default to overt subjects
- Could reflect model architecture bias OR environmental evidence
- Most models acquire English-like preferences after approx. 1.5 epochs.

# Universal Early Null Subject Stage

**Surprising Finding: All Models Show Initial Null Subject Preference**

Despite English being overt-subject, ALL models prefer null subjects early in training

**Theoretical Implications:**

- Consistent with some child acquisition patterns (null-first accounts)
- Contradicts Bloom's prediction that English children default to overt subjects
- Could reflect model architecture bias OR environmental evidence
- Most models acquire English-like preferences after approx. 1.5 epochs.

**Question:** Is this a learning bias or evidence from the input environment?

# Evidence Types: Shortcuts vs. Deep Learning

## Direct Evidence

- **Subject pronouns:** Critical
- Remove pronouns → near-chance performance
- Supports Hyams' direct evidence account

## Indirect Evidence

- **Determiners:** Provide "shortcuts"
- **Verbal morphology:** Affects final strength
- **Expletives:** Minimal effect

## Grokking Hypothesis

Removing shortcuts (determiners) forces slower but potentially more robust generalization

# Broader Theoretical Implications

**What This Study Challenges**

Multiple acquisition theories do not capture model learning behavior

**Challenges:**

- Yang's variational learning
- Simple parameter-setting

**Supports:**

- Hyams' direct evidence account
- Duguine's implication of the article system
- Gradual, evidence-based learning

**Future Work:** Test these patterns with human participants and cross-linguistic data

# Chapter 2: Transfer Effects in Bilingual Acquisition

**Core Research Questions:**

- Are large language models capable of maintaining competence when trained multilingually?

# Chapter 2: Transfer Effects in Bilingual Acquisition

**Core Research Questions:**

- Are large language models capable of maintaining competence when trained multilingually?
- Does the order of language presentation impact learnability of subject drop?

# Chapter 2: Transfer Effects in Bilingual Acquisition

**Core Research Questions:**

- Are large language models capable of maintaining competence when trained multilingually?
- Does the order of language presentation impact learnability of subject drop?
- Do L1 transfer effects vary based on evidence strength?

# Chapter 2: Transfer Effects in Bilingual Acquisition

**Core Research Questions:**

- Are large language models capable of maintaining competence when trained multilingually?
- Does the order of language presentation impact learnability of subject drop?
- Do L1 transfer effects vary based on evidence strength?

**Approach:**

- Sequential bilingual training: English $\leftrightarrow$ Italian

# Chapter 2: Transfer Effects in Bilingual Acquisition

**Core Research Questions:**

- Are large language models capable of maintaining competence when trained multilingually?
- Does the order of language presentation impact learnability of subject drop?
- Do L1 transfer effects vary based on evidence strength?

**Approach:**

- Sequential bilingual training: English $\leftrightarrow$ Italian
- 2 x 2 design: L1 language x Training duration

# Chapter 2: Transfer Effects in Bilingual Acquisition

**Core Research Questions:**

- Are large language models capable of maintaining competence when trained multilingually?
- Does the order of language presentation impact learnability of subject drop?
- Do L1 transfer effects vary based on evidence strength?

**Approach:**

- Sequential bilingual training: English $\leftrightarrow$ Italian
- 2 x 2 design: L1 language x Training duration
- Evaluate null-subject competence in both languages

# Chapter 2: Four Bilingual Training Experiments

**Training Protocol: 90M Dataset**

*L1 training (1 or 2 epochs) → L2 training (5 epochs opposite language)*

**English First Models:**

1. **Experiment 0:** English 1 epoch → Italian 5 epochs
2. **Experiment 1:** English 2 epochs → Italian 5 epochs

**Italian First Models:**

3. **Experiment 2:** Italian 1 epoch → English 5 epochs
4. **Experiment 3:** Italian 2 epochs → English 5 epochs

# Chapter 2: Four Bilingual Training Experiments

**Training Protocol: 90M Dataset**

*L1 training (1 or 2 epochs) → L2 training (5 epochs opposite language)*

**English First Models:**

1. **Experiment 0:** English 1 epoch → Italian 5 epochs
2. **Experiment 1:** English 2 epochs → Italian 5 epochs

**Italian First Models:**

3. **Experiment 2:** Italian 1 epoch → English 5 epochs
4. **Experiment 3:** Italian 2 epochs → English 5 epochs

# Chapter 2: Four Bilingual Training Experiments

**Training Protocol: 90M Dataset**
*L1 training (1 or 2 epochs) → L2 training (5 epochs opposite language)*

**English First Models:**

1. **Experiment 0:** English 1 epoch → Italian 5 epochs
2. **Experiment 1:** English 2 epochs → Italian 5 epochs

**Italian First Models:**

3. **Experiment 2:** Italian 1 epoch → English 5 epochs
4. **Experiment 3:** Italian 2 epochs → English 5 epochs

# Chapter 2: Four Bilingual Training Experiments

**Training Protocol: 90M Dataset**
*L1 training (1 or 2 epochs) → L2 training (5 epochs opposite language)*

**English First Models:**

1. **Experiment 0:** English 1 epoch → Italian 5 epochs
2. **Experiment 1:** English 2 epochs → Italian 5 epochs

**Italian First Models:**

3. **Experiment 2:** Italian 1 epoch → English 5 epochs
4. **Experiment 3:** Italian 2 epochs → English 5 epochs

# Chapter 2: Four Bilingual Training Experiments

**Training Protocol: 90M Dataset**
*L1 training (1 or 2 epochs) → L2 training (5 epochs opposite language)*

**English First Models:**

1. **Experiment 0:** English 1 epoch → Italian 5 epochs
2. **Experiment 1:** English 2 epochs → Italian 5 epochs

**Italian First Models:**

3. **Experiment 2:** Italian 1 epoch → English 5 epochs
4. **Experiment 3:** Italian 2 epochs → English 5 epochs

**Key Question:** Is there asymmetric transfer/impairment between languages?

# Italian Stimuli: Pro-Drop Language

**Core Contrasts (Italian pro-drop)**

**Person/Number**: Anna ha finito il libro. ∅/%Lei pensa che il finale sia perfetto.
*Anna has finished the book. ∅/She thinks that the ending is perfect.*

**Control**: Maria ha convinto suo fratello ∅/*lui a partire presto.
*Maria convinced her brother ∅/*him to leave early.*

**Expletives**: ∅/*Sembra che gli studenti abbiano superato l'esame.
*∅/*It seems that the students have passed the exam.*

**Conjunctions**: Giovanni si è svegliato tardi e ∅/lui ha perso il treno.
*Giovanni woke up late and ∅/he missed the train.*

*Italian allows null subjects where English requires overt realization*

# The 'Default' Account and Predictions

**Theoretical Background**

Children initially default to null subjects, then learn overt subject requirements

**Asymmetry Prediction:**

- **Italian→English:** Pro-drop L1 should not differentially impact English L2 learning

# The 'Default' Account and Predictions

**Theoretical Background**

Children initially default to null subjects, then learn overt subject requirements

**Asymmetry Prediction:**

- **Italian→English:** Pro-drop L1 should not differentially impact English L2 learning
- **English→Italian:** Non-pro-drop L1 should impair Italian L2 learning

# The 'Default' Account and Predictions

**Theoretical Background**

Children initially default to null subjects, then learn overt subject requirements

**Asymmetry Prediction:**

- **Italian→English:** Pro-drop L1 should not differentially impact English L2 learning
- **English→Italian:** Non-pro-drop L1 should impair Italian L2 learning
- **Result:** Italian-first models outperform English-first models

# The 'Default' Account and Predictions

**Theoretical Background**

Children initially default to null subjects, then learn overt subject requirements

**Asymmetry Prediction:**

- **Italian→English:** Pro-drop L1 should not differentially impact English L2 learning
- **English→Italian:** Non-pro-drop L1 should impair Italian L2 learning
- **Result:** Italian-first models outperform English-first models

**Critical Test:** Do we see greater transfer benefits vs. interference costs?

# Training Methodology: Four Bilingual Models

## Training Protocol

Four models trained for 6-7 epochs total with systematic checkpointing

**Checkpoint Strategy:**

- **Log-step checkpoints** within first epoch: 1, 2, 4, 8, 16, 32...

# Training Methodology: Four Bilingual Models

**Training Protocol**

Four models trained for 6-7 epochs total with systematic checkpointing

**Checkpoint Strategy:**

- **Log-step checkpoints** within first epoch: 1, 2, 4, 8, 16, 32...
- **Regular checkpoints** throughout L1 and L2 training

# Training Methodology: Four Bilingual Models

## Training Protocol

Four models trained for 6-7 epochs total with systematic checkpointing

**Checkpoint Strategy:**

- **Log-step checkpoints** within first epoch: 1, 2, 4, 8, 16, 32...
- **Regular checkpoints** throughout L1 and L2 training
- **Final 5 epochs** (L2 phase) densely sampled for analysis

# Training Methodology: Four Bilingual Models

**Training Protocol**

Four models trained for 6-7 epochs total with systematic checkpointing

**Checkpoint Strategy:**

- **Log-step checkpoints** within first epoch: 1, 2, 4, 8, 16, 32...
- **Regular checkpoints** throughout L1 and L2 training
- **Final 5 epochs** (L2 phase) densely sampled for analysis

**Evaluation Across Training:**

- Models evaluated on both English and Italian stimuli

# Training Methodology: Four Bilingual Models

**Training Protocol**

Four models trained for 6-7 epochs total with systematic checkpointing

**Checkpoint Strategy:**

- **Log-step checkpoints** within first epoch: 1, 2, 4, 8, 16, 32...
- **Regular checkpoints** throughout L1 and L2 training
- **Final 5 epochs** (L2 phase) densely sampled for analysis

**Evaluation Across Training:**

- Models evaluated on both English and Italian stimuli
- Learning curves constructed for L1 and L2 acquisition

# Training Methodology: Four Bilingual Models

## Training Protocol

Four models trained for 6-7 epochs total with systematic checkpointing

**Checkpoint Strategy:**

- **Log-step checkpoints** within first epoch: 1, 2, 4, 8, 16, 32...
- **Regular checkpoints** throughout L1 and L2 training
- **Final 5 epochs** (L2 phase) densely sampled for analysis

**Evaluation Across Training:**

- Models evaluated on both English and Italian stimuli
- Learning curves constructed for L1 and L2 acquisition
- Age of Acquisition (AoA) and end-state performance measured

# Predictions: Asymmetric Transfer Effects

**Key Hypothesis:**

- **English background** should *impair* Italian L2 learning

# Predictions: Asymmetric Transfer Effects

**Key Hypothesis:**

- **English background** should *impair* Italian L2 learning
- **Italian background** should *not impair* English L2 learning

# Predictions: Asymmetric Transfer Effects

**Key Hypothesis:**

- **English background** should *impair* Italian L2 learning
- **Italian background** should *not impair* English L2 learning
- Asymmetry reflects different constraint directions

# Predictions: Asymmetric Transfer Effects

**Key Hypothesis:**

- **English background** should *impair* Italian L2 learning
- **Italian background** should *not impair* English L2 learning
- Asymmetry reflects different constraint directions

**Specific Predictions:**

**English L1 Models:**

- More English training $\rightarrow$ slower Italian L2
- Delayed AoA for Italian null subjects
- Stronger interference effects

**Italian L1 Models:**

- Amount of Italian training has minimal effect
- Consistent English L2 acquisition
- No systematic interference

# Chapter 3: Cross-Linguistic Priming of Null Subjects

**Research Questions:**

- Do large language models form cross-linguistic abstract representations?

# Chapter 3: Cross-Linguistic Priming of Null Subjects

**Research Questions:**

- Do large language models form cross-linguistic abstract representations?
- Can structural priming reveal competence beyond performance?

# Chapter 3: Cross-Linguistic Priming of Null Subjects

**Research Questions:**

- Do large language models form cross-linguistic abstract representations?
- Can structural priming reveal competence beyond performance?
- How do bilingual models represent the 'absence' of subjects?

# Chapter 3: Cross-Linguistic Priming of Null Subjects

**Research Questions:**

- Do large language models form cross-linguistic abstract representations?
- Can structural priming reveal competence beyond performance?
- How do bilingual models represent the 'absence' of subjects?

**Method:**

- Prime with null/overt subjects in Language A

# Chapter 3: Cross-Linguistic Priming of Null Subjects

**Research Questions:**

- Do large language models form cross-linguistic abstract representations?
- Can structural priming reveal competence beyond performance?
- How do bilingual models represent the 'absence' of subjects?

**Method:**

- Prime with null/overt subjects in Language A
- Measure surprisal on target verbs in Language B

# Chapter 3: Cross-Linguistic Priming of Null Subjects

**Research Questions:**

- Do large language models form cross-linguistic abstract representations?
- Can structural priming reveal competence beyond performance?
- How do bilingual models represent the 'absence' of subjects?

**Method:**

- Prime with null/overt subjects in Language A
- Measure surprisal on target verbs in Language B
- Compare cross-linguistic priming effects

# Cross-Linguistic Structural Priming

## Theoretical Motivation

If models have abstract syntactic representations, they should show cross-linguistic priming

**Key Innovation:**

- Priming reveals representations that performance tasks cannot

# Cross-Linguistic Structural Priming

## Theoretical Motivation

If models have abstract syntactic representations, they should show cross-linguistic priming

**Key Innovation:**

- Priming reveals representations that performance tasks cannot
- Cross-linguistic design eliminates lexical confounds

# Cross-Linguistic Structural Priming

## Theoretical Motivation

If models have abstract syntactic representations, they should show cross-linguistic priming

**Key Innovation:**

- Priming reveals representations that performance tasks cannot
- Cross-linguistic design eliminates lexical confounds
- Measures abstract 'absence' of subjects across languages

# Cross-Linguistic Structural Priming

## Theoretical Motivation

If models have abstract syntactic representations, they should show cross-linguistic priming

**Key Innovation:**

- Priming reveals representations that performance tasks cannot
- Cross-linguistic design eliminates lexical confounds
- Measures abstract 'absence' of subjects across languages

**Prediction:**

- Null subjects in Language A prime null preference in Language B

# Cross-Linguistic Structural Priming

## Theoretical Motivation

If models have abstract syntactic representations, they should show cross-linguistic priming

**Key Innovation:**

- Priming reveals representations that performance tasks cannot
- Cross-linguistic design eliminates lexical confounds
- Measures abstract 'absence' of subjects across languages

**Prediction:**

- Null subjects in Language A prime null preference in Language B
- Overt subjects in Language A prime overt preference in Language B

# Cross-Linguistic Structural Priming

**Theoretical Motivation**

If models have abstract syntactic representations, they should show cross-linguistic priming

**Key Innovation:**

- Priming reveals representations that performance tasks cannot
- Cross-linguistic design eliminates lexical confounds
- Measures abstract 'absence' of subjects across languages

**Prediction:**

- Null subjects in Language A prime null preference in Language B
- Overt subjects in Language A prime overt preference in Language B
- Effects independent of lexical overlap or surface similarity

# Structural Priming Methodology

## Surprisal-Based Measurement

Following Sinclair et al. (2022) and Momma et al. (2025)

**Calculation:**

- Prime: Null/overt subject in Language A

# Structural Priming Methodology

## Surprisal-Based Measurement

Following Sinclair et al. (2022) and Momma et al. (2025)

**Calculation:**

- Prime: Null/overt subject in Language A
- Target: Verb prediction in Language B

# Structural Priming Methodology

## Surprisal-Based Measurement

Following Sinclair et al. (2022) and Momma et al. (2025)

**Calculation:**

- Prime: Null/overt subject in Language A
- Target: Verb prediction in Language B
- Compare: -logProb(verb—null_prime) - -logProb(verb—overt_prime)

# Structural Priming Methodology

## Surprisal-Based Measurement

Following Sinclair et al. (2022) and Momma et al. (2025)

**Calculation:**

- Prime: Null/overt subject in Language A
- Target: Verb prediction in Language B
- Compare: -logProb(verb—null_prime) - -logProb(verb—overt_prime)

**Critical Insight:**

- No lexical overlap between prime and target

# Structural Priming Methodology

## Surprisal-Based Measurement

Following Sinclair et al. (2022) and Momma et al. (2025)

**Calculation:**

- Prime: Null/overt subject in Language A
- Target: Verb prediction in Language B
- Compare: -logProb(verb—null_prime) - -logProb(verb—overt_prime)

**Critical Insight:**

- No lexical overlap between prime and target
- No shared surface structure between languages

# Structural Priming Methodology

## Surprisal-Based Measurement

Following Sinclair et al. (2022) and Momma et al. (2025)

**Calculation:**

- Prime: Null/overt subject in Language A
- Target: Verb prediction in Language B
- Compare: -logProb(verb—null_prime) - -logProb(verb—overt_prime)

**Critical Insight:**

- No lexical overlap between prime and target
- No shared surface structure between languages
- Pure test of abstract syntactic representation

# From Parallel Stimuli to Cross-Linguistic Priming

## Leveraging Bilingual Evaluation Sets

Transform parallel English/Italian stimuli into priming experiments

**Standard Parallel Evaluation:**

- English: *Anna finished. She/*∅ thinks...*

# From Parallel Stimuli to Cross-Linguistic Priming

## Leveraging Bilingual Evaluation Sets

Transform parallel English/Italian stimuli into priming experiments

**Standard Parallel Evaluation:**

- English: *Anna finished. She/*∅ thinks...*
- Italian: *Anna ha finito. ∅/Lei pensa...*

# From Parallel Stimuli to Cross-Linguistic Priming

## Leveraging Bilingual Evaluation Sets

Transform parallel English/Italian stimuli into priming experiments

**Standard Parallel Evaluation:**

- English: *Anna finished. She/*∅ thinks...*
- Italian: *Anna ha finito. ∅/Lei pensa...*
- Separate within-language comparisons

# From Parallel Stimuli to Cross-Linguistic Priming

**Leveraging Bilingual Evaluation Sets**

Transform parallel English/Italian stimuli into priming experiments

**Standard Parallel Evaluation:**

- English: *Anna finished. She/*∅ thinks...*
- Italian: *Anna ha finito. ∅/Lei pensa...*
- Separate within-language comparisons

**Cross-Linguistic Priming Design:**

- **Cross-join** English and Italian sentences

# From Parallel Stimuli to Cross-Linguistic Priming

## Leveraging Bilingual Evaluation Sets

Transform parallel English/Italian stimuli into priming experiments

**Standard Parallel Evaluation:**

- English: *Anna finished. She/*∅ thinks...*
- Italian: *Anna ha finito. ∅/Lei pensa...*
- Separate within-language comparisons

**Cross-Linguistic Priming Design:**

- **Cross-join** English and Italian sentences
- Prime with Italian *null* → Target English verb

# From Parallel Stimuli to Cross-Linguistic Priming

## Leveraging Bilingual Evaluation Sets

Transform parallel English/Italian stimuli into priming experiments

**Standard Parallel Evaluation:**

- English: *Anna finished. She/*∅ thinks...*
- Italian: *Anna ha finito. ∅/Lei pensa...*
- Separate within-language comparisons

**Cross-Linguistic Priming Design:**

- **Cross-join** English and Italian sentences
- Prime with Italian *null* → Target English verb
- Prime with Italian *overt* → Target English verb

# From Parallel Stimuli to Cross-Linguistic Priming

## Leveraging Bilingual Evaluation Sets

Transform parallel English/Italian stimuli into priming experiments

**Standard Parallel Evaluation:**

- English: *Anna finished. She/*∅ thinks...*
- Italian: *Anna ha finito. ∅/Lei pensa...*
- Separate within-language comparisons

**Cross-Linguistic Priming Design:**

- **Cross-join** English and Italian sentences
- Prime with Italian *null* → Target English verb
- Prime with Italian *overt* → Target English verb
- Compare surprisal differences across prime conditions

# Cross-Linguistic Priming Matrix

**Experimental Design:**

| Prime Language | Target Language | Measurement |
|---|---|---|
| Italian null | English | Surprisal on English verb |
| Italian overt | English | Surprisal on English verb |
| English null | Italian | Surprisal on Italian verb |
| English overt | Italian | Surprisal on Italian verb |

**Key Advantages:**

- Syntactic priming should occur with **no lexical overlap** between prime and target

# Cross-Linguistic Priming Matrix

**Experimental Design:**

| Prime Language | Target Language | Measurement |
|---|---|---|
| Italian null | English | Surprisal on English verb |
| Italian overt | English | Surprisal on English verb |
| English null | Italian | Surprisal on Italian verb |
| English overt | Italian | Surprisal on Italian verb |

**Key Advantages:**

- Syntactic priming should occur with **no lexical overlap** between prime and target
- Simple stimuli construction, as we can use any parallel eval set to assess abstract representations as well as preferences

# Priming as a Window into Abstract Syntax

## What Cross-Linguistic Priming Reveals

Abstract syntactic knowledge that transcends surface linguistic differences

**Theoretical Predictions:**

- If the model's are not showing robust cross-linguistic priming effects, that indicates that they are developing generalizations in a shallow, concrete way.

# Priming as a Window into Abstract Syntax

## What Cross-Linguistic Priming Reveals

Abstract syntactic knowledge that transcends surface linguistic differences

**Theoretical Predictions:**

- If the model's are not showing robust cross-linguistic priming effects, that indicates that they are developing generalizations in a shallow, concrete way.
- If the model is forming strong syntactic generalizations, but these are not in the same direction as would be expected for human learners, that is problematic for an approach relying on a Platonic assumption

# Priming as a Window into Abstract Syntax

## What Cross-Linguistic Priming Reveals

Abstract syntactic knowledge that transcends surface linguistic differences

**Theoretical Predictions:**

- If the model's are not showing robust cross-linguistic priming effects, that indicates that they are developing generalizations in a shallow, concrete way.
- If the model is forming strong syntactic generalizations, but these are not in the same direction as would be expected for human learners, that is problematic for an approach relying on a Platonic assumption

# Priming as a Window into Abstract Syntax

## What Cross-Linguistic Priming Reveals

Abstract syntactic knowledge that transcends surface linguistic differences

**Theoretical Predictions:**

- If the model's are not showing robust cross-linguistic priming effects, that indicates that they are developing generalizations in a shallow, concrete way.
- If the model is forming strong syntactic generalizations, but these are not in the same direction as would be expected for human learners, that is problematic for an approach relying on a Platonic assumption

**Expected Results:**

- Italian null subjects prime English null preference

# Priming as a Window into Abstract Syntax

## What Cross-Linguistic Priming Reveals

Abstract syntactic knowledge that transcends surface linguistic differences

**Theoretical Predictions:**

- If the model's are not showing robust cross-linguistic priming effects, that indicates that they are developing generalizations in a shallow, concrete way.
- If the model is forming strong syntactic generalizations, but these are not in the same direction as would be expected for human learners, that is problematic for an approach relying on a Platonic assumption

**Expected Results:**

- Italian null subjects prime English null preference
- Italian overt subjects prime English overt preference

# Priming as a Window into Abstract Syntax

## What Cross-Linguistic Priming Reveals

Abstract syntactic knowledge that transcends surface linguistic differences

**Theoretical Predictions:**

- If the model's are not showing robust cross-linguistic priming effects, that indicates that they are developing generalizations in a shallow, concrete way.
- If the model is forming strong syntactic generalizations, but these are not in the same direction as would be expected for human learners, that is problematic for an approach relying on a Platonic assumption

**Expected Results:**

- Italian null subjects prime English null preference
- Italian overt subjects prime English overt preference
- and Vice Versa

# References I


Arnett, C., Chang, T. A., Michaelov, J. A., & Bergen, B. K. (2025). On the acquisition of shared grammatical representations in bilingual language models. *arXiv [cs.CL]*.

Bloom, L. (1970). *Language development*. The MIT Press, Massachusetts Institute of Technology.

Bloom, P. (1990). Subjectlees sentences in child language. *Linguistic Inquiry*, *21*(4), 491–504.

Bock, K. (1986). Syntactic persistence in language production. *Cogn. Psychol.*, *18*(3), 355–387.

Cao, R., & Yamins, D. (2021). Explanatory models in neuroscience: Part 1 – taking mechanistic abstraction seriously. *arXiv [q-bio.NC]*.

Chomsky, N. (1959). A review of B.f. skinner's verbal behavior. *Language*, *35*(1), 26–58.

# References II


Duguine, M. (2017). Reversing the approach to null subjects: A perspective from language acquisition. *Front. Psychol.*, *8*, 27.

Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. (2020). A systematic assessment of syntactic generalization in neural language models. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 1725–1744). Association for Computational Linguistics.

Hyams, N. (1986, August). *Language acquisition and the theory of parameters*. Kluwer Academic.

Hyams, N., & Wexler, K. (1993). On the grammatical basis of null subjects in child language. *Linguist. Inq.*, *24*(3), 421–459.

# References III


Michaelov, J., Arnett, C., Chang, T., & Bergen, B. (2023). Structural priming demonstrates abstract grammatical representations in multilingual language models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 3703–3720.

Momma, S., Slevc, L. R., & Phillips, C. (2018). Unaccusativity in sentence production. *Linguist. Inq.*, *49*(1), 181–194.

Rizzi, L. (1994). Early null subjects and root null subjects. In B. Lust (Ed.), *Language acquisition studies in generative grammar* (p. 151, Vol. 2). John Benjamins Publishing Company.

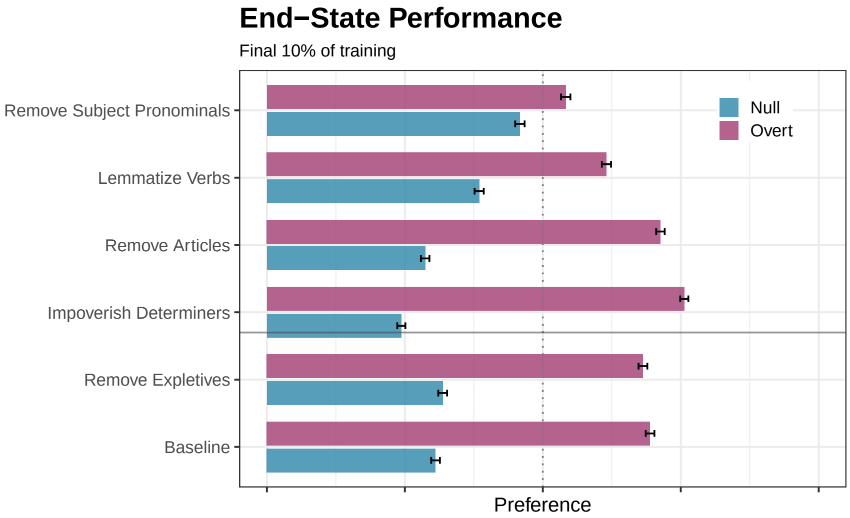
Warstadt, A., & Bowman, S. R. (2020). Can neural networks acquire a structural bias from raw linguistic data? *Proceedings of the 42nd Annual Conference of the Cognitive Science Society.*

Yang, C. D. (2003, February). *Knowledge and learning in natural language*. Oxford University Press.

# References IV

Yang, C. D. (2004). Universal grammar, statistics or both? *Trends Cogn. Sci.*, *8*(10), 451–456.

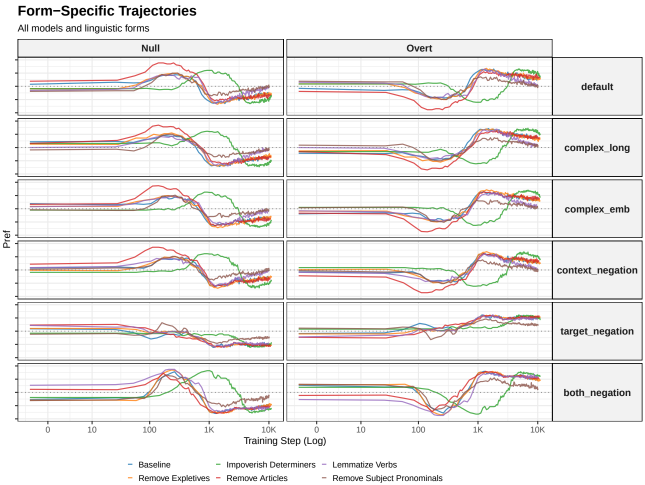
# End-State Performance Comparison



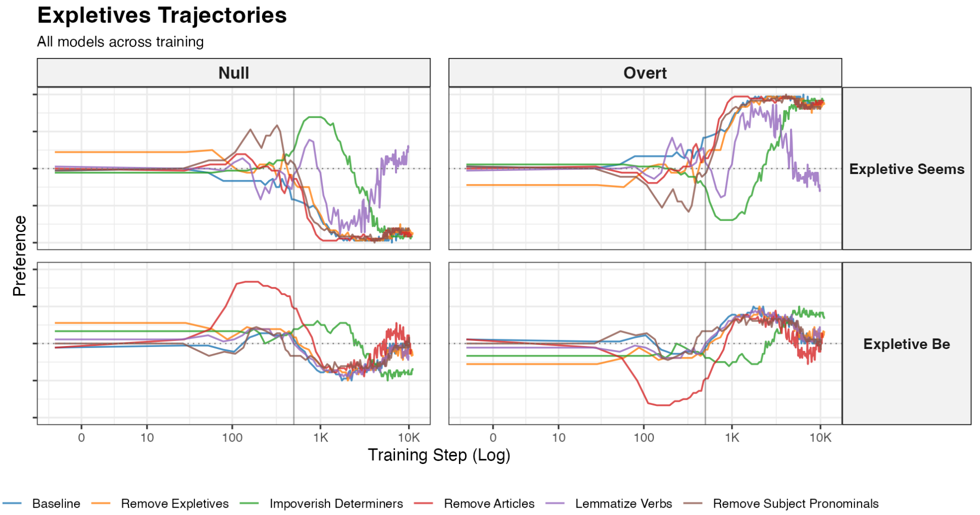
# Processing Forms vs Default Performance

| Form | Baseline | Rmv. Expletives | Impvr. Detrmn. | Rmv. Articles | Lemmatize Verbs | Rmv. Subject Pronominals |
|---|---|---|---|---|---|---|
| Complex Long | ✓ | | ✓ | | | |
| Complex Emb | | | ✓ | | | |
| Context Negation | | | | | | |
| Target Negation | ✓ | ✓ | | ✓ | ✓ | ✓ |
| Both Negation | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

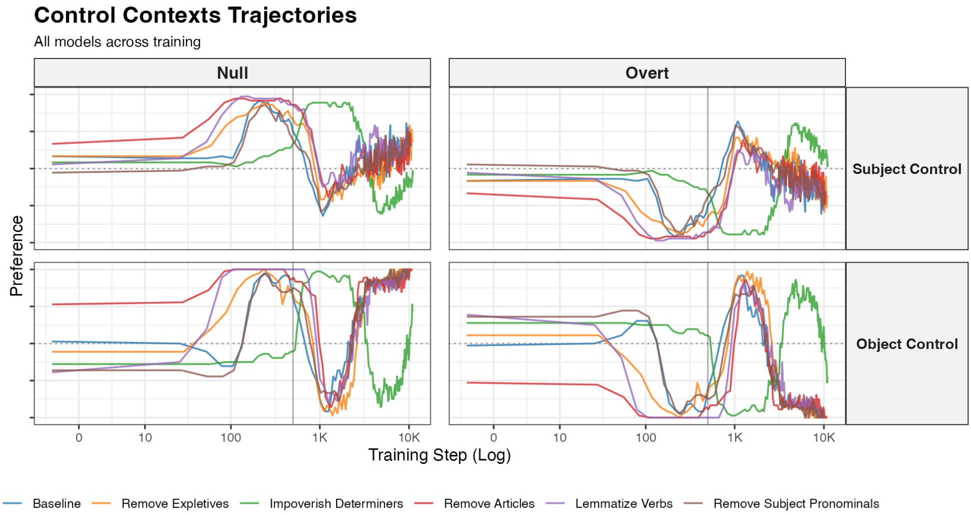
# Developmental Trajectories by Processing Manipulation



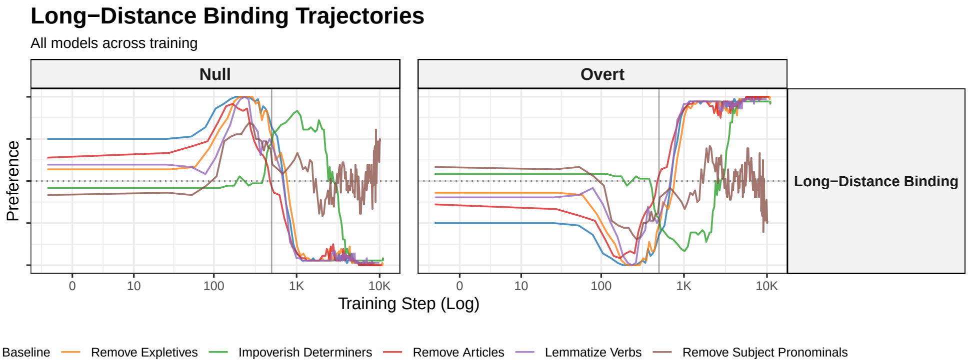
# Learning Trajectories: Expletive Constructions



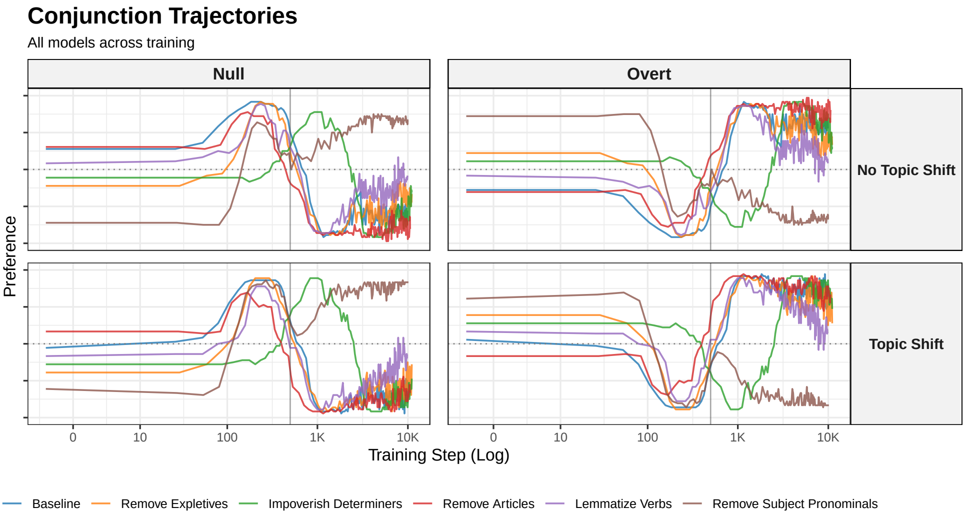
# Learning Trajectories: Control Constructions



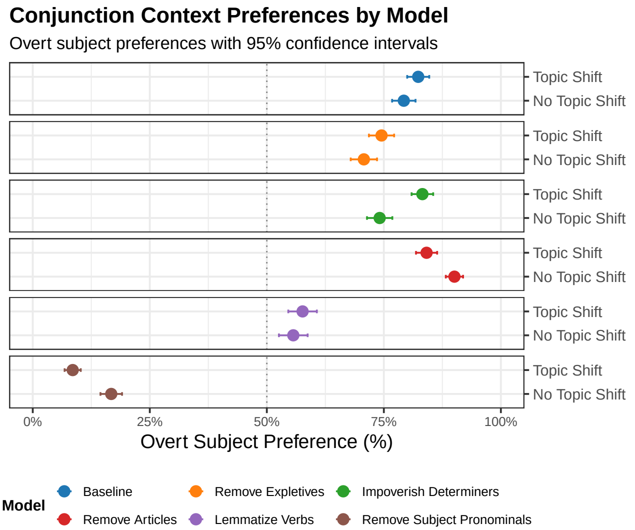
# Learning Trajectories: Long-Distance Binding



# Learning Trajectories: Conjunction Contexts



# Conjunction Context Performance by Model



# Model Preferences: Null vs Overt Subjects

| Model | Null Pref | Overt Pref |
|---|---|---|
| Baseline | 0.326 | 0.674 |
| Remove Expletives | 0.328 | 0.672 |
| Impoverish Determiners | 0.353 | 0.647 |
| Remove Articles | 0.336 | 0.664 |
| Lemmatize Verbs | 0.378 | 0.622 |
| Remove Subject Pronominals | 0.439 | 0.561 |

# Age of Acquisition by Model

| Model | AOA | CI |
|---|---|---|
| Lemmatize Verbs | 705.00 | [661, 749] |
| Baseline | 727.00 | [665, 792] |
| Rmv. Expletives | 767.00 | [709, 821] |
| Rmv. Subject Pronominals | 775.00 | [707, >5000] |
| Rmv. Articles | 808.00 | [759, 861] |
| Impvr. Detrmn. | 3400.00 | [3307, 3499] |