

Repetition Without Exclusivity: Scale Sensitivity of Referential Mechanisms in Child-Scale Language Models

Jon-Paul Cacioli

Independent Researcher

Melbourne, Australia

synthium@hotmail.com

Abstract

We present the first systematic evaluation of mutual exclusivity (ME), the bias to map novel words to novel referents, in text-only language models trained on child-directed speech. We operationalise ME as referential suppression and train 45 GPT-2 models (2.9M–33.5M parameters; 5–20 epochs on AO-CHILDES; 5 seeds) on a pre-registered ME battery. Across all 9 cells of the grid, models show robust anti-ME repetition priming, the opposite of ME, in 85–100% of items (all $p < 2.5 \times 10^{-13}$), attenuating with improved language modelling (Spearman $\rho = -0.533$) but never crossing zero. A positive control on pretrained GPT-2 (124M–1.5B) shows the same priming, attenuating with scale but never reversing into significant ME, confirming the battery can detect suppression when present. The dissociation also holds on a size-matched non-CDS corpus (WikiText). A context-dependence diagnostic further shows that apparent ME-like patterns with nonce tokens reflect embedding similarity, not referential disambiguation. We connect this to the grounded cognition literature, where ME does emerge in visually grounded models, and argue that referential grounding, or another mechanism instantiating competition among candidate referents, may be required for ME-like behaviour to emerge.

1 Introduction

The BabyLM paradigm trains language models on child-scale data and evaluates linguistic competence (Warstadt et al., 2023; Charpentier et al., 2025). Evaluations focus on syntax (BLiMP; Warstadt et al., 2020) and natural language understanding (GLUE; Wang et al., 2018). But referential pragmatics, how children use words to pick out objects in the world, remains untested. We address this gap using mutual exclusivity (ME) as the test case.

ME is the bias to map novel words to novel referents rather than objects that already have a name

(Markman and Wachtel, 1988). Intuitively, if a child sees a familiar dog and an unfamiliar object and hears “look at the dax!”, ME leads the child to assume “dax” refers to the unfamiliar object, because the dog already has a name. The bias is central to early word learning, developmentally graded and modulated by vocabulary size (Lewis et al., 2020), weaker in bilinguals (Byers-Heinlein and Werker, 2009), and has been studied computationally in grounded models (Gulordava et al., 2020; Ohmer et al., 2022; Nortje et al., 2024; Oneață et al., 2025; Thai et al., 2025). A key feature of all computational work showing positive ME results is that the model has referents (images, object representations, or symbolic referent sets) to which labels can be mapped.

BabyLM-class models are text-only. Can ME emerge from distributional learning on child-directed speech alone, without referential grounding? This is not a question about whether text-only models *should* exhibit ME; it is a question about whether distributional statistics are *sufficient* to induce it.

The question is sharpened by recognising that ME and a competing pressure act on the same measurement. When a noun is used as a label in a short discourse, two forces bear on its probability at a later completion position. Repetition priming (or givenness) raises it: recently mentioned words become more likely. Mutual exclusivity lowers it: once an object is named, a novel label should attach elsewhere, suppressing the just-used noun as a candidate for a new referent. These pressures point in opposite directions, so the sign of the net effect is diagnostic. Child-directed speech is dense with repetition (§3.2), which should teach priming; whether text-only learning *also* extracts the referential-contrast signal underlying ME is precisely what is at issue. We therefore measure whether labelling a noun raises its later probability (priming, anti-ME) or lowers it (suppression,

ME-consistent).

We test this across 45 autoregressive models spanning a $12\times$ parameter range and $3.8\times$ perplexity range, plus a masked language model (BabyBERTa; Huebner et al., 2021), and validate the measure with a positive control on pretrained GPT-2 up to 1.5B parameters. We find no evidence of ME at any scale tested; instead, models develop repetition priming that attenuates with improved language modelling but never reverses sign. We provide a mechanistic account grounded in the statistics of child-directed speech and introduce a context-dependence diagnostic that rules out artefactual ME from embedding similarity, a reusable methodological contribution for prompt-based pragmatic evaluation.

2 Background

2.1 ME in Developmental Psychology

ME was first characterised by Markman and Wachtel (1988) as a constraint on word learning: children assume that each object has only one label, so a novel word is mapped to the unnamed object rather than one that already has a name. The bias emerges around 17 months (Halberda, 2003), strengthens with vocabulary growth (Lewis et al., 2020), and is modulated by language experience: bilingual children show weaker ME, consistent with their routine experience of multiple labels for the same referent (Byers-Heinlein and Werker, 2009). An important distinction drawn by McMurray et al. (2012) separates referent selection (an in-the-moment disambiguation) from word learning (retention over time); our evaluation targets the former.

Bayesian accounts formalise ME as probabilistic inference over word–referent mappings (Xu and Tenenbaum, 2007), where the critical ingredient is the referent set: the learner must represent distinct objects, each requiring a label.

2.2 ME in Computational Models

Cross-situational neural models exhibit ME when lexical competition is built into learning or selection (Gulordava et al., 2020); pragmatic agents combining RSA reasoning with gradient learning also develop it (Ohmer et al., 2022), as do meta-learning systems for few-shot word learning (Gandhi and Lake, 2020).

The most relevant recent work comes from visually grounded speech (VGS) models. Nortje et al. (2024) showed that ME emerges from contrastive

learning on speech–image pairs, with stronger ME when models have more visual prior knowledge. Extending this to the bilingual case, Oneață et al. (2025) found that bilingual VGS models exhibit weaker ME, paralleling human developmental data. Thai et al. (2025) introduced MEBench and reported that even state-of-the-art vision-language models show only weak ME bias.

The key insight from this literature is that every positive ME result involves models with referents. ME has never been evaluated in text-only models trained on naturalistic child-directed speech.

2.3 Child-Scale Language Models

The BabyLM Challenge (Warstadt et al., 2023; Charpentier et al., 2025) trains models on 10M–100M-word corpora, approximating children’s input, and evaluates them on syntax (BLiMP) and NLU (GLUE). BabyBERTa (Huebner et al., 2021) trains RoBERTa-base on AO-CHILDES, American-English CDS (~ 4 M tokens). These evaluations target syntax and NLU; referential pragmatics is untested. We fill this gap.

3 Method

3.1 Pre-Registration and Transparency

We conducted a pilot study on a single autoregressive model and BabyBERTa (described in §3.5) that yielded three mechanistic findings. Based on these exploratory results, we pre-registered a confirmatory scale-sensitivity experiment on the Open Science Framework (OSF; <https://osf.io/zu7af>) before training any additional models. The pre-registration specifies four hypotheses (§3.6), statistical tests, significance thresholds, and the exact training grid. Train and evaluation scripts are publicly available at <https://github.com/synthiumjp/study2-me-eval>. Pilot findings are reported as exploratory with full transparency about the analysis path.

3.2 Corpus

AO-CHILDES (Huebner et al., 2021) contains 893,989 sentences (~ 4.07 M word tokens, 6.27M BPE tokens) of American-English child-directed speech from the CHILDES database (MacWhinney, 2000). The corpus is overwhelmingly composed of single-referent naming contexts; multi-referent disambiguation constructions (e.g., “the X or the Y”) are rare.

A property of CDS that is directly relevant to our findings is its high rate of noun repetition. Among the nine concrete nouns used in our evaluation battery (*ball, book, car, cup, hat, dog, cat, fish, bird*), the probability of a noun reappearing within the next 3 sentences averages 36.4% (range: 30.6% for *dog* to 42.2% for *ball*), and every test noun participates in hundreds to thousands of 10-sentence windows containing 3 or more mentions (e.g., *ball*: 4,236 windows; *cup*: 1,050). This repetition structure, characteristic of how caregivers scaffold naming (“that’s a ball. see the ball?. can you get the ball?”), provides a strong distributional signal favouring repetition priming. By contrast, sentences containing two or more of our target nouns account for only 524 of 893,989 sentences (0.06%), and the majority of these are compound expressions (*kitty cat*) rather than contrastive reference. The distributional environment of CDS thus strongly favours label repetition over label contrast.

3.3 Models

Masked LM (pilot only). BabyBERTa (Huebner et al., 2021): RoBERTa-base architecture, 8K vocabulary, trained on AO-CHILDES individual sentences. We use the published checkpoint.

Autoregressive models (confirmatory). GPT-2 architecture (Radford et al., 2019) trained from scratch on AO-CHILDES. We train a 3×3 grid of model sizes and training durations, with 5 random seeds per cell (45 models total):

The three sizes are Small (4 layers, 128 dim, 4 heads; 2.86M params), Medium (6 layers, 256 dim, 8 heads; 8.88M), and Large (8 layers, 512 dim, 8 heads; 33.50M).

All models share a BPE vocabulary of 8,020 tokens (8,000 merges + 20 nonce tokens added with space-prefix encoding). Vocabulary size was fixed at 8,000 learned BPE units as a practical child-scale setting, matching the BabyBERTa convention for AO-CHILDES, keeping the embedding matrix in proportion to the transformer, and ensuring the common concrete nouns used in evaluation are single tokens; it was not selected through vocabulary-size optimisation. Training durations: 5, 10, and 20 epochs. Optimizer: AdamW, learning rate 5×10^{-4} , cosine schedule, batch size 32, context length 128 tokens. Seeds 0–4. Total wall time: 31.4 hours on an AMD RX 7900 GRE (16 GB) via DirectML.

Parameter counts differ from pre-registration estimates ($\sim 2\text{M}$, $\sim 7\text{M}$, $\sim 30\text{M}$) due to the embedding

matrix dimension with the 8,020-token vocabulary. We report actual counts throughout.

3.4 ME Evaluation Battery

3.4.1 From Referent Selection to Probability Suppression

Our test does not reproduce the perceptual conditions of developmental ME experiments; it asks whether a text-only model exhibits the probability redistribution an exclusivity-like inference would require in a minimally contrastive discourse. The preamble introduces two referents, x and y ; the labelling sentence identifies one as x . If a system applies an exclusivity-like inference, that referent becomes less eligible as the target of a subsequent contrasting reference, so probability mass at the completion position should shift away from x toward y . Suppression of x is thus a textual signature of ME-like inference, not a recreation of the looking-time paradigm. Repetition priming predicts the opposite shift, since recently mentioned material becomes more probable. The test therefore pits two opposing pressures on the same completion position: referential exclusion and repetition-based continuation.

3.4.2 Familiar–Familiar Suppression Track (H1)

We test whether labelling one of two introduced referents as a familiar noun increases or decreases the probability of predicting that noun at a subsequent completion position.

Baseline: “there is a [FAM₁] and a [FAM₂]. this is a _____”

ME condition: “there is a [FAM₁] and a [FAM₂]. that is a [FAM₁]. this is a _____”

ME predicts $P(\text{FAM}_1 \mid \text{ME}) < P(\text{FAM}_1 \mid \text{baseline})$ (suppression) and $P(\text{FAM}_2 \mid \text{ME}) > P(\text{FAM}_2 \mid \text{baseline})$ (boost). Anti-ME predicts the opposite. We test 10 noun pairs in both directions (20 items per model seed). Log-probabilities are extracted via teacher-forced next-token prediction.

3.4.3 Context-Dependence Diagnostic (H2)

To distinguish genuine ME from embedding-level artefacts, we evaluate nonce-token predictions under five conditions:

- **full_context:** “there is a [FAM] and a [NONCE₁]. the [NONCE₂] is the _____” (standard ME design)
- **swap_context:** Same with FAM and NONCE₁ order reversed

- **nonce_only**: “there is a [NONCE₁]. the [NONCE₂] is the ____” (no familiar noun)
- **fam_only**: “there is a [FAM]. the [NONCE₂] is the ____” (no nonce in context)
- **no_preamble**: “the [NONCE₂] is the ____” (no context at all)

ME-consistent = $P(\text{NONCE}_2) > P(\text{FAM})$ at the completion position. 8 items $\times$ 10 embedding initialisation seeds per model. Nonce tokens use Strategy B embedding initialisation: each nonce is anchored to a random rare-noun embedding with 10% Gaussian noise. This is the most conservative available strategy: centroid-based approaches produce artificially clustered nonce embeddings that inflate similarity priming, while purely random initialisation places nonces far from the noun region, making probability comparisons uninformative. Strategy B produces nonce embeddings that are plausibly noun-like but not systematically similar to each other.

3.4.4 Dose–Response Track (H3)

We vary the number of labelling repetitions (0, 1, 2, 3) to test whether the priming advantage (the log-probability difference between the labelled and unlabelled noun) increases monotonically with dose. Evaluated on 5 noun pairs $\times$ 5 model seeds per cell (25 item–seed pairs per cell).

3.4.5 Scoring

All scoring uses log-probability of target tokens at the completion position, extracted via teacher-forced next-token prediction. All target tokens are verified to be single BPE tokens in the shared vocabulary.

3.5 Pilot Findings (Exploratory)

Three findings from the single-model pilot motivate the confirmatory design:

Pilot Finding 1: BabyBERTa cross-sentence insensitivity. BabyBERTa’s $\langle \text{mask} \rangle$ predictions are invariant to prior discourse context ($< 10^{-6}$ difference across all conditions). This is a consequence of the training regime (individual sentences with masked token prediction), not a limitation of masked LMs as an architecture class; masked LMs trained on concatenated multi-sentence inputs might show different sensitivity. ME evaluation via multi-sentence suppression design is therefore impossible for this specific checkpoint. We report this as a methodological finding.

Pilot Finding 2: AR repetition priming. On the pilot AR model (6L-256d, PPL 36.4), all 6/6 familiar–familiar pairs show anti-ME: labelling FAM₁ increases $P(\text{FAM}_1)$ and decreases $P(\text{FAM}_2)$. Priming ranges from +1.3 to +2.9 nats. This is consistent with CDS statistics: parents repeatedly name objects (“that’s a ball. see the ball. get the ball”).

Pilot Finding 3: Nonce ME is embedding similarity. The context-dependence diagnostic reveals that nonce_only (7/8 ME-consistent) exceeds full_context (2/8). Introducing any nonce in context primes nonce-region tokens; introducing a familiar noun suppresses them via frequency advantage. The disambiguation context contributes nothing.

3.6 Pre-Registered Hypotheses

Based on the pilot, we pre-registered four hypotheses:

H1 (confirmatory): Anti-ME rate $> 50\%$ in all 9 cells of the training grid. *Test*: One-tailed sign test per cell, $\alpha = 0.05$. *Criterion*: All 9 cells must reach significance.

H2 (confirmatory): nonce_only $\geq$ full_context in all 9 cells. *Test*: Wilcoxon signed-rank test (5 paired seeds), $\alpha = 0.05$. *Criterion*: All 9 cells.

H3 (confirmatory): $>50\%$ of item–seed pairs show strictly monotonic dose–response in all 9 cells. *Test*: Kendall’s τ for trend, bootstrap 95% CI on slope, $\alpha = 0.05$. *Criterion*: All 9 cells $>50\%$ monotonic.

H4 (confirmatory): Conjunction of H1: no cell shows ME-consistent suppression.

Each hypothesis is treated as a separate pre-registered question. No family-wise error rate correction is applied across hypotheses, as each addresses a distinct theoretical prediction with a distinct control comparison.

4 Results

We report confirmatory results followed by exploratory analyses. All confirmatory tests use the pre-registered statistical procedures and significance thresholds.

In summary: anti-ME repetition priming is significant in all 9 cells (H1 confirmed); the context-dependence diagnostic replicates in all 9 cells (H2 confirmed); dose–response priming shows significant positive trends in all 9 cells, with strict monotonicity in 8/9 (H3 partially confirmed); and no cell

Table 1: Perplexity (mean $\pm$ SD across 5 seeds).

	5 epochs	10 epochs	20 epochs
Small (2.9M)	57.8 $\pm$ 0.1	48.5 $\pm$ 0.1	42.0 $\pm$ 0.1
Medium (8.9M)	46.3 $\pm$ 0.1	36.6 $\pm$ 0.1	28.9 $\pm$ 0.0
Large (33.5M)	39.3 $\pm$ 0.1	27.0 $\pm$ 0.1	15.2 $\pm$ 0.1

Table 2: H1 results. Anti-ME rate = proportion of item-seed pairs where labelling FAM₁ increases $P(\text{FAM}_1)$. Priming in nats (natural log units; negative = anti-ME direction).

Cell	Anti-ME	Priming (nats)	Sign test p
Small, 5 ep	100/100	-1.14	7.89×10^{-31}
Small, 10 ep	100/100	-1.47	7.89×10^{-31}
Small, 20 ep	99/100	-1.49	7.97×10^{-29}
Medium, 5 ep	100/100	-1.46	7.89×10^{-31}
Medium, 10 ep	100/100	-1.43	7.89×10^{-31}
Medium, 20 ep	98/100	-0.81	3.98×10^{-27}
Large, 5 ep	100/100	-1.20	7.89×10^{-31}
Large, 10 ep	98/100	-1.06	3.98×10^{-27}
Large, 20 ep	85/100	-0.58	2.41×10^{-13}

shows ME-consistent suppression (H4 confirmed). Priming attenuates with improved language modelling ($\rho = -0.533$) but never crosses zero.

4.1 Training Grid

All 45 models converged without exclusions. Table 1 reports perplexity on a held-out AOCHILDES test set.

PPL spans a $3.8\times$ range (15.2 to 57.8). Seed variance is negligible (all SDs ≤ 0.1).

4.2 H1: Anti-ME at All Scales (Confirmed)

Anti-ME rates range from 85% to 100%, with all sign tests significant at $p < 2.5 \times 10^{-13}$ (Table 2; Figure 1). Not a single cell shows majority ME-consistent suppression.

Priming attenuates with improved language modelling but never crosses zero. **H1 confirmed.**

4.3 H1 Discourse Ablation

To test whether the anti-ME direction depends on the two-referent contrastive structure or reflects simpler repetition, we re-scored the labelled noun under four preamble variants (Medium, 10 epochs, 5 seeds), measuring priming against a no-labelling baseline. Priming was positive in every condition: full contrastive preamble (+1.43 nats), no preamble (+2.22), single referent, i.e. no competitor (+2.34), and non-labelling repetition, “I saw a x ” (+1.11). Across the three conditions that retain labelling, *all* 100 item-seed observations show priming and none shows suppression. The

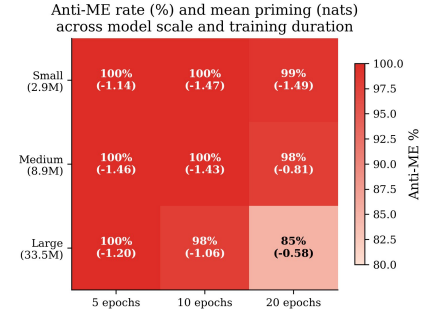


Figure 1: Anti-ME rate (%) and mean priming (nats) across model scale and training duration. All cells show strong anti-ME; priming magnitude decreases toward the bottom-right (larger, more-trained models).

Table 3: H2 results. Mean ME-consistent items out of 8.

Cell	nonce_only	full_context	fam_only
Small, 5 ep	5.0	2.2	~ 0
Small, 10 ep	5.7	2.7	~ 0
Small, 20 ep	5.3	2.8	~ 0
Medium, 5 ep	6.7	2.4	~ 0
Medium, 10 ep	6.9	3.3	~ 0
Medium, 20 ep	6.4	3.2	~ 0
Large, 5 ep	6.0	2.0	~ 0
Large, 10 ep	6.2	2.9	~ 0
Large, 20 ep	5.7	2.6	~ 0

contrastive preamble significantly *reduces* priming relative to the no-preamble and single-referent conditions (Wilcoxon $p < 10^{-17}$), a shift in the ME-predicted direction, but in no labelling condition does priming approach zero. The anti-ME direction is thus not an artefact of the contrastive setup: it survives removing the preamble and the competitor, and is only attenuated, never reversed, when the mention is stripped of its labelling function.

4.4 H2: Context-Dependence Diagnostic Replicates (Confirmed)

nonce_only exceeds full_context in all 9 cells (Table 3; Figure 2). All Wilcoxon tests reach $p = 0.0312$ (the minimum achievable for 5 paired observations all in the same direction).

Aggregated across all 45 models, the five-condition ordering is: nonce_only (6.0/8) > no_preamble (3.1/8) > full_context (2.7/8) > swap_context (2.3/8) > fam_only (0.7/8). This gradient reflects embedding-region priming: nonce context boosts nonce-region predictions; familiar-noun context suppresses them. The disambiguation preamble adds nothing.

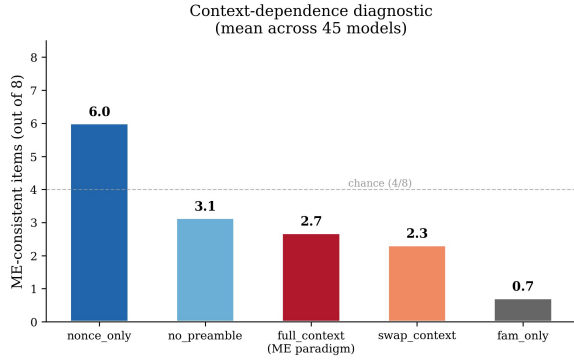


Figure 2: Context-dependence diagnostic (mean across 45 models). The nonce_only condition produces the highest apparent “ME,” despite containing no familiar noun and no ME paradigm. The full ME design (full_context) scores below chance. Dashed line: chance (4/8).

Table 4: H3 results.

Cell	Monotonic	Slope	Kendall τ	p
Small, 5 ep	25/25	+0.576	+0.391	2.07×10^{-7}
Small, 10 ep	25/25	+0.847	+0.480	1.98×10^{-10}
Small, 20 ep	22/25	+0.965	+0.421	2.39×10^{-8}
Medium, 5 ep	25/25	+1.080	+0.497	4.19×10^{-11}
Medium, 10 ep	24/25	+1.138	+0.451	2.13×10^{-9}
Medium, 20 ep	15/25	+0.896	+0.241	1.36×10^{-3}
Large, 5 ep	20/25	+1.159	+0.477	2.51×10^{-10}
Large, 10 ep	15/25	+0.976	+0.391	2.14×10^{-7}
Large, 20 ep	12/25	+1.067	+0.317	2.66×10^{-5}

H2 confirmed.

4.5 H3: Dose–Response (Partially Confirmed)

8/9 cells exceed the 50% monotonic threshold (the exception is Large at 20 epochs, 48%). All Kendall’s τ are positive and significant (all $p < 0.002$), and all bootstrap 95% CIs on slope exclude zero (Table 4; Figure 3).

Priming reliably increases with dose at all scales (all slopes positive, all trends significant), but strict item-level monotonicity weakens in the best language models. **H3 partially confirmed.**

4.6 H4: No ME at Any Scale (Confirmed)

Conjunction of H1. No cell shows ME-consistent suppression. **H4 confirmed.**

4.7 Exploratory: PPL–Priming Relationship

Across the 45 models, perplexity and priming magnitude are correlated (Spearman $\rho = -0.533$, $p = 0.0002$): better language models show weaker repetition priming. Decomposing this, model size contributes independently ($\rho = +0.434$, $p = 0.003$) as does training duration ($\rho = +0.335$,

$p = 0.024$). The relationship is approximately log-linear ($R^2 = 0.44$, vs. linear $R^2 = 0.36$). We do not extrapolate a zero-crossing perplexity from this fit, as any such estimate lies far below the observed range and would not be trustworthy; we note only that within the observed range priming attenuates but remains strictly positive (Figure 4). The GPT-2 positive control (§4.9) probes this trajectory directly at much larger scale and finds attenuation toward zero, not a reversal into ME.

4.8 Exploratory: Per-Item Analysis

Items vary in priming magnitude. The strongest priming pair is cup/book (mean -2.10 nats, 98% anti-ME across all 45 models). The weakest are dog/cat (-0.42 nats, 96% anti-ME) and cat/dog (-0.48 nats, 93% anti-ME). Both weak-priming pairs have natural synonyms in AO-CHILDES (dog/puppy, cat/kitty), potentially diluting the repetition signal. This parallel to bilingual attenuation of ME (Byers-Heinlein and Werker, 2009) is suggestive but requires targeted follow-up.

Per-noun corpus frequency does not predict per-noun priming magnitude (Spearman $\rho = 0.07$, $p = 0.86$, $n = 9$), nor does per-noun repetition rate ($\rho = -0.15$, $p = 0.70$). This suggests that while CDS repetition structure explains the *direction* of the effect (priming rather than suppression), per-item variation in magnitude is driven by other factors, likely properties of the unlabelled competitor noun or pair-specific co-occurrence patterns rather than the labelled noun’s frequency alone.

4.9 Positive Control: Pretrained GPT-2

A natural concern is whether the battery could detect ME-consistent suppression if it were present, or whether it is structurally biased toward anti-ME readings. As a positive control, we ran the identical H1 suppression battery on four off-the-shelf pretrained GPT-2 models (Radford et al., 2019) spanning 124M–1.5B parameters, an adult-scale web-text regime 4–45 $\times$ larger than our largest child-scale model. These models use the same architecture family but are trained on orders of magnitude more data of a very different kind.

The control reveals a clear scale-dependent trend (Table 5). Mean priming falls monotonically with scale: gpt2 (124M) shows strong anti-ME priming (+2.46 nats, 20/20 items), attenuating through gpt2-medium (+0.97) and gpt2-large (+0.69), until gpt2-xl (1.5B) shows no significant priming (mean -0.31 nats, 7/20 anti-ME, sign test $p = 0.94$). Two

Dose-response: priming increases with repetitions at all scales

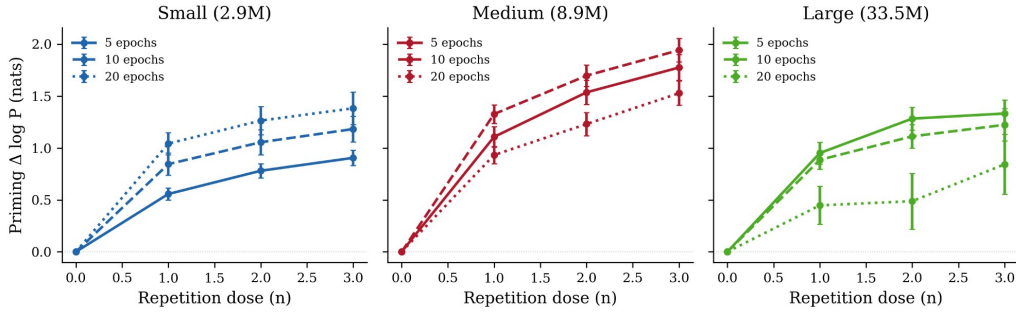


Figure 3: Dose-response curves by model size. Priming advantage (labelled minus unlabelled log-probability) increases with repetition dose at all scales. Error bars: SEM across item-seed pairs. Concavity in the 20-epoch curves (dotted) reflects declining marginal returns on repetition.

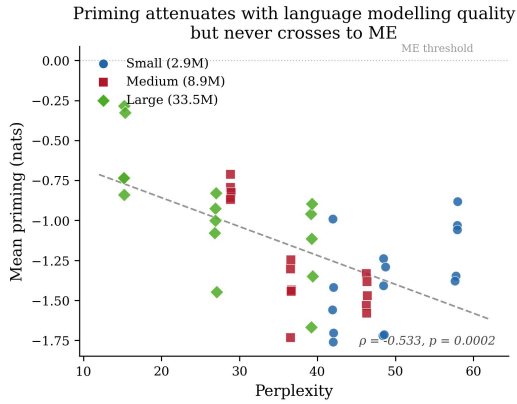


Figure 4: PPL vs. mean priming across 45 models. Each point is one model; shapes indicate size class. Dashed line: linear regression. Dotted line at 0: ME threshold. Priming attenuates with lower perplexity but remains strictly negative (anti-ME).

observations matter. First, the battery is plainly *capable* of moving off the anti-ME regime: the gpt2→gpt2-xl contrast is large and monotonic, so the child-scale null is a genuine measurement rather than a measurement floor. Second, gpt2-xl does not exhibit *ME* so much as an extinguished priming advantage: its slightly negative mean is driven almost entirely by two taxonomic-neighbour pairs (dog/cat -2.03 ; bird/fish -1.69), and excluding these four ordered pairs returns the mean to -0.08 nats, indistinguishable from zero. The residual suppression is best read as semantic competition between category co-hyponyms that a large model has learned co-occur, not as referential exclusivity. This directly instantiates the discourse-coherence confound that any prompt-based ME measure must contend with, and shows the diagnostic surfaces it rather than masking it.

Model	Params	Anti-ME	Mean (nats)
gpt2	124M	20/20	+2.46
gpt2-medium	355M	18/20	+0.97
gpt2-large	774M	18/20	+0.69
gpt2-xl	1.5B	7/20	-0.31

Table 5: Positive control on pretrained GPT-2. Anti-ME priming attenuates monotonically with scale and is extinguished, but does not reverse into significant ME, by 1.5B parameters. The gpt2-xl mean is dominated by two taxonomic pairs; excluding them yields -0.08 nats (see text).

Taken together with the child-scale grid, the priming advantage attenuates smoothly from +2.5 nats at 2.9M–124M parameters toward zero at 1.5B, but no model at any scale, from child-scale CDS training to adult-scale web text, crosses into significant ME-consistent suppression. This extends the paper’s central claim across nearly three orders of magnitude of scale while confirming the instrument’s validity.

4.10 Generalisation to a Non-CDS Corpus

To test whether the dissociation is specific to child-directed speech, we trained the Medium architecture (5 seeds, 10 epochs) on a size-matched slice of WikiText-103 (Merity et al., 2017), adult encyclopedic prose, truncated to the same token count as AO-CHILDES. This holds data volume fixed and varies register.

Anti-ME priming persists: 96/100 items across the five WikiText models (mean -0.72 nats), with no model showing ME-consistent suppression. The context-dependence diagnostic also replicates, with the same ordering (nonce_only highest at 5.0/8, fam_only lowest at 2.2/8). The dissociation is there-

fore not an artefact of CDS.

Priming is, however, significantly weaker on WikiText than on the matched CDS cell (-0.72 vs -1.43 nats; Mann–Whitney $U = 0$, $p = 0.008$; every CDS model primes more than every WikiText model). Two points bear on the interpretation. First, this attenuation is *not* a repetition-frequency effect: the nine battery nouns have nearly identical mean 3-sentence repetition rates in the two corpora (35.2% in WikiText vs 36.4% in AO-CHILDES). Rather, in encyclopedic text these nouns occur predominantly in non-object senses and frames (“book” as a published work, “cup” as a trophy, “car” as a railway carriage), weakening the object-naming pattern that drives priming. Second, the WikiText models are substantially weaker language models (PPL ≈ 198 vs ≈ 37 for the CDS Medium cell). This runs *opposite* to the within-CDS quality trend, where *better* language models prime *less* ($\rho = -0.533$, §4.7): if model quality drove the difference, the far weaker WikiText models should prime *more*, not less. The attenuation is thus attributable to register and lexical sense rather than to model quality. Critically, the qualitative dissociation (priming, never ME) is genre-invariant.

5 Discussion

5.1 Repetition Priming Without Mutual Exclusivity

The central finding is a dissociation: autoregressive language models trained on child-directed speech develop robust repetition priming but do not develop mutual exclusivity at any scale tested. Better language models attenuate their priming (§4.7), and the GPT-2 control (§4.9) extends this to 1.5B parameters and adult web text: priming is extinguished but never reverses into significant ME, with no sign of an approaching sign change at any scale.

5.2 What Text-Only Learning Lacks

Our findings converge with the grounded models literature. Nortje et al. (2024) demonstrated ME in visually grounded speech models, Oneață et al. (2025) showed weaker ME in bilingual VGS models paralleling human data, and Thai et al. (2025) found weak ME bias even in state-of-the-art VLMs. The pattern across these results and ours is consistent: positive ME demonstrations have involved models with referents, whereas over text alone the available mechanisms (embedding similarity and discourse-frequency effects) do not produce lexical

exclusivity. This aligns with the Bayesian word-learning framework (Xu and Tenenbaum, 2007), in which the critical ingredient is the referent set, and with evidence that ME in children is graded and experience-modulated (Lewis et al., 2020). The tested text-only models do not represent the competing referent set that standard accounts of ME assume. An alternative reading, however, is that ME follows not from grounding per se but from an explicit lexical-competition objective (Gulordava et al., 2020); we cannot fully separate the two, since grounded models typically also introduce competition. Both accounts locate the missing ingredient outside text-only next-token prediction, the level at which our claim operates, and our cross-study grounding evidence is correlational rather than a direct within-architecture manipulation.

5.3 The Diagnostic and the Scope of the Claim

Beyond the ME-specific findings, the context-dependence diagnostic has broader utility: if an apparent pragmatic effect survives removal of the pragmatically relevant context, it is not pragmatic. Stripping the disambiguation preamble while retaining nonce context *increased* the apparent ME signal, revealing embedding similarity as the driver, a test applicable to any prompt-based claim of pragmatic competence in LMs. The H1 ablation (§4.3) applies the same logic to the suppression track, confirming the anti-ME direction is repetition-driven rather than a response to the referential contrast.

We do not claim that distributional learning is fundamentally inadequate for referential pragmatics, nor that ME requires innate machinery. The claim is narrower: input modality constrains what referential mechanisms can be learned, and under child-scale text-only input the distributional signal favours repetition over exclusivity. Whether different architectures, objectives, or larger scales could overcome this is open; our data neither support nor rule it out, consistent with the BabyLM multimodal track (Warstadt et al., 2023; Charpentier et al., 2025).

6 Conclusion

Child-scale language models trained on CDS develop repetition-based reference tracking rather than mutual exclusivity, a dissociation robust across scale (to 1.5B parameters) and genre. The tested text-only models do not explicitly represent the competing referent set that standard accounts

of ME assume, while prior positive demonstrations involve explicit referents or objectives that impose lexical competition. Whether grounding, referent-set representation, or lexical competition is the operative ingredient is the question these results motivate.

Limitations

Several constraints bound our conclusions. Both corpora are English (AO-CHILDES and the Wiki-Text replication of §4.10); our largest from-scratch model is 33.5M parameters, and though the GPT-2 control reaches 1.5B without reversal, we cannot rule out change at larger scales. We hold training-data volume fixed at the developmentally plausible $\sim 4\text{M}$ -token scale rather than varying it; the GPT-2 control, trained on orders of magnitude more data, shows the dissociation persists far beyond this regime, but we do not isolate data quantity as a factor within our controlled grid. We did not optimise architecture for AO-CHILDES, though the effect’s consistency across a $12\times$ parameter range argues against architecture-specificity. Most importantly, a direct manipulation of grounding within one architecture, rather than the cross-study comparison we rely on, remains the key next step.

References

- Krista Byers-Heinlein and Janet F. Werker. 2009. Monolingual, bilingual, trilingual: Infants’ language experience influences the development of a word-learning heuristic. *Developmental Science*, 12(5):815–823.
- Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Y. Hu, Jing Liu, Jaap Jumelet, Tal Linzen, Aaron Mueller, Candace Ross, Raj Sanjay Shah, Alex Warstadt, Ethan Gotlieb Wilcox, and Adina Williams. 2025. Findings of the third BabyLM challenge: Accelerating language modeling research with cognitively plausible data. In *Proceedings of the First BabyLM Workshop*, pages 399–420.
- Kanishk Gandhi and Brenden M. Lake. 2020. Mutual exclusivity as a challenge for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 33.
- Kristina Gulordava, Thomas Brochhagen, and Gemma Boleda. 2020. Deep daxes: Mutual exclusivity arises through both learning biases and pragmatic strategies in neural networks. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, pages 2089–2095.
- Justin Halberda. 2003. The development of a word-learning strategy. *Cognition*, 87(1):B23–B34.
- Philip A. Huebner, Elior Sulem, Cynthia Fisher, and Dan Roth. 2021. BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646.
- Molly Lewis, Veronica Cristiano, Brenden M. Lake, Tammy Kwan, and Michael C. Frank. 2020. The role of developmental change and linguistic experience in the mutual exclusivity effect. *Cognition*, 198:104191.
- Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*, 3rd edition. Erlbaum.
- Ellen M. Markman and Gwyn F. Wachtel. 1988. Children’s use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, 20(2):121–157.
- Bob McMurray, Jessica S. Horst, and Larissa K. Samuelson. 2012. Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review*, 119(4):831–877.
- Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *International Conference on Learning Representations*.
- Leanne Nortje, Dan Oneață, Yevgen Matusevych, and Herman Kamper. 2024. Visually grounded speech models have a mutual exclusivity bias. *Transactions of the Association for Computational Linguistics*, 12:755–770.
- Xenia Ohmer, Michael Franke, and Peter König. 2022. Mutual exclusivity in pragmatic agents. *Cognitive Science*, 46(1):e13069.
- Dan Oneață, Leanne Nortje, Yevgen Matusevych, and Herman Kamper. 2025. [The mutual exclusivity bias of bilingual visually grounded speech models](#). In *Proceedings of Interspeech 2025*, pages 5043–5047.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.
- Anh Thai, Stefan Stojanov, Zixuan Huang, Bikram Boote, and James M. Rehg. 2025. MEBench: A novel benchmark for understanding mutual exclusivity bias in vision-language models. *arXiv preprint arXiv:2505.20122*.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP*.
- Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 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.

Fei Xu and Joshua B. Tenenbaum. 2007. Word learning as Bayesian inference. *Psychological Review*, 114(2):245–272.