Filling in the Mechanisms: How do LMs Learn Filler-Gap Dependencies under Developmental Constraints?

Anonymous ACL submission

Abstract

For humans, filler-gap dependencies require a shared representation across different syntactic constructions. Although causal analyses suggest this may also be true for LLMs (Boguraev et al., 2025), it is still unclear if such a representation also exists for language models trained on developmentally feasible quantities of data. We applied Distributed Alignment Search (DAS, Geiger et al. (2024)) to LMs trained on varying amounts of data from the BabyLM challenge (Warstadt et al., 2023), to evaluate whether representations of filler-gap dependencies transfer between wh-questions and topicalization, which greatly vary in terms of their input frequency. Our results suggest shared, yet item-sensitive mechanisms may develop with limited training data. More importantly, LMs still require far more data than humans to learn comparable generalizations, highlighting the need for language-specific biases in models of language acquisition.

1 Introduction

A major question in language acquisition asks how learners can apply their linguistic knowledge to generalize to unseen information. The Poverty of the Stimulus argument (Chomsky, 1965) claims that children require specialized representations to learn language due to sparse input. Cognitive models of language acquisition have proposed and revised many claims about the nature of the representations humans could use to solve this learning problem (Yang, 2004; Perkins et al., 2022; Pearl, 2023). Language models (LMs), however, are trained on domain-general next-word prediction tasks, yet still posit some generalizations over syntactic structure (Futrell et al., 2019; Warstadt et al., 2020; Linzen and Baroni, 2021; Wilcox et al., 2024). These successes have led some researchers to question the need for language-specific representations altogether (Piantadosi, 2023; Futrell and Mahowald, 2025).

One particular case that evaluates this debate involves *filler-gap dependencies*. Filler-gap dependencies are a type of syntactic relation formed when a constituent (or the filler) is displaced from its canonical position (the gap) and interpreted in another position. These gaps can exist across various constructions including, but not limited to, wh-questions (*"What did the student make _?"*), relative clauses (*"the robot that the student made _"*), and topicalization (*"This robot, the student made _"*).

Filler-gap dependencies are a useful test case for evaluating representations necessary for language acquisition, as they not only require knowledge of hierarchical structure but also the recognition of an empty syntactic position. Many linguistic theories claim that despite superficial differences, these constructions may share an abstract underlying mechanism (Chomsky, 1977; Culicover et al., 1977; Gazdar, 1982; Kaplan and Bresnan, 1982; Postal, 1999). However, results from LMs have led others to argue that statistical patterns over language input may be sufficient (Wilcox et al., 2018, 2024).

These results have since been challenged, highlighting that LMs may not represent filler-gap dependencies in a human-like manner (Bhattacharya and van Schijndel, 2020; Lan et al., 2024; Howitt et al., 2024; Chang et al., 2025). However, most previous work relies largely on evaluating language model probabilities, rather than explicitly finding causal patterns in LM representations. Boguraev et al. (2025) do apply causal interpretability methods (Geiger et al., 2024; Mueller et al., 2025), and identify shared underlying structure across various types of filler-gap dependencies. However, even if Boguraev et al. (2025) show a shared representation of filler-gap dependencies may be learnable in principle, their results do not address whether inductive biases are needed for humans, as they evaluated LMs trained on data that matches nei-

ther the content nor the quality of data available to human learners (Wilcox et al., 2025).

We run causal interventions on LMs (Boguraev et al., 2025) trained on data from the BabyLM challenge (Warstadt et al., 2023), which reflects material that English-speaking children could be exposed to, up to 12 years of age. Our experiments show that LMs may posit a shared representation of filler-gap dependencies across high and low frequency constructions, but this representation is far less general than the ones proposed by linguists, showing systematic variability across items and constructions. Importantly, the model fails to posit this representation at the point it is expected to emerge in human learners (Perkins and Lidz, 2021). This result suggests that human learners need a combination of domain-general statistical learning mechanisms, along with language-specific inductive biases (Yang, 2004; Portelance and Jasbi, 2024).

2 Background

Filler-gap dependencies are shared across syntactic constructions that show different surface forms, even if they may serve different semantic and discourse functions (Schütze et al., 2015). Looking at the following sentences, (1) is a *wh-question*, while (3) is an example of *topicalization*.

(1) *Who* did the teacher like _ ?

(2) Did the teacher like?*

(3) *The author*, the teacher liked _ .

(4) The teacher liked.*

In (1) *who* is the object of *like*, and in (3), *The author* is the object of *liked*. These constituents, which are **fillers**, are fronted to form a dependency with the **gaps**, which are unpronounced but marked with _ for readability. Learning the generalization also involves recognizing where the dependency may *not* be valid, such as (2) and (4), which show that (1) and (3) are respectively ungrammatical without the fillers. Syntactic configurations called *islands* make extracting a filler ungrammatical (Chomsky, 1977).[1] This has made recognizing filler-gap licensing a particularly relevant test case when evaluating syntactic structure in language models.

[1]Refer to Wilcox et al. (2024), Howitt et al. (2024), and Chang et al. (2025) for further discussion relating to language models and island constraints.

2.1 LM Surprisal and Filler-Gap Dependencies

Many studies use LMs to compute the *surprisal*, or negative log probability, of a word given its context. Surprisal quantifies the effect of processing difficulty (Levy, 2008), and evaluating LMs' surprisals at particular points in a sentence effectively identifies which parts are expected to be more difficult to process (Futrell et al., 2019).[2] Work evaluating LMs on syntactic structure has often relied on comparing two minimal pairs of sentences, such as (1) vs. (2) and (3) vs. (4), where the ungrammatical version should typically have a higher surprisal than the grammatical version (Marvin and Linzen, 2018; Warstadt et al., 2020; Gauthier et al., 2020).

In English sentences with embedded wh-movement, LSTM language models have shown positive results in recognizing the presence and absence of fillers and gaps (Wilcox et al., 2018). These results have been extended to Transformer models and various island constraints (Wilcox et al., 2024).[3]

Ozaki et al. (2022) evaluated LSTMs across a range of other filler-gap constructions, including topicalization, and found that model performance for each construction is highly correlated with its frequency. They do not present evidence whether this generalization is shared *across* constructions.

To this end, other approaches involve providing LMs with additional training examples. Simulated priming studies on wh-movement (Bhattacharya and van Schijndel, 2020; Prasad et al., 2019) have shown some evidence for a shared representation of filler-gap dependency, but not constraints on the dependency. More recent studies have retrained LMs by augmenting their training data with positive examples of a dependency. Lan et al. (2024) show that augmentation improves LMs' performance on complicated filler-gap constructions (parasitic gaps and across-the-board movement). Extending this approach across constructions, Howitt et al. (2024) adopted their methodology and found that augmenting LSTMs' training data with instances of one construction (clefting and topicalization) failed to improve performance on detecting filler-gap licensing

[2]However, see Huang et al. (2024) for evidence that LM surprisal cannot reflect the quantitative effects of processing difficulty for particular types of syntactically complex sentences.

[3]Results are more mixed in Norwegian (Kobzeva et al., 2023) and Dutch (Suijkerbuijk et al., 2023), which have different filler-gap structures from English.

and islands for other constructions (wh-movement, clefting, topicalization, and tough-movement).

Although surprisal can effectively show whether LMs are able to assign probabilities appropriately for cases of filler-gap licensing, it may not necessarily directly reflect LMs' internal representations. Retraining LMs can help address whether they posit a shared representation for filler-gap dependencies, but results have been mixed.

2.2 Causal Interventions

In order to evaluate whether LM representations can indeed encode linguistic features of interest, several studies have adopted methods from mechanistic interpretability, using *causal interventions*. Broadly speaking, these approaches manipulate an LM's internal representations of individual items to find aspects of its representational space that are causally responsible for a desired feature (Wang et al., 2021; Lasri et al., 2022; Hao and Linzen, 2023; Kryvosheieva et al., 2025, among others, see Mueller et al. (2025) for a review). Arora et al. (2024) evaluated causal intervention methods on models from the Pythia series (Biderman et al., 2023), run on a large-scale suite of syntactic constructions (Gauthier et al., 2020), including wh-movement. Arora et al. (2024)'s best-performing method, Distributed Alignment Search (DAS) (Geiger et al., 2024), learns a rotation of the representation space to identify a direction; when intervened upon, this vector maximizes the probability of a counterfactual output label. We provide further description of this method in 4.3. Boguraev et al. (2025) applied DAS to Pythia-6.9B, finding evidence for a shared representation across seven types of filler-gap constructions. In their analysis, they identify an underlying "transfer network" that evaluates whether the representation from one filler-gap construction can successfully be implanted in another. More frequent constructions, such as wh-questions, act as "source" nodes in the network towards infrequent "sink" nodes such as topicalization. Despite unequal contributions to the shared representation, evidence for such a representation did exist. They also identify a "lexical boost"[4] effect based on whether the examples involved in the intervention share the same level of animacy, that is; stronger causal effects when both sentences in the intervention describe animate or inanimate subjects.

[4]This term is borrowed from psycholinguistic studies using syntactic priming, where human processing gets facilitated by similar sentence structures (Traxler et al., 2014).

2.3 Developmentally Plausible Data

Although Boguraev et al. (2025) shows evidence for an LM positing a shared mechanism for filler-gap dependencies, they relied on a Transformer model with billions of parameters, trained on a Web-scale corpus (Gao et al., 2020). These results make the case that filler-gap dependencies are learnable with a domain-general mechanism in principle, but have less to say about questions of *human* language acquisition (Wilcox et al., 2025). Human learners demonstrate sensitivity to filler-gap dependencies with exposure to exponentially less input than modern LMs (Gagliardi et al., 2016; Atkinson et al., 2018; Perkins and Lidz, 2021). The BabyLM Challenge (Warstadt et al., 2023) trains models on approximately 100 million tokens, the estimated linguistic input of a child around 12 years of age (Gilkerson et al., 2017). The data and models from BabyLM allow researchers to answer questions about human language acquisition that would be difficult to test experimentally, while providing learners with training data mimicking children's input. A prior evaluation of BabyLM models' performance on filler-gap licensing and island constraint violations found partial but incomplete acquisition, with performance varying substantially across island types (Chang et al., 2025). However, their findings are limited to wh-movement, which frequently occurs in children's input (Furrow et al., 1979), but models should also be evaluated across other constructions that appear less frequently to see if they posit a shared representation.

3 Research Questions and Hypotheses

Our study seeks to evaluate whether a causal cross-construction representation for filler-gap dependencies is learnable under a domain-general mechanism, provided that a learner has access to child-like amounts of input data. To this end, we aim to replicate findings from Boguraev et al. (2025) on a model trained on the corpus used in the BabyLM challenge, in service of extending results from linguistically informed mechanistic interpretability studies to questions about language development. We evaluate the model on both a high-frequency construction: wh-questions, compared to a low-frequency construction, topicalization. We propose the following three research questions:

- **RQ1:** Can a language model with relatively

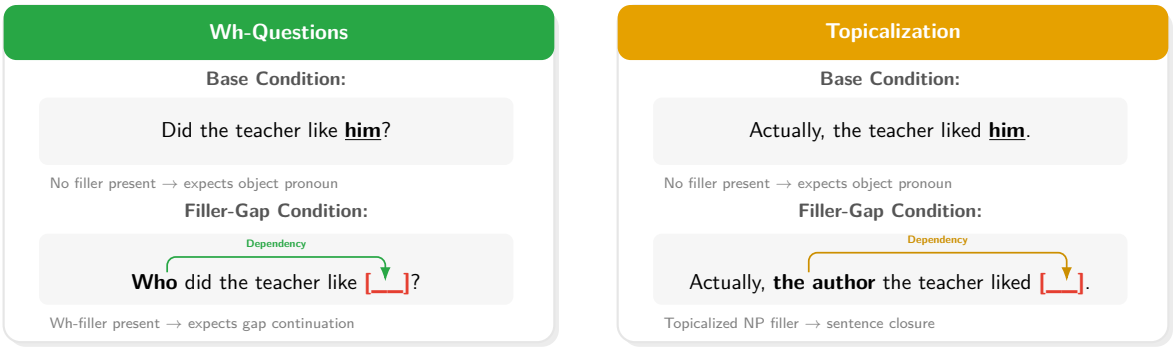

Figure 1: The diagram contrasts Wh-Questions (left, green) and Topicalization structures (right, orange).

few parameters and human-like training data still posit a causal representation of filler-gap dependencies? If so, when does this effect emerge?

- **RQ2:** Are representations localized within examples of the same construction?
- **RQ3:** Do representations transfer from low-frequency to high-frequency constructions?

We create three hypotheses based on these questions. First, the version of the model trained on all 100 million tokens in the BabyLM training corpus *should* learn an abstract filler-gap mechanism that is detectable and transferable by DAS from one construction to another, based on some of the positive results from Chang et al. (2025). However, due to the lack of language-specific inductive biases, we do not predict a strong causal effect before 10 million tokens. This is because the BabyLM checkpoint for 10 million tokens mimics the linguistic input of children from ages 2-5 years (Warstadt et al., 2023), while English-speaking children are able to recognize filler-gap dependencies around 18 months (Perkins and Lidz, 2021) (**H1: Misaligned Emergence Hypothesis**). Second, since LMs likely learn the filler-gap dependency in a piecemeal fashion (Ozaki et al., 2022; Howitt et al., 2024), we expect stronger causal effects for within-construction interventions compared to cross-construction interventions. We also hypothesize that transfer improves when both constructions share the same level of animacy, to replicate (Boguraev et al., 2025)'s findings for lexical boost effects. (**H2: Construction-Specificity Hypothesis**). Third, given the greater prevalence of wh-questions in language corpora, as opposed to topicalization, and existing work showing learning correlates with input frequency (Ozaki et al., 2022; Boguraev et al., 2025), we expect an asymmetrical and one-way transfer from high-frequency wh-questions to low-frequency Topicalization (**H3: Frequency Modulation Hypothesis**).

4 Methods

4.1 Model

We use the BabyLM-100M model (Warstadt et al., 2023). This model uses the GPT-2-small architecture (Radford et al., 2019) trained on the BabyLM Strict-100M corpus, consisting of a set of approximately 100 million words as training data designed to mimic the total linguistic input received by an English-speaking child until early adolescence (around 12 years of age) (Gilkerson et al., 2017; Warstadt et al., 2023).

The corpus consists of relevant data from the British National Corpus, CHILDES language acquisition database, Switchboard Dialog Act Corpus, subtitles from children's TV shows, and simplified Wikipedia articles (Charpentier et al., 2025).

Multiple checkpoints for the model across the training process were also released. We evaluate several checkpoints where the model received increasing amounts of input (1M–100M tokens): 10 checkpoints from 0–10M tokens and 9 checkpoints between 10M–100M tokens. Additionally, nine additional checkpoints (100M–1000M) are used in extended analysis in the appendix.[5]

4.1.1 Constructions

We use two filler-gap constructions that have different levels of frequency: matrix wh-questions (high frequency) and topicalization (low frequency), based on Ozaki et al. (2022). This combination

[5] If accepted, code and data for this paper will be publicly released.

allows us to systematically evaluate whether knowledge of high-frequency constructions can be transferred to less frequent ones that may barely be present in children's input.

Wh-Questions (High Frequency). We use single-clause matrix wh-questions, such as *"What did the doctor do ___?"* and *"What did the student read ___?"* Wh-questions are very common in natural language and exist in child-directed speech (Furrow et al., 1979). Prior work demonstrates that neural LMs readily acquire sensitivity to wh-dependencies in English (Wilcox et al., 2024). This makes Wh-questions a plausible construction for a high-frequency "source" in transfer experiments.

Topicalization (Low Frequency). We employ fronted object topicalization with an optional discourse marker, such as *"The student, the teacher liked ___"* or *"Actually, the book the author read ___"* Topicalization is extremely uncommon in natural language and almost absent from child-directed speech (Roland et al., 2007). Related works find that LMs fail to display the correct behavior for topicalization (Ozaki et al., 2022), even when augmented with examples in the training data (Howitt et al., 2024).[6] In addition, when analyzing transfer of the filler-gap mechanism across constructions, Boguraev et al. (2025) found topicalization had a low "out-degree" in their transfer network: despite receiving other constructions, topicalization rarely transfers to them. This makes topicalization a strong candidate for a low-frequency "sink" construction for our experiments.

We created minimal pairs following the original template and methodology of Boguraev et al. (2025). Each pair compares a base sentence with no filler-gap dependency with a filler-gap sentence containing a matrix wh-question or a topicalization setup. The model's expected next-token prediction differs based on the two predicted gaps (marked with _):

4.2 Materials

Replicating the experimental structure of Boguraev et al. (2025), we test sentences with both animate and inanimate fillers. Animate fillers use *who* (wh-questions) or NPs with perceived life or agency, such as *the author* (topicalization). In contrast, inanimate fillers use what (wh-questions) or nonliving and nonsentient NPs such as *the book*. This creates four dataset template variants: *wh_animate*, *wh_inanimate*, *topic_animate*, and *topic_inanimate*.

Different combinations of the following lexical items are used in respective templates: **Subject NPs** (50 animate nouns: *teacher, doctor, manager*, etc.), **Verbs** (30 transitive verbs: *like, admire, follow*, etc.), **Auxiliaries** for wh-questions (7 verbs: *did, will, could*, etc.), and **Licensing adverbs** for topicalization (25 adverbs: *Actually, Frankly, Surprisingly*, etc.). We verified all materials consist of one token for the model and occur within the BabyLM corpus.[7]

This yielded approximately 21,000 unique sentence pairs for wh-questions and 1,875,000 for topicalization per animacy condition. Topicalization randomly selects both the sentence-initial adverb and the topicalized filler phrase, increasing the amount of unique sentences. However, both pools remain substantially larger than 2000 pairs sampled for DAS training, maintaining sentence diversity for generalization.

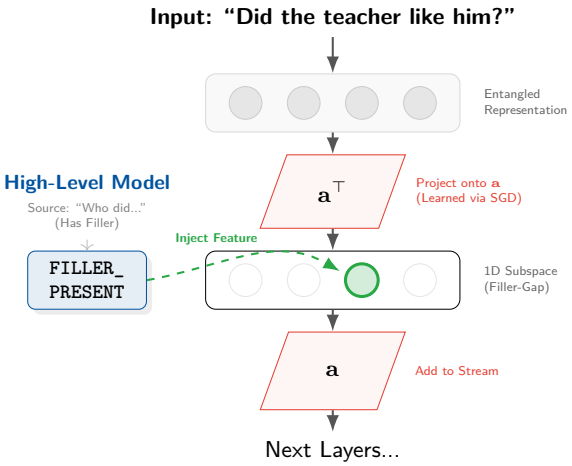

Figure 2: To create a DAS vector, we learn a direction $\mathbf{a}$ to align neural representations with the binary variable FILLER_PRESENT. We intervene by projecting the difference between the source and base representations onto $\mathbf{a}$ and injecting it into the base sentence.

4.3 Distributed Alignment Search (DAS)

Distributed Alignment Search (DAS) is a causal intervention method to test if a high-level concept aligns with the internal weights of a language model (Geiger et al., 2024). We can define a mini-

[6] Howitt et al. (2024) show that including the discourse marker leads to slight qualitative improvements, but the generalization is only learned in one direction, when the filler is present.

[7] See Nair and Resnik (2023), Giulianelli et al. (2024), and Oh and Schuler (2025) for discussion of tokenization and psycholinguistic applications.

mum causal model for a filler-gap dependency using a binary variable: FILLER_PRESENT $\in \{0, 1\}$. This variable causally influences gap expectations and the model's next-token predictions.

Given a *source* sentence with a filler (1) and a *base* sentence without a filler (1), we can intervene based on the internal representations of the base sentence by implanting the learned filler-gap DAS feature from the source, as seen in figure 2. A successful implementation should shift the prediction of the model from the base label (*him*) toward the source label (*?*). This would suggest that the filler-gap dependency is encoded at the intervention site (Wu et al., 2024).

4.3.1 Training

Following prior work, we use a 1-dimensional variant of DAS vector (Geiger et al., 2024; Arora et al., 2024; Boguraev et al., 2025). Given an embedding space with dimensionality d, for each layer ℓ and token position p, we learn a *direction vector* $\mathbf{a}_{\ell,p} \in \mathbb{R}^d$ that defines a one-dimensional subspace in which the filler-gap feature is encoded, between the model's representations of the base construction at ℓ and p ($\mathbf{h}_{\text{base},\ell,\mathbf{p}} \in \mathbb{R}^d$) and the source construction ($\mathbf{h}_{\text{source},\ell,\mathbf{p}} \in \mathbb{R}^d$). Once the vector $\mathbf{a}$ is learned, the intervention projects the difference between the source and base representations onto $\mathbf{a}$ and adds it to the base representation:

$$\tilde{\mathbf{h}} = \mathbf{h}_{\text{base}} + \mathbf{a}\mathbf{a}^\top(\mathbf{h}_{\text{source}} - \mathbf{h}_{\text{base}}) \qquad (1)$$

This intervention preserves the orthogonal dimensions of the base representation and only modifies the value along the learned feature direction $\mathbf{a}$. The direction is optimized to minimize cross-entropy loss between the counterfactual predictions of the model after intervention and the source sentence labels.

We use a batch size of 25, 80 training steps per layer-position combination, and a learning rate of 5×10^{-3} to train each DAS vector. We provide further information on hyperparameter selection in Appendix A. Due to the comparatively low number of layers in BabyLM's architecture, we test all 12 layers across 6 token positions aligned to template slots: prefix (position 0), filler (position 1), auxiliary/complementizer (position 2), article (position 3), subject NP (position 4), and verb (position 5).

4.3.2 Evaluation Metrics

We primarily use the **ODDS** metric to quantify the magnitude of the causal effect through measuring how much the intervention shifts log-probabilities toward the counterfactual outcome. This is given by the formula:

$$\begin{aligned}\textbf{ODDS} &= \log \frac{P(\text{base} \mid \text{clean})}{P(\text{source} \mid \text{clean})} \\ &+ \log \frac{P(\text{source} \mid \text{int})}{P(\text{base} \mid \text{int})}\end{aligned} \qquad (2)$$

In Equation (2), 'clean' refers to a standard forward pass without an intervention, while 'int' refers to a forward pass where the DAS intervention is applied.

A positive **ODDS** value suggests the intervention successfully shifts predictions toward the source; in addition, the higher the **ODDS** values, the stronger the causal effects.

Utilizing empirical findings from Arora et al. (2024) on the Pythia model family, we establish the following qualitative thresholds: values near 0 show little to no causal effect (comparable to random baselines), values in the 3–6 range demonstrate emerging to moderate causal structure (as seen in smaller models such as Pythia-14M), and values greater than 8 indicate strong causal mechanisms (as seen in larger models such as Pythia-6.9B).

We primarily report **MAX ODDS** , or the maximum **ODDS** value across all layers at a given position. The goal of DAS is to localize the feature to specific layers, so the maximum represents the layer-position combination with the most effective causal effect.

4.3.3 Experiments

We run two types of experiments: *localization* and *transfer*, to evaluate generalizations within and across construction types.

1. **Wh → Wh (within-construction localization)**: Train DAS on wh-questions, test on held-out wh-questions
2. **Topic → Topic (within-construction localization)**: Train DAS on topicalization, test on held-out topicalization
3. **Wh → Topic (forward transfer)**: Train DAS on wh-questions, test on topicalization
4. **Topic → Wh (backward transfer)**: Train DAS on topicalization, test on wh-questions

The localization experiments (Wh→Wh, Topic→Topic) quantify the extent to which

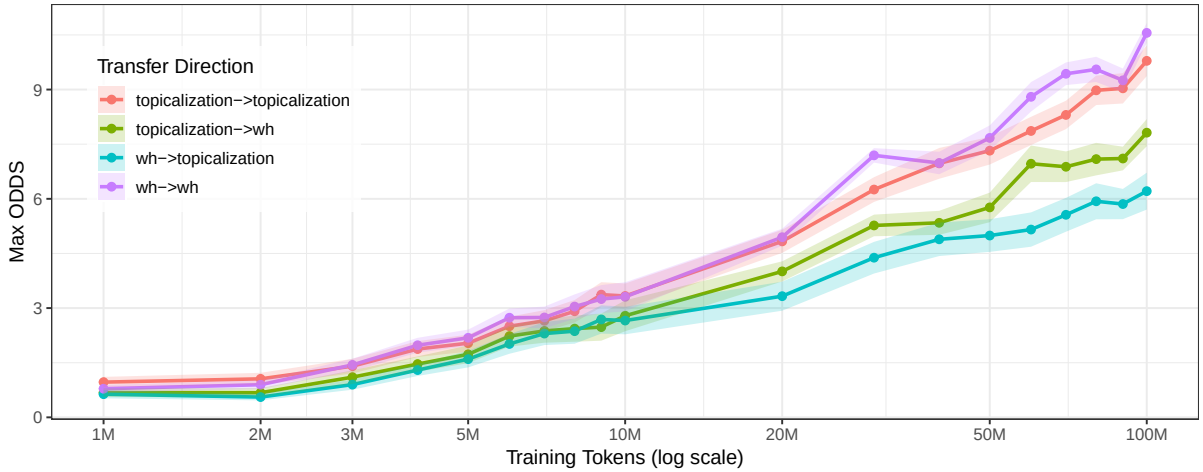

Figure 3: Developmental trajectory of all filler-gap mechanisms across training. Error bands show ±1 SE across a minimum of 6 seeds. All four conditions show a monotonic increase with training tokens.

the DAS can identify filler-gap representations within each construction type. Likewise, cross-construction transfers (Wh→Topic, Topic→Wh) test if representations of a given construction are generalizable across the filler-gap constructions. If the transfer is symmetric, **ODDS** retention should be similar in both directions. If frequency modulates transfer, we predict asymmetrical transfer towards the high-frequency (Wh → Topic) direction.

4.4 Statistical Analysis

We fit a linear model predicting **MAX ODDS** from number of training tokens, transfer direction, and animacy:

$$y = \beta_0 + \beta_t \mathbf{x}_{\text{tokens}} + \beta_d \mathbf{x}_{\text{dir}} + \beta_a \mathbf{x}_{\text{anim}} + \epsilon \quad (3)$$

Where y is **MAX ODDS** , and $\mathbf{x}$ represents the fixed effects for token count, transfer direction, and animacy. Post-hoc contrasts used estimated marginal means with Holm-Bonferroni correction for 171 pairwise comparisons (Lenth and Piaskowski, 2017). Cohen's d is used to measure effect sizes using the residual standard deviation.

5 Results

We present the results from 19 BabyLM checkpoints in the developmentally plausible range (1M–100M tokens) across multiple constructions (wh-questions, topicalization) and animacy conditions (animate, inanimate). All experiments were repeated on a minimum of 6 seeds to establish tighter bounds of confidence. The linear model achieved $R^2 = 0.53$, $F(22, 4537) = 235$, $p < .001$.

5.1 Developmental Trajectory

Figure 3 shows **MAX ODDS** across training on all four transfer directions. Filler-gap localization significantly increased with training duration ($F(18, 4537) = 264.3$, $p < .001$), from near zero at 1M tokens (**MAX ODDS** ≈ 0.8) to a robust effect by 100M (**MAX ODDS** ≈ 10.6 for Wh→Wh). Qualitatively, relatively stronger causal effects (**MAX ODDS** > 8) began emerging after around 50M tokens, while the effects are weaker (**MAX ODDS** ≈ 3) around 10M tokens.

5.2 Within-Construction Localization

Both constructions displayed successful within-construction localization, increasing with training tokens. At 100M tokens, Wh→Wh achieved **MAX ODDS** = 10.56 (SD = 2.11) and Topic→Topic achieved **MAX ODDS** = 9.79 (SD = 3.24). Compared to the results of the transfer experiments, within-construction localization effects consistently exceeded cross-construction transfer effects (mean difference = 1.33 **MAX ODDS** , t = 17.4, $p < .001$, d = 0.52), showing that high performance on one dependency type does not completely generalize to others.

5.3 Cross-Construction Transfer

Unlike the frequency-based prediction in Hypothesis 3, Topic→Wh transfer *exceeded* Wh→Topic transfer throughout training (difference = 0.57 **MAX ODDS** , t = -5.29, $p < .001$, d = -0.22). At

Contrast	Est.	SE	t	p	d
Within–Across	1.33	0.08	17.4	<.001	0.52
Wh→Topic–Topic→Wh	-0.57	0.11	-5.3	<.001	-0.22
Animate–Inanimate	0.68	0.08	8.6	<.001	0.26

Table 1: Post-hoc comparisons from linear model. Negative asymmetry results suggest Topic→Wh outperformed Wh→Topic.

the 100M checkpoint, Topic→Wh achieved **MAX ODDS** = 7.82 versus Wh→Topic's **MAX ODDS** = 6.21.

This transfer asymmetry could be due to wh-questions developing a more item-specific representation that transfers less effectively across construction types. Conversely, a more general mechanism may be reflected in topicalization, when it is learned, due to little to no presence in the input. We evaluate this behavior on the model when it is trained on up to 1 billion tokens in Appendix A.3.

5.4 Animacy Effects

We find further evidence of a "lexical boost" effect predicted by Boguraev et al. (2025). DAS transfer was significantly stronger when animacy of training and evaluation matched relative to animacy-mismatched conditions (difference = 0.67 **MAX ODDS** , $t = 4.86$, $p < .001$, $d = 0.28$).

6 Discussion and Conclusions

This study applies causal interventions to determine how a language model learns filler-gap dependencies when provided with developmentally realistic amounts of training data from the BabyLM corpus, determining if findings from larger LMs (Boguraev et al., 2025) apply in this setting. Using DAS, we evaluated BabyLM-100M's generalizations in four experimental conditions: localization within constructions and transfer across constructions, for two types of filler-gap dependencies, wh-questions and topicalization. Our results show that the model learns a shared representation for filler-gap dependencies, but still requires far more data than children would, and is still highly sensitive to variation across items.

Regarding **RQ1**, which asks both whether and when LMs learn a causal representation, we find evidence in favor of our misaligned emergence hypothesis. Although the BabyLM-100M model showed strong causal effects when trained on the full corpus, they failed to emerge when the model received human-like quantities of training data. The full corpus was comparable to the input available to English-speaking adolescents (up to around 12 years) (Warstadt et al., 2023), while children exhibit robust knowledge of filler-gap dependencies prior to two years of age (Gagliardi et al., 2016; Atkinson et al., 2018; Perkins and Lidz, 2021). In their description of the BabyLM corpus, Warstadt et al. (2023) report that the 10M checkpoint corresponds to children's linguistic knowledge between the ages of 2 and 5. If the model *was* learning with a human-like mechanism, we would expect strong causal effects prior to this checkpoint, yet only report weak effects, if any.

RQ2 asks whether the learned representations are specific to particular constructions. We found evidence for this construction-specificity hypothesis because the localization experiments consistently performed better than the transfer experiments. That is, generalization within examples of the same construction was far more effective than transferring representations across constructions. We also replicated the lexical boost effects when animacy gets matched during the intervention, suggesting this feature may transfer across items. Regarding the direction of transfer, as discussed in **RQ3**, we found improved performance generalizing from topicalization to wh-questions, which was the opposite of our predictions in the frequency modulation hypothesis. Future work can determine if this happens because LMs may learn more item-specific representations for wh-questions since they are more frequent, and deploy a more generalized mechanism to represent topicalization.

Overall, instead of positing a single, general representation of filler-gap constructions, LMs learn item and construction-specific representations. Future work should extend DAS to evaluate learning the filler-gap dependency in both directions and sensitivity to island constraints, across more diverse construction types.

When modeling human language acquisition, however, our results show LMs trained solely on next-word prediction are not sufficient. Instead, we emphasize the need to model learning with explicit inductive biases over structured hypothesis spaces. LMs may still play a role in this enterprise, through reflecting inductive biases architecturally (Murty et al., 2023) or specifying possible hypothesis spaces (Misra and Kim, 2024; Portelance and Jasbi, 2024). Thus, cognitive models of language acquisition should build on advances from statistical learning, while acknowledging their limits.

Limitations

This study only focused on English, limiting the generalizability of these results to other languages where filler-gap dependencies behave differently under LMs (Kobzeva et al., 2023; Suijkerbuijk et al., 2023). Additionally, the training corpus is based on text input alone, while children learn from spoken data, multimodal environments, and social interaction (Meylan et al., 2023; Vong et al., 2024). Since this study focused on extending Boguraev et al. (2025)'s results to a BabyLM-scale model, we did not evaluate whether it could recognize the absence of filler-gap dependencies, and for island constraints, which have been used in surprisal-based studies (Ozaki et al., 2022; Wilcox et al., 2024; Howitt et al., 2024; Chang et al., 2025). More complex materials would also be useful to ensure results are not associated with confounding factors like punctuation, since we extract representations from periods and question marks. Lastly, although DAS was the best performing method from Arora et al. (2024), measures like Boundless DAS operate over subspaces instead of single dimensions (Wu et al., 2023; Geiger et al., 2024), and could have yielded stronger causal effects.

Ethical Considerations

This work used publicly available data and models, which are described further in the original publications. We do not foresee any risks associated with this work, as we used the data for their intended purpose to study human language acquisition. Generative AI (GenAI) was used in this project. We used Antigravity[8] to design plots and refactor code, and Claude Opus 4.5 to refine paper writing for brevity. We never use GenAI for writing text from scratch in this paper. We take complete responsibility for any GenAI errors. By discussing GenAI usage here, we aim to encourage other researchers to do the same.

References

Aryaman Arora, Dan Jurafsky, and Christopher Potts. 2024. CausalGym: Benchmarking causal interpretability methods on linguistic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14638–14663, Bangkok, Thailand. Association for Computational Linguistics.

Emily Atkinson, Matthew W Wagers, Jeffrey Lidz, Colin Phillips, and Akira Omaki. 2018. Developing incrementality in filler-gap dependency processing. *Cognition*, 179:132–149.

Debasmita Bhattacharya and Marten van Schijndel. 2020. Filler-gaps that neural networks fail to generalize. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 486–495, Online. Association for Computational Linguistics.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Sasha Boguraev, Christopher Potts, and Kyle Mahowald. 2025. Causal interventions reveal shared structure across English filler–gap constructions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25021–25042, Suzhou, China. Association for Computational Linguistics.

Chi-Yun Chang, Xueyang Huang, Humaira Nasir, Shane Storks, Olawale Akingbade, and Huteng Dai. 2025. Mind the gap: How babylms learn filler-gap dependencies. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15060–15076.

Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Hu, Jaap Jumelet, Tal Linzen, Jing Liu, Aaron Mueller, Candace Ross, Raj Sanjay Shah, Alex Warstadt, Ethan Wilcox, and Adina Williams. 2025. Babylm turns 3: Call for papers for the 2025 babylm workshop.

Noam Chomsky. 1965. Aspects of the theory of syntax.

Noam Chomsky. 1977. On Wh-Movement. *Formal Syntax*, pages 71–132. Publisher: Academic Press.

Peter W Culicover, Thomas Wasow, and Adrian Akmajian. 1977. *Formal syntax*. Academic Press.

David Furrow, Katherine Nelson, and Helen Benedict. 1979. Mothers' speech to children and syntactic development: Some simple relationships. *Journal of child language*, 6(3):423–442.

Richard Futrell and Kyle Mahowald. 2025. How linguistics learned to stop worrying and love the language models. *arXiv preprint arXiv:2501.17047*.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of NAACL-HLT*, pages 32–42.

[8] https://antigravity.google/

Annie Gagliardi, Tara M Mease, and Jeffrey Lidz. 2016. Discontinuous development in the acquisition of filler-gap dependencies: Evidence from 15-and 20-month-olds. *Language Acquisition*, 23(3):234–260.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. Syntaxgym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76.

Gerald Gazdar. 1982. Phrase Structure Grammar. In Pauline Jacobson and Geoffrey K. Pullum, editors, *The Nature of Syntactic Representation*, Synthese Language Library, pages 131–186. Springer Netherlands, Dordrecht.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. 2024. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning*, pages 160–187. PMLR.

Jill Gilkerson, Jeffrey A Richards, Steven F Warren, Judith K Montgomery, Charles R Greenwood, D Kimbrough Oller, John HL Hansen, and Terrance D Paul. 2017. Mapping the early language environment using all-day recordings and automated analysis. *American journal of speech-language pathology*, 26(2):248–265.

Mario Giulianelli, Luca Malagutti, Juan Luis Gastaldi, Brian DuSell, Tim Vieira, and Ryan Cotterell. 2024. On the Proper Treatment of Tokenization in Psycholinguistics. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18556–18572, Miami, Florida, USA. Association for Computational Linguistics.

Sophie Hao and Tal Linzen. 2023. Verb conjugation in transformers is determined by linear encodings of subject number. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4531–4539.

Katherine Howitt, Sathvik Nair, Allison Dods, and Robert Melvin Hopkins. 2024. Generalizations across filler-gap dependencies in neural language models. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 269–279, Miami, FL, USA. Association for Computational Linguistics.

Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen. 2024. Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, 137:104510.

Ronald Kaplan and Joan Bresnan. 1982. Lexical-Functional Grammar: A Formal System for Grammatical Representation. In Joan Bresnan, editor, *The Mental Representation of Grammatical Relations*, pages 173–281. MIT Press.

Anastasia Kobzeva, Suhas Arehalli, Tal Linzen, and Dave Kush. 2023. Neural Networks Can Learn Patterns of Island-insensitivity in Norwegian. In *Proceedings of the Society for Computation in Linguistics 2023*, pages 175–185, Amherst, MA. Association for Computational Linguistics.

Daria Kryvosheieva, Andrea de Varda, Evelina Fedorenko, and Greta Tuckute. 2025. Different types of syntactic agreement recruit the same units within large language models.

Nur Lan, Emmanuel Chemla, and Roni Katzir. 2024. Large language models and the argument from the poverty of the stimulus. *Linguistic Inquiry*, pages 1–28.

Karim Lasri, Tiago Pimentel, Alessandro Lenci, Thierry Poibeau, and Ryan Cotterell. 2022. Probing for the usage of grammatical number. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8818–8831.

Russell V. Lenth and Julia Piaskowski. 2017. emmeans: Estimated marginal means, aka least-squares means.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7:195–212.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.

Stephan C Meylan, Ruthe Foushee, Nicole H Wong, Elika Bergelson, and Roger P Levy. 2023. How adults understand what young children say. *Nature human behaviour*, 7(12):2111–2125.

Kanishka Misra and Najoung Kim. 2024. Generating novel experimental hypotheses from language models: A case study on cross-dative generalization. *arXiv preprint arXiv:2408.05086*.

Aaron Mueller, Jannik Brinkmann, Millicent Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, et al. 2025. The quest for the right mediator: Surveying mechanistic interpretability for nlp through the lens of causal mediation analysis. *Computational Linguistics*, pages 1–48.

Shikhar Murty, Pratyusha Sharma, Jacob Andreas, and Christopher D Manning. 2023. Pushdown layers: Encoding recursive structure in transformer language

models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3233–3247.

Sathvik Nair and Philip Resnik. 2023. Words, subwords, and morphemes: What really matters in the surprisal-reading time relationship? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11251–11260.

Byung-Doh Oh and William Schuler. 2025. The impact of token granularity on the predictive power of language model surprisal. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4150–4162.

Satoru Ozaki, Dan Yurovsky, and Lori Levin. 2022. How well do lstm language models learn filler-gap dependencies? In *Proceedings of the Society for Computation in Linguistics 2022*, pages 76–88.

Lisa Pearl. 2023. Computational cognitive modeling for syntactic acquisition: Approaches that integrate information from multiple places. *Journal of Child Language*, 50(6):1353–1373.

Laurel Perkins, Naomi H Feldman, and Jeffrey Lidz. 2022. The power of ignoring: Filtering input for argument structure acquisition. *Cognitive Science*, 46(1):e13080.

Laurel Perkins and Jeffrey Lidz. 2021. Eighteen-month-old infants represent nonlocal syntactic dependencies. *Proceedings of the National Academy of Sciences*, 118(41):e2026469118.

Steven T Piantadosi. 2023. Modern language models refute chomsky's approach to language. *From field-work to linguistic theory: A tribute to Dan Everett*, 15:353–414.

Eva Portelance and Masoud Jasbi. 2024. The roles of neural networks in language acquisition. *Language and Linguistics Compass*, 18(6):e70001.

Paul M. Postal. 1999. *Three Investigations of Extraction*. The MIT Press.

Grusha Prasad, Marten van Schijndel, and Tal Linzen. 2019. Using priming to uncover the organization of syntactic representations in neural language models. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 66–76.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Douglas Roland, Frederic Dick, and Jeffrey L Elman. 2007. Frequency of basic english grammatical structures: A corpus analysis. *Journal of memory and language*, 57(3):348–379.

Carson T. Schütze, Jon Sprouse, and Ivano Caponigro. 2015. Challenges for a theory of islands: A broader perspective on ambridge, pine, and lieven. *Language*, 91(2):31–39.

Michelle Suijkerbuijk, Peter de Swart, and Stefan L Frank. 2023. The learnability of the wh-island constraint in dutch by a long short-term memory network. In *Proceedings of the Society for Computation in Linguistics 2023*, pages 321–331.

Matthew J Traxler, Kristen M Tooley, and Martin J Pickering. 2014. Syntactic priming during sentence comprehension: Evidence for the lexical boost. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(4):905.

Wai Keen Vong, Wentao Wang, A Emin Orhan, and Brenden M Lake. 2024. Grounded language acquisition through the eyes and ears of a single child. *Science*, 383(6682):504–511.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2021. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. In *The Eleventh International Conference on Learning Representations*.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, et al. 2023. Findings of the babylm challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM challenge at the 27th conference on computational natural language learning*, pages 1–34.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.

Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. 2024. Using computational models to test syntactic learnability. *Linguistic Inquiry*, 55(4):805–848.

Ethan Gotlieb Wilcox, Michael Y Hu, Aaron Mueller, Alex Warstadt, Leshem Choshen, Chengxu Zhuang, Adina Williams, Ryan Cotterell, and Tal Linzen. 2025. Bigger is not always better: The importance of human-scale language modeling for psycholinguistics. *Journal of Memory and Language*, 144:104650.

Zhengxuan Wu, Atticus Geiger, Aryaman Arora, Jing Huang, Zheng Wang, Noah Goodman, Christopher Manning, and Christopher Potts. 2024. pyvene: A library for understanding and improving PyTorch

models via interventions. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pages 158–165, Mexico City, Mexico. Association for Computational Linguistics.

Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah Goodman. 2023. Interpretability at scale: Identifying causal mechanisms in alpaca. *Advances in neural information processing systems*, 36:78205–78226.

Charles D Yang. 2004. Universal grammar, statistics or both? *Trends in cognitive sciences*, 8(10):451–456.

A Appendix

A.1 Hyperparameter Selection

Early experiments found that the default hyperparameters reported in the Boguraev et al. (2025) codebase (batch size 25 $\times$ 16 steps) resulted in undertrained DAS vectors. This is possible because the Pythia 1.4B model has far more parameters and was trained on far more data compared to BabyLM-100M.

We conducted a hyperparameter sweep to determine optimal DAS training parameters. Figure 4 and 5 show **MAX ODDS** across different batch sizes (8, 16, 25, 32) and training steps (40, 60, 80, 100, 120) for the Wh→Wh within-construction condition at the 100M checkpoint. Based on these results, we selected a batch size of 25 with 80 training steps (2000 total samples) for all experiments.

The learning rate was fixed at the default value of 5×10^{-3} used in Arora et al. (2024).

A.2 Animacy Figures

Supplementing the statistical tests for animacy effects in 5.4, we plot the increase in **MAX ODDS** across training split by lexical matching conditions. Figure 6 compares the causal performance when the intervention source and target base sentences share the same animacy status (Animate → Animate) versus differing animacy status (Animate → Inanimate). Results show a consistent gap between the two conditions, suggesting the learned representation may retain sensitivity to lexical features, such as animacy, through the pretraining process.

To separate animacy effects from construction-specific variance, the reported metrics for both Figure 6 and statistical testing are averaged across wh-question and topicalization constructions.

A.3 Beyond Developmental Constraints

We also present our results for the full 1 billion token training trajectory (100M–1000M tokens)[9] based on additional checkpoints released for the BabyLM model to better understand how filler-gap mechanisms continue to develop beyond developmentally plausible input levels. We see cross-construction generalizations plateau after 100M tokens, and performance begins to overlap for both sets of results. LMs could require more input to learn a representation specific to topicalization, since it rarely shows up in children's input. These results confirm our overall claims localization within examples of one construction is stronger than transfer across constructions, while showing inconclusive evidence for our third hypothesis regarding interactions between construction frequency and transfer.

[9]The model received 1000M tokens because it was trained on the 100M token dataset for 10 epochs.

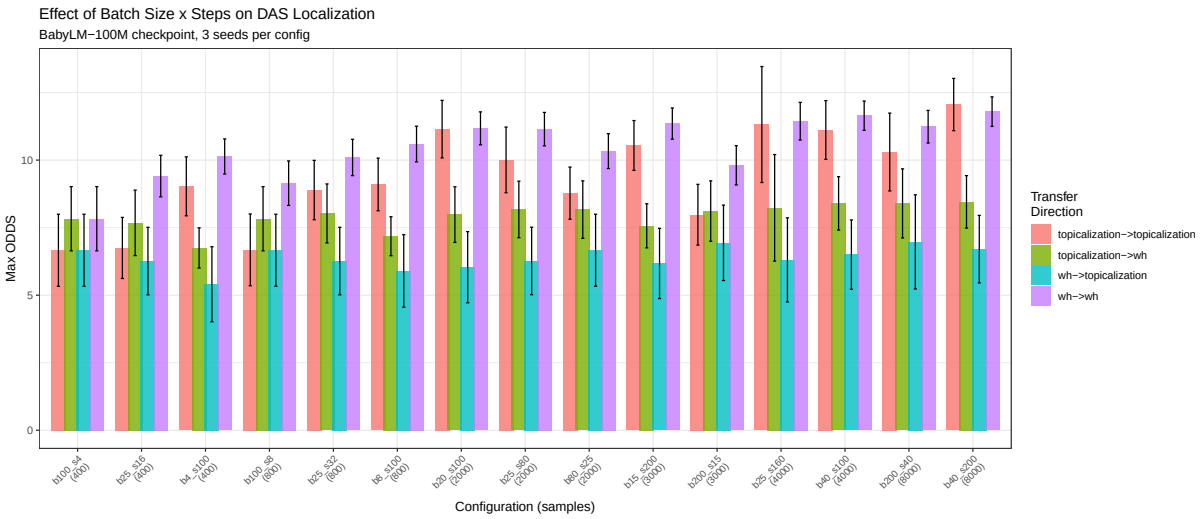

Figure 4: Hyperparameter sweep for DAS training. **MAX ODDS** increases with training samples and stabilizes around 2000 samples (batch size 25 $\times$ 80 steps).

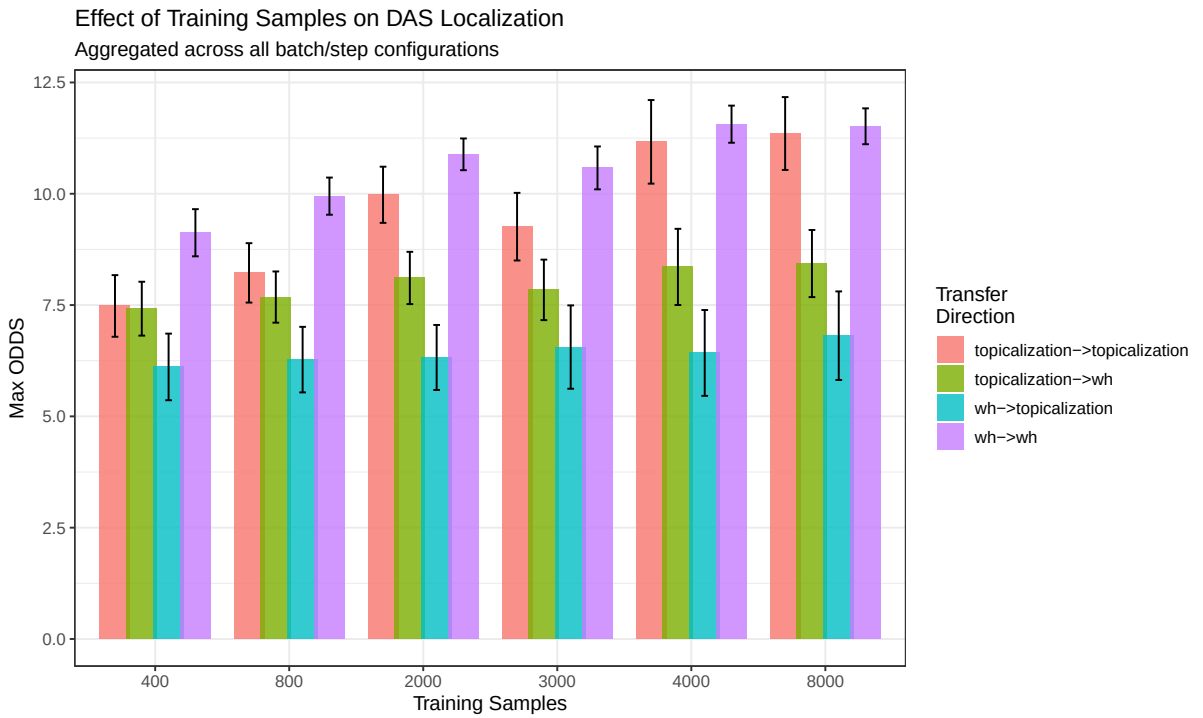

Figure 5: **MAX ODDS** as a function of total training samples, collapsed across batch sizes. Performance plateaus around 2000–2500 samples.

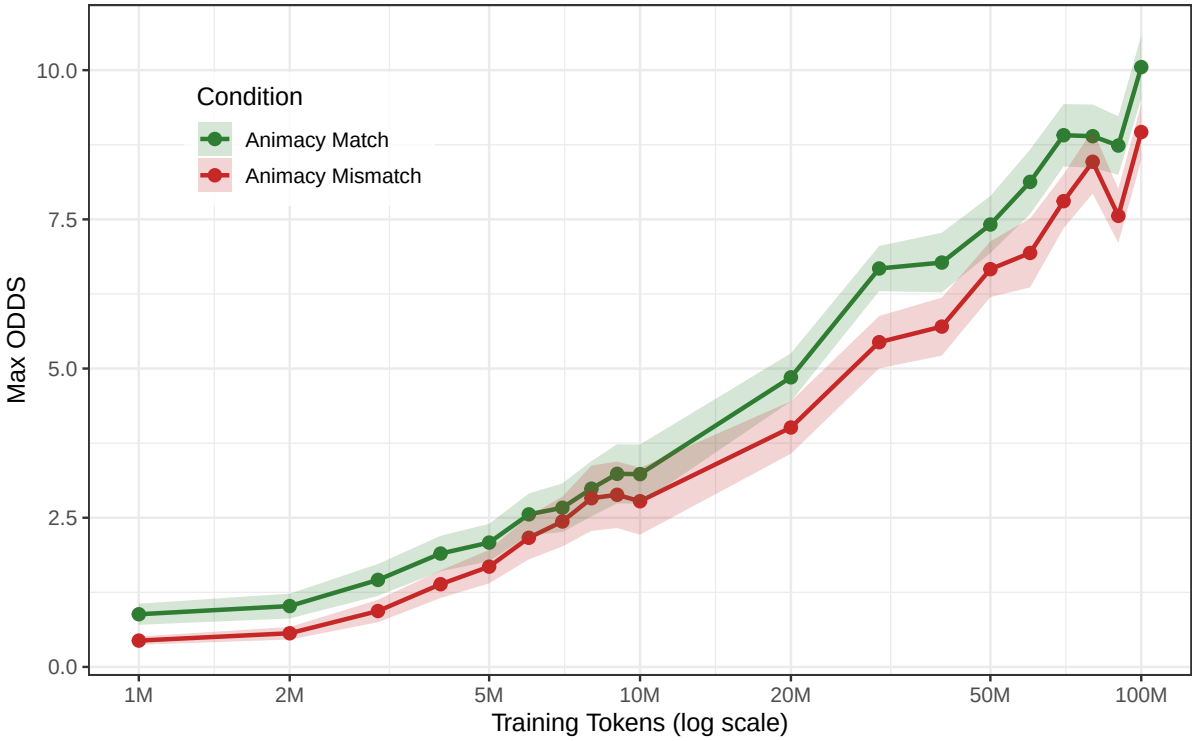

Figure 6: Developmental trajectory of lexical boost across training. Error bands show ±1 SE across a minimum of 2 seeds.

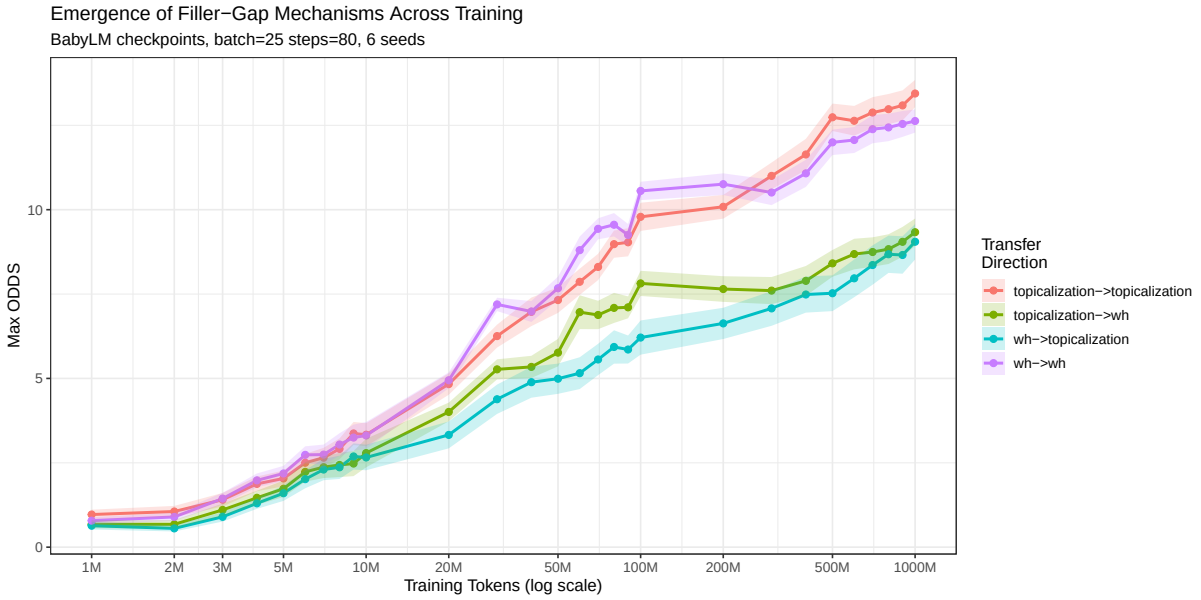

Figure 7: Full developmental trajectory from 1M to 1000M tokens. Filler-gap mechanisms continue to improve but begin to plateau around 500M–700M tokens.

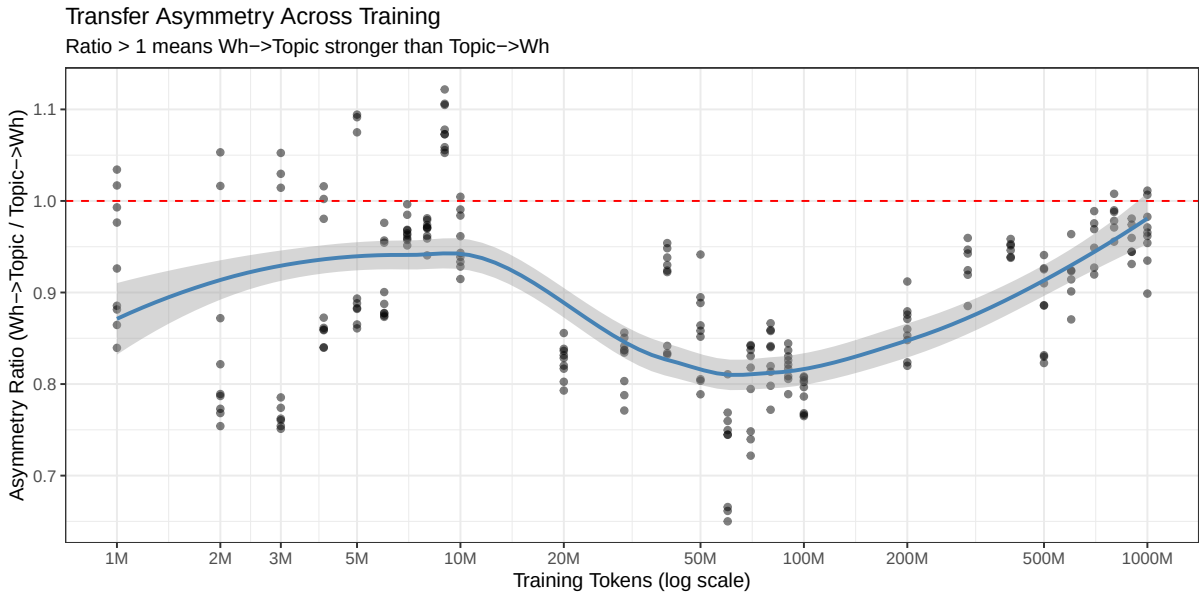

Figure 8: Transfer asymmetry across full training range. The asymmetry (Topic→Wh > Wh→Topic) persists and slightly increases at later checkpoints.