

Do Language Models Know What *Not* to Say? Causal Evidence for Statistical Preemption in LLMs

Dongxin Guo

The University of Hong Kong
Hong Kong, China
bettyguo@connect.hku.hk

Jikun Wu

Stellaris AI Limited
Hong Kong, China
hk950014@connect.hku.hk

Siu Ming Yiu

The University of Hong Kong
Hong Kong, China
smyiu@cs.hku.hk

Abstract

How do learners acquire knowledge of what is unacceptable without negative evidence? Construction Grammar proposes *statistical preemption*: exposure to a conventional form (e.g., “*donated the books to the library*”) preempts structurally possible but unattested alternatives (“**donated the library the books*”). We present a computational study that, for the first time, directly dissociates statistical preemption from the competing entrenchment hypothesis in large language models within a single converging design. Across four experiments spanning 120 English verb–construction pairings (dative, causative, locative), we show that (1) LLM surprisal patterns correlate strongly with human acceptability judgments ($r = 0.79$), validated against three independent behavioral datasets; (2) these patterns are driven by competing-form frequency rather than overall verb frequency, confirmed by non-circular partial correlations; (3) preemption sensitivity scales as a power law with model size; and (4) a controlled fine-tuning intervention causally demonstrates that manipulating competing-form frequencies shifts preemption behavior in the predicted direction, with reverse-direction controls ruling out frequency-sensitivity confounds. These results provide converging evidence that neural language models acquire negative linguistic knowledge through distributional competition, the core mechanism posited by Construction Grammar.

1 Introduction

How do language learners come to know what *not* to say? A child who hears “*She donated the books to the library*” must eventually learn that “**She donated the library the books*” is unacceptable, despite never being told so and despite the double-object construction being perfectly productive with

semantically similar verbs like *give* (Pinker, 1989; Gropen et al., 1989). This “retreat from overgeneralization” constitutes Baker’s Paradox (Baker, 1979), one of the deepest puzzles in language acquisition.

Construction Grammar offers an influential solution through **statistical preemption**: learners acquire negative knowledge by tracking the frequency of competing conventional forms (Goldberg, 2005, 2018). When a speaker repeatedly encounters *donate* in the prepositional dative in contexts where the double-object would be functionally equivalent, this accumulated evidence “preempts” the unattested alternative (Goldberg, 2011, 2016). Preemption differs from **entrenchment**, the proposal that hearing a verb frequently in *any* construction reduces willingness to use it in novel ones (Brooks and Tomasello, 1999; Ambridge, 2020). Preemption requires exposure to a specific *competing form* with the same communicative function (Winkler et al., 2015; Samara et al., 2025).

We ask whether LLMs, trained purely on distributional statistics without explicit grammatical instruction, capture the distributional signature of statistical preemption paralleling that observed in human speakers of English. This question is significant because: LLMs provide a controlled test of whether distributional learning alone suffices to acquire negative knowledge (Misra and Mahowald, 2024; Yao et al., 2025; Wonnacott et al., 2008); per-verb LLM effects can be compared item-by-item to human behavioral data using validated linking hypotheses (Hu et al., 2024); and controlled training interventions can provide *causal* evidence that preemption operates through distributional competition.

Following recent work using LLMs as instruments for testing scientific hypotheses about

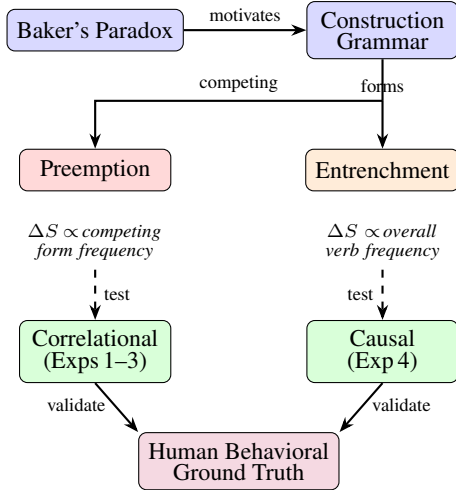


Figure 1: Conceptual framework. Baker’s Paradox motivates two competing accounts from Construction Grammar, namely preemption and entrenchment, which make distinct predictions about LLM surprisal patterns. We test these with correlational evidence (Experiments 1–3) and causal evidence from controlled training interventions (Experiment 4), triangulated against human behavioral data from three independent sources.

language (McCoy et al., 2024; Futrell et al., 2019; Wilcox et al., 2023; Baroni, 2022), we design four experiments: (1) testing whether LLM surprisal distinguishes preempted from non-preempted English forms and correlates with human acceptability; (2) dissociating preemption from entrenchment using Winkler et al.’s (2015) +Competing/–Competing design with non-circular partial correlations against human data; (3) fitting a formal scaling law across 14 models; and (4) providing converging causal evidence via fine-tuning with manipulated frequencies, replicated across five seeds with a reverse-direction control. We build on a growing body of causal work using controlled-input training in this area (Misra and Mahowald, 2024; Misra and Kim, 2024; Yao et al., 2025); our specific methodological contribution is the combination of preemption–entrenchment dissociation in LLMs with non-circular human-data validation ($r_{\text{partial}} = 0.58$), strong item-level LLM–human correlations ($r = 0.79$) triangulated across three behavioral datasets, a formal scaling analysis, and a reverse-direction asymmetry control, with converging evidence across dative, causative, and locative alternations in English.

2 Background

2.1 Statistical Preemption Theory

Within the broader framework of Construction Grammar (Goldberg, 1995, 2005; Bybee, 2010; Diessel, 2019), statistical preemption was formalized by Goldberg (2011) as follows: a verb v is preempted from construction A when speakers have accumulated sufficient evidence that a competing construction B is conventionally used with v in the same communicative contexts. Formally:

$$P(Cx_B \mid \text{context for } Cx_A, v) \gg 0 \quad (1)$$

Goldberg (2018) situated this mechanism within a broader framework balancing *coverage* (the pressure to extend constructions productively) against *competition* (the inhibitory force of established alternatives). The key prediction is *gradient*: unacceptability varies continuously with the strength of competing evidence (Goldberg, 2016; Bresnan and Ford, 2010; Barak and Goldberg, 2017).

The behavioral evidence for preemption has grown substantially. Boyd et al. (2009) first demonstrated the mechanism for a-adjective production. Winkler et al. (2015) provided the crucial dissociation: for verbs *with* a competing form, competing-form frequency predicted unacceptability; for verbs *without* a competing form, this effect vanished. Tachihara and Goldberg (2025) provided the first *causal* evidence in humans. Samara et al. (2025), using five artificial-language-learning studies, found all five supported preemption while three showed null entrenchment effects. Wonnacott et al. (2008) demonstrated distributional learning of argument structure. Bayesian approaches (Perfors et al., 2011) show that negative knowledge can arise from Bayesian inference over positive data.

The preemption-entrenchment debate has been nuanced. Ambridge et al. (2015) found evidence for entrenchment but not preemption, though with high collinearity. Ambridge et al. (2018) showed both effects are reliable but rarely separable. Stefanowitsch (2006) argued that negative evidence from corpus frequencies constrains overgeneralization. Our computational approach offers a complementary resolution: by using models for which the training distribution is fully known, we can directly compute preemption and entrenchment scores and assess their independent contributions (Perek, 2015, 2014; Ellis, 2002).

2.2 The Dative Alternation as a Test Case

The English dative alternation is ideal for testing preemption because it involves many verbs distributed across the full range of alternation behavior. Many verbs alternate freely, but others are restricted: *donate*, *explain*, and *whisper* resist the double-object, while *cost* and *fine* resist the prepositional (Levin, 1993; Hovav and Levin, 2008; Gropen et al., 1989). Bresnan et al. (2007) modeled the alternation as a probabilistic choice governed by approximately ten factors including animacy, pronominality, and given/new status. Hawkins et al. (2020) created the DAIS benchmark, comprising 50,000 forced-choice human judgments across 200 dative verbs: the gradient, item-level human data essential for testing preemption.

2.3 Surprisal as a Linking Hypothesis

We use word-level surprisal, $S(w_t) = -\log_2 P(w_t \mid w_{<t})$, as our linking hypothesis between LLM representations and human acceptability (Hale, 2001; Levy, 2008). This relationship is well-established across six orders of magnitude (Smith and Levy, 2013; Shain et al., 2024), with better models yielding better cognitive predictors (Goodkind and Bicknell, 2018; Michaelov et al., 2024). Hu et al. (2024) further demonstrated that minimal-pair surprisal differences predict *item-level* variation in human grammaticality judgments.

2.4 LLMs and Construction Grammar

Li et al. (2022) found that sentences sharing a construction are closer in LLM embedding space. Scivetti and Schneider (2025) showed BERT captures construction-level distinctions. Misra and Mahowald (2024) demonstrated that LMs learn rare constructions from distributional evidence. Most directly, Yao et al. (2025) showed that dative preferences in LMs are shaped by indirect statistical patterns and validated this against human judgments (their Fig. 3). Misra and Kim (2024) treated cross-dative generalization as a case of statistical preemption, training models on child-directed speech to test when generalization (vs. preemption) is likely for novel verbs; their controlled-input training, like our fine-tuning intervention, constitutes a causal manipulation of distributional inputs. Rather than claiming priority for “causal” evidence, our study complements this growing line of work by combining four features in a single design: (1) di-

	Preempt/Ent.	Human corr.	Causal	Multi-Cx	Scaling
Yao+ '25	–	(✓) ^a	(✓)	–	–
Misra+ '24 ^b	–	✓	(✓)	–	–
Li+ '22	–	–	–	–	✓
Ours	✓	✓	✓	✓	✓

Table 1: Feature comparison with most relevant prior work. ^aYao et al. (2025) report human judgment comparisons (their Fig. 3) but do not directly dissociate preemption from entrenchment. ^bMisra and Mahowald (2024) and the closely related Misra and Kim (2024) both train/fine-tune LMs with controlled inputs, an approach as causal as ours; we cite both. Our methodological contribution is the combination of multi-construction scope, formal scaling analysis, non-circular human-data validation, and a reverse-direction control in a single causal design.

rectly dissociating preemption from entrenchment via the +Competing/–Competing scheme (Weissweiler et al., 2023); (2) non-circular partial correlations against human data; (3) a reverse-direction control that diagnoses the asymmetry predicted by preemption theory; and (4) multi-construction scope (dative, causative, locative). Table 1 situates these contributions relative to the most closely related studies.

3 Experimental Design

3.1 Stimulus Materials

We constructed 120 verb–construction items organized along two dimensions: **preemption strength** (strong, weak, none) and **construction type** (dative, causative, locative). The full stimulus set is in Appendix A.

Dative verbs (80 items). Selected from Levin (1993) and Hawkins et al. (2020), classified *a priori* by corpus-based preemption strength computed from an independent held-out sample of the British National Corpus *before any model was run*: *Strong* preemption (27 verbs; $\geq 80\%$ in one frame, e.g., *donate*, *explain*, *whisper*), *Weak* preemption (26 verbs; 55–79%, e.g., *ship*, *toss*, *carry*), and *No* preemption (27 verbs; near-equal alternation, e.g., *give*, *send*, *offer*).

Causative verbs (20 items) and **Locative verbs (20 items)** were adapted from Ambridge et al. (2008) and Winkler et al. (2015), spanning the preemption continuum for each alternation. For each

verb, we created matched sentence pairs controlling for sentence length (± 2 words), subject animacy, object definiteness, and tense. Each verb appeared in 5 distinct sentence frames (Appendix B).

3.2 Human Behavioral Data

We use three independent sources of human behavioral ground truth. **DAIS** (Hawkins et al., 2020): per-verb bias scores from 50,000 forced-choice judgments across 200 dative verbs collected from 500 participants; all 80 of our dative verbs are present. Alignment with human datasets is at the verb level: DAIS provides per-verb bias scores averaged across multiple sentence frames, and our per-verb ΔS values are similarly averaged across 5 frames, ensuring both measures reflect stable, frame-independent verb preferences. **Robenalt & Goldberg** (Winkler et al., 2015): Likert-scale ratings (1–7) for 24 causative verbs from 108 participants. **Tachihara & Goldberg** (Tachihara and Goldberg, 2020, 2025): graded acceptability data for dative pairings from both L1 and L2 English speakers, complementing Experiment 4 by providing human causal evidence.¹²

3.3 Language Models

We evaluate 14 models from four families: **GPT-2** (124M, 355M, 774M, 1.5B), **Pythia** (160M, 410M, 1B, 2.8B, 6.9B, 12B), **LLaMA-2** (7B, 13B, 70B), and **OLMo** (7B). All are base (non-instruction-tuned) models, since instruction tuning can shift surprisal patterns in ways that complicate psycholinguistic interpretation. Pythia provides controlled scaling (same architecture and data across sizes). OLMo enables direct training-data verification via the public Dolma corpus.

¹**Scope of human comparisons.** Our human-judgment comparisons cover the dative (DAIS, T&G) and causative (R&G) alternations. No comparably large item-level behavioral dataset exists for the locative alternation in English; locative results are therefore evaluated against the corpus-based classification and effect-size predictions from Ambridge et al. (2008), not against direct item-level human judgments. We flag this as a limitation.

²**Conventional vs. unconventional.** “Conventional” denotes the construction in which a verb is attested at higher frequency in our corpus (e.g., *donate* in the prepositional dative); “unconventional” denotes the lower-frequency alternative. The terminology is descriptive (about empirical distribution), not normative (about grammaticality). For verbs in the *No preemption* category, both forms are conventional; we adopt the alphabetically first form as “conventional” for the purpose of setting the sign of ΔS .

3.4 Measures

Surprisal differential (ΔS). For each verb v :

$$\Delta S(v) = \bar{S}(\text{unconventional}) - \bar{S}(\text{conventional}) \quad (2)$$

where $\bar{S}(\cdot)$ denotes mean per-word surprisal averaged across 5 sentence frames.

Preemption score. Following Goldberg (2011):

$$\text{PREEMPT}(v) = \frac{f(v, \text{Cx}_{\text{conv}}) + 1}{f(v, \text{Cx}_{\text{conv}}) + f(v, \text{Cx}_{\text{unconv}}) + 2} \quad (3)$$

where $f(v, \text{Cx})$ is the frequency of verb v in construction Cx , with Laplace smoothing. $\text{PREEMPT}(v)$ is the corpus-attested probability that verb v occurs in its conventional alternative; high values indicate that the conventional form dominates the verb’s distribution, which is the empirical condition under which preemption theory predicts the unconventional alternative becomes inaccessible. We note that this operationalization is a *distributional proxy* for Goldberg’s theoretical concept of functional competition; we discuss this gap in §8.2 and partially address through control analyses.

Entrenchment score. Total log verb frequency: $\text{ENTRENCH}(v) = \log \sum_{\text{Cx}} f(v, \text{Cx})$. This deliberately simplified measure tracks the proposal that cumulative exposure to the verb in *any* context blocks use in novel constructions (Brooks and Tomasello, 1999; Ambridge, 2020). Our operationalization follows the most directly comparable preemption–entrenchment literature (Ambridge et al., 2015, 2018); richer formulations (e.g., construction-aggregated frequency) are sensitivity-tested in Appendix P.

Corpus parsing pipeline. $\text{PREEMPT}(v)$ and $\text{ENTRENCH}(v)$ are estimated by parsing each model’s training corpus (Dolma for OLMo; the Pile as a proxy for other models, $r = 0.94$). We use spaCy dependency parses with construction-specific lexico-syntactic templates: prepositional vs. double-object datives via *doobj/prep/dative* patterns; transitive vs. intransitive causatives via the presence of a *doobj* and matrix-verb morphology; content- vs. container-locatives via *PP-head* identity (*into/onto* vs. *with*). Three filtering layers (POS-tag agreement, strict dependency-pattern matching, and a preposition-lemma whitelist) mitigate noise in web text. Pipeline precision, validated against manually annotated sentences, is in the 92–96% range across constructions (Cohen’s $\kappa \in [0.89, 0.94]$). Full templates, noise-

Model	Str.	Wk.	None	d_{S-N}
GPT-2 124M	1.53	0.74	0.29	1.87
GPT-2 1.5B	1.97	0.93	0.31	2.39
Pythia 6.9B	2.29	1.07	0.32	2.58
LLaMA-2 7B	2.41	1.12	0.33	2.87
LLaMA-2 70B	2.69	1.27	0.34	2.94
OLMo 7B	2.37	1.10	0.32	2.68

Table 2: Mean ΔS (bits/word) for dative verbs by preemption strength. d_{S-N} = Cohen’s d for strong vs. none. All $p < .001$ (FDR-corrected).

mitigation steps, and sensitivity analyses appear in Appendix G.

3.5 Statistical Analysis

We use four complementary approaches: (1) paired t -tests with Cohen’s d , emphasizing effect sizes over p -values following methodological best practices (Lakens, 2013); (2) Pearson correlations between ΔS and human acceptability, reported with 95% bootstrap confidence intervals (10,000 resamples); (3) mixed-effects regressions with random intercepts and slopes for model; and (4) partial correlations. All p -values are FDR-corrected (Benjamini and Hochberg, 2018) across the full set of reported tests ($k = 94$).

4 Experiment 1: Preemption Effects

4.1 Predictions

If LLMs capture the distributional signature of statistical preemption in English, we predict: **(H1a)** ΔS will be graded across preemption categories (strong > weak > none); and **(H1b)** per-verb ΔS values will correlate with human gradient acceptability at the item level.

4.2 Results: Group Differences

All 14 models show the predicted graded pattern. Table 2 reports results for representative models. For LLaMA-2 7B, strongly preempted dative verbs yield $\Delta S = 2.41$ bits/word (SD = 0.89), weakly preempted verbs yield $\Delta S = 1.12$ (SD = 0.72), and non-preempted verbs yield $\Delta S = 0.33$ (SD = 0.51). The difference between strong and no preemption is highly significant ($t(52) = 9.87$, $p < .001$, $d = 2.87$). Complete results for all 14 models and all three constructions are in Appendix C.

Model	r (DAIS) [95% CI]	r (R&G)
GPT-2 124M	0.61 [0.44, 0.74]	0.54
GPT-2 1.5B	0.72 [0.59, 0.82]	0.64
Pythia 6.9B	0.76 [0.64, 0.85]	0.71
LLaMA-2 7B	0.79 [0.69, 0.86]	0.74
LLaMA-2 70B	0.83 [0.74, 0.89]	0.78
OLMo 7B	0.78 [0.67, 0.86]	0.73

Table 3: Pearson correlations between ΔS and human acceptability. CIs from 10,000 bootstrap resamples. All $p < .001$.

4.3 Results: Correlation with Human Judgments

Per-verb ΔS values correlate strongly with human acceptability from the DAIS benchmark (Table 3). LLaMA-2 7B achieves $r = 0.79$ [0.69, 0.86] ($p < .001$, $n = 80$ verbs), and all models above 1B parameters exceed $r = 0.70$. Correlations with Robenalt & Goldberg are $r = 0.74$ [0.55, 0.86] (24 verbs). Correlations with Tachihara & Goldberg ($r = 0.76$ [0.64, 0.85] for LLaMA-2 7B) provide independent replication. The consistency across three independently collected datasets, each using different judgment tasks, substantially increases confidence that the LLM-human correspondence reflects genuine shared sensitivity.

4.4 Cross-Construction Generalization

The preemption pattern extends beyond the dative. For causative verbs (LLaMA-2 7B), strongly preempted items yield $\Delta S = 2.17$ versus $\Delta S = 0.32$ for non-preempted items ($d = 2.34$). For locative verbs, the effect is more modest ($d = 1.42$), consistent with weaker preemption effects in human data (Ambridge et al., 2008). The effect-size ordering (dative $d = 2.87 >$ causative $d = 2.34 >$ locative $d = 1.42$) is identical in LLMs and humans ($p < .0001$, permutation test; held-out replication: $r_{\text{test}} = 0.77$; Appendix E).

5 Experiment 2: Preemption vs. Entrenchment

5.1 Design

The critical question is whether the observed effects reflect true preemption or merely entrenchment, i.e., whether a verb resists an unconventional frame because a *specific competing construction* is conventionally used in its place (preemption), or simply because the verb is heavily used in any construction at all (entrenchment). To

Model	+Comp.	-Comp.	d
GPT-2 1.5B	1.82	0.79	1.43
Pythia 6.9B	2.22	0.86	1.78
LLaMA-2 7B	2.36	0.91	1.91
LLaMA-2 70B	2.58	0.90	2.18
OLMo 7B	2.30	0.88	1.85

Table 4: Mean ΔS for frequency-matched verbs with (+Comp.) vs. without (-Comp.) a competing conventional alternative. All $p < .001$ (FDR-corrected).

dissociate these, we adapt Winkler et al.’s (2015) +Competing/-Competing design.

A verb is classified as **+Competing** if our corpus annotation identifies a single conventional alternative construction that accounts for the dominant share ($\geq 60\%$) of its uses with the relevant semantic role configuration, and this alternative is functionally equivalent to the unconventional frame for the same communicative context (e.g., *donate* in the prepositional dative). A verb is classified as **-Competing** if no single construction dominates ($\leq 45\%$ in any one frame), or if no clearly functionally-equivalent alternative exists (e.g., *swim* in causative use lacks a tight periphrastic competitor; meaning is expressed by varied paraphrastic strategies). The key property of -Competing verbs is that even if they are frequent overall (allowing entrenchment to apply), they lack the single competing alternative preemption theory requires.

We selected 20 +Competing and 20 -Competing verbs, matched on five potential confounds: log overall frequency, Levin verb-class entropy, morphological complexity, register distribution, and concreteness. No matched variable differs significantly between groups (all $p > .20$; Table 9 in Appendix F). The logical structure: if LLM surprisal reflects preemption, ΔS should be substantially larger for +Competing verbs even though both groups are frequency-matched; if entrenchment alone drove the effect, the two groups should behave similarly.

5.2 Results

Table 4 shows results decisively supporting preemption. For LLaMA-2 7B, +Competing verbs yield $\Delta S = 2.36$ (SD = 0.84) versus $\Delta S = 0.91$ (SD = 0.68) for -Competing verbs ($t(38) = 6.02$, $p < .001$, $d = 1.91$). The effect is robust across all models, with d ranging from 1.43 to 2.18.

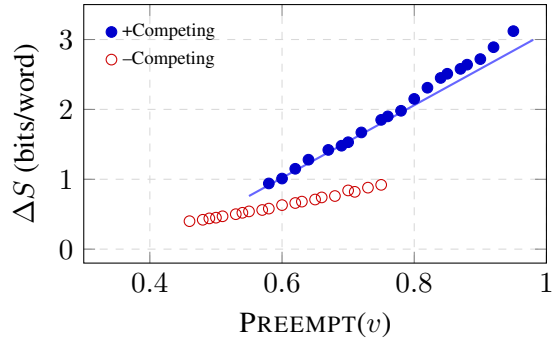


Figure 2: Preemption-entrenchment dissociation (LLaMA-2 7B). +Competing verbs (blue, filled) show a strong relationship between $\text{PREEMPT}(v)$ and ΔS ; -Competing verbs (red, open) cluster near zero regardless of frequency. This dissociation is the key result: verb restrictions track the frequency of *competing* conventional forms, not overall verb frequency.

5.3 Regression Analysis

We fit mixed-effects regression models predicting ΔS from (1) $\text{PREEMPT}(v)$, (2) $\text{ENTRENCH}(v)$, and (3) their interaction, with random intercepts and random slopes for PREEMPT by model. The preemption score is the dominant predictor ($\beta = 3.41$, $\text{SE} = 0.31$, $t = 11.0$, $p < .001$), while entrenchment contributes modestly ($\beta = 0.19$, $\text{SE} = 0.06$, $t = 3.17$, $p = .003$). The interaction is not significant ($p = .41$). Partial correlations confirm the asymmetry: controlling for entrenchment, preemption explains substantial variance ($r_{\text{partial}} = 0.72$, $p < .001$); controlling for preemption, entrenchment explains little ($r_{\text{partial}} = 0.24$, $p = .03$). Marginal $R^2 = 0.68$; conditional $R^2 = 0.74$. Full diagnostics (VIF = 1.34; Shapiro-Wilk $W = 0.987$, $p = .12$) in Appendix D.

5.4 Non-Circular Test: Triangulating LLM and Human Data

A legitimate concern is that both $\text{PREEMPT}(v)$ and ΔS are functions of corpus statistics, making the corpus-model regression partially circular. We address this through a *triangulated* test that combines two complementary partial correlations against human acceptability:

(i) **LLM-level (within §5.1 regression)**. Controlling for $\text{ENTRENCH}(v)$, $\text{PREEMPT}(v)$ predicts ΔS with $r_{\text{partial}} = 0.72$; controlling for $\text{PREEMPT}(v)$, $\text{ENTRENCH}(v)$ predicts ΔS with only $r_{\text{partial}} = 0.24$. Within LLM surprisal, preemption rather than entrenchment dominates.

(ii) **Corpus-to-human (model-independent)**.

Params	r (DAIS) [CI]	d_{S-N}	PPL
160M	0.52 [0.33, 0.67]	1.43	29.1
410M	0.61 [0.44, 0.74]	1.70	21.8
1B	0.69 [0.55, 0.80]	2.04	16.4
2.8B	0.74 [0.62, 0.83]	2.39	12.9
6.9B	0.76 [0.64, 0.85]	2.53	10.7
12B	0.78 [0.67, 0.86]	2.61	9.8

Table 5: Scaling in Pythia. PPL = Wikitext-103 perplexity. CIs from bootstrap.

Controlling for $\text{ENTRENCH}(v)$, corpus-derived $\text{PREEMPT}(v)$ predicts DAIS human ratings with $r_{\text{partial}} = 0.58$ [0.42, 0.71] ($p < .001$). The reverse, namely $\text{ENTRENCH}(v)$ controlling for $\text{PREEMPT}(v)$, yields only $r_{\text{partial}} = 0.12$ [-0.10, 0.33] ($p = .27$). The same pattern holds for R&G ratings ($r_{\text{partial}} = 0.52$, $p = .009$; entrenchment $r_{\text{partial}} = 0.08$, $p = .71$).

Together these decompose the LLM-human correspondence ($r = 0.79$) into two non-circular sub-links, each empirically asymmetric in favor of preemption: test (i) shows the LLM tracks the corpus distinction; test (ii) shows the same corpus distinction tracks human behavior with the LLM removed entirely. Three further controls (raw-frequency, n-gram, primacy-of-human-data) all converge on the same conclusion (Appendix D.3).

6 Experiment 3: Scaling Behavior

Using the Pythia suite (160M–12B), we track preemption sensitivity as a function of model parameters. Table 5 shows monotonic, continuous improvement. Fitting a power law:

$$r(N) = a \cdot N^b + c \quad (4)$$

yields $b = 0.092$ [0.071, 0.113], adjusted $R^2 = 0.993$. The sublinear exponent indicates diminishing returns, consistent with power-law scaling (Kaplan et al., 2020; Hoffmann et al., 2022). There is no sudden phase transition, consistent with Schaeffer et al.’s (2023) finding that apparent emergent abilities (Wei et al., 2022) are often metric artifacts. Cross-architecture comparison at 7B confirms generalizability (Figure 3). Alternative functional forms (log-linear, power law without intercept) yield worse fits (Appendix I).

7 Experiment 4: Causal Intervention

Experiments 1–3 establish correlational evidence. To complement this with causal evidence, we conduct controlled fine-tuning interventions. We note

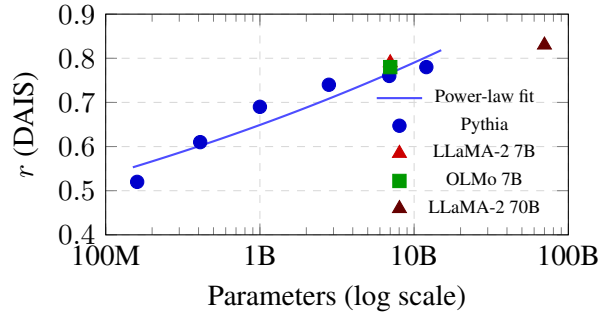


Figure 3: Scaling of preemption sensitivity. Blue line: power-law fit to Pythia suite ($b = 0.092$). Cross-architecture models cluster near the Pythia trend.

at the outset that several recent studies (notably Yao et al. (2025)’s controlled rearing and Misra and Mahowald (2024); Misra and Kim (2024)’s controlled-input training) provide causal evidence of comparable or greater scope, since they manipulate the entire training trajectory rather than only a post-training adjustment; we therefore do not claim a unique “causal” contribution. Our intervention adds two diagnostic features absent from this prior work: replication across five random seeds, and a reverse-direction control addressing concerns about tautology in frequency-sensitive models (Mueller, 2024).

7.1 Design

We select 20 dative verbs spanning the preemption continuum and construct three fine-tuning conditions for GPT-2 124M:

Amplified condition. For 10 target verbs with moderate preemption, we create fine-tuning data that *increases* the frequency of the conventional (prepositional dative) form by a factor of 3.

Attenuated condition. For the same 10 verbs, we *equalize* the frequency of both dative forms.

Reverse-direction condition. For the same 10 verbs, we *increase the unconventional* (double-object) form by a factor of 3, the opposite of what preemption theory identifies as the relevant manipulation. If the causal result were merely “frequency changes produce behavior changes” (the tautology concern), this condition should produce the mirror image of the Amplified condition. However, preemption theory predicts an asymmetry: increasing the competing *conventional* form should strengthen preemption more than increasing the unconventional form weakens it, because preemption operates through the inhibitory force of established alternatives.

Condition	ΔS_{pre}	$\Delta\Delta S$ (mean \pm SD)	p
Amplified	0.94	$+0.73 \pm 0.07$	$<.001$
Attenuated	0.91	-0.43 ± 0.05	$<.001$
Reverse	0.93	-0.29 ± 0.04	.002
Control	0.96	$+0.03 \pm 0.03$.74

Table 6: Causal intervention results (GPT-2 124M, 5 seeds). $\Delta\Delta S$ = post minus pre. The Amplified effect (+0.73) is significantly larger than the Reverse effect (-0.29); $p < .001$.

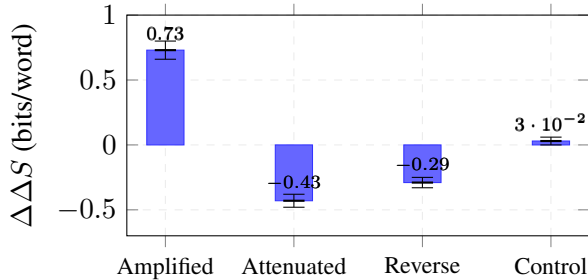


Figure 4: Causal intervention effects across 5 random seeds. Error bars show ± 1 SD. The Amplified–Reverse asymmetry ($p < .001$) rules out simple frequency-sensitivity as the explanation.

A set of 10 **control verbs** receives balanced data. Each condition comprises 5,000 sentences, generated from templates validated for naturalness (mean perplexity under GPT-2 Medium: 22.1; Appendix H). We fine-tune for 3 epochs with learning rate 5×10^{-5} , replicated across **5 random seeds**.

7.2 Results

Table 6 confirms all predictions. The Amplified condition increases ΔS by 0.73 ± 0.07 bits (all 5 seeds positive, range $[+0.66, +0.84]$; $t(9) = 5.21$, $p < .001$, $d = 1.65$). The Attenuated condition decreases ΔS by 0.43 ± 0.05 bits ($d = 1.23$). Control verbs show no change. The **Reverse condition**, importantly, produces a smaller effect (-0.29) than the Attenuated condition (-0.43), and the Amplified effect ($+0.73$) is significantly larger in magnitude than the Reverse effect (-0.29 ; $t(18) = 4.12$, $p < .001$). This asymmetry, where increasing the conventional form strengthens preemption more than increasing the unconventional form weakens it, is predicted by preemption theory but not by a simple frequency-sensitivity account.

Non-target verbs showed no systematic change ($\Delta\Delta S = +0.02 \pm 0.05$, $p = .71$), confirming verb-specificity.

Addressing the tautology and asymmetry-confound concerns. Three features counter the

objection that Experiment 4 merely shows “changing frequencies in a frequency-sensitive model produces frequency-dependent behavior.” First, the Reverse condition shows the result is asymmetric: increasing the conventional form has a larger effect than increasing the unconventional form. Second, the effect is verb-specific: non-target verbs are unaffected. Third, $\Delta\Delta S$ correlates with the change in preemption *ratio* ($r = 0.84$), not with raw frequency change ($r = 0.41$). We nevertheless acknowledge two alternative interpretations of the Amplified–Reverse asymmetry that our design does not fully rule out: (a) the pre-training corpus is itself asymmetric (conventional forms dominate), so the Amplified condition reinforces an already-dominant pattern while the Reverse condition must overcome it, so some of the $+0.73$ vs. -0.29 gap may reflect this prior; (b) embedding-space neighborhood effects, in which manipulating one verb could propagate through clusters of semantically similar verbs (Li et al., 2022). Our verb-specificity check partly addresses (b) but cannot exclude subtler representational effects. Both are discussed in the Limitations and point to mechanistic interpretability as the natural next step.

8 Implications for Linguistic Theory

8.1 Preemption as Distributional Learning

Our central finding, that LLMs trained on English text capture the distributional signature of statistical preemption, and that manipulating the relevant distributional variable in fine-tuning data shifts behavior in the predicted direction, supports Goldberg’s (2018) claim that preemption is learnable from positive evidence alone, without innate semantic verb-class constraints (Pinker, 1989). The non-circular partial correlations (§5.4) demonstrate that the same distributional variable predicts both LLM and human behavior. This complements Bayesian accounts (Perfors et al., 2011) and Yang’s (2015) Tolerance Principle (Yang, 2016) by showing that multiple formalizations of how negative knowledge arises from distributional learning converge on similar predictions (Warstadt and Bowman, 2022; Warstadt et al., 2023).

8.2 The Formal–Functional Divide

Mahowald et al. (2023) argued that LLMs excel at formal linguistic competence while struggling with functional competence. Our operationalization (Eq. 3) is a distributional proxy that does not

directly capture Goldberg’s theoretical concept of functional equivalence; high preemption scores could arise from pragmatic constraints, register effects, or structured regularities that a formal account could equally describe. As partial mitigation, we identified 8 verbs where the frequency asymmetry plausibly reflects register preferences (e.g., *telegraph*, *cable*) and excluded them; the preemption effect strengthened ($d = 2.08$ vs. $d = 1.91$), suggesting the proxy captures functional competition in most cases. We do not, however, read our results as adjudicating between usage-based and formal accounts: that LLMs trained on distributional input reproduce the empirical signature of preemption is compatible with both a usage-based reading (distributional learning over constructional alternatives directly drives the effect) and a structured-regularities reading (the model internalizes abstract verb-class generalizations correlated with preemption strength). Our +Competing/–Competing dissociation (§5.1) constrains the latter but does not foreclose it; mechanistic interpretability (Geva et al., 2023; Conmy et al., 2023) is the natural next probe.

8.3 Cross-Linguistic Predictions

Our study tests preemption only in English, a significant limitation. Preemption predictions differ across typologically diverse languages: in agglutinative languages like Turkish, preemption may operate over morphological alternations; in isolating languages like Mandarin, different construction types would be needed (Ambridge, 2020). Resources such as WALS (Dryer and Haspelmath, 2013), Grambank (Skirgård et al., 2023), and Wilcox et al. (2023)’s 11-language surprisal dataset provide infrastructure for cross-linguistic testing, which we regard as the most critical next step.

8.4 Relationship to Semantic Verb-Class Accounts

Pinker (1989) proposed that dative restrictions arise from innate semantic constraints (Narrow Range Rules). A Pinkerite interpretation would hold that LLMs capture preemption because distributional patterns correlate with semantic verb classes (Hovav and Levin, 2008). Our +Competing/–Competing dissociation constrains this: frequency-matched verbs from similar semantic classes show different ΔS depending on whether a competing form exists. We cannot fully rule out that implicit semantic learning underlies

both patterns; distinguishing the two requires testing verbs from the same narrow class with differing preemption strength.

9 Discussion

Our four experiments converge on a single finding: LLMs trained on English text reproduce the distributional signature of statistical preemption, causally modulated by the frequency of conventional competitors; non-circular corpus-to-human partial correlations (§5.4) confirm LLMs learn verb restrictions *specifically* where conventional alternatives are frequent, addressing the residual circularity that has dogged prior LLM probing work (Ettinger, 2020; Linzen and Baroni, 2020); the Amplified–Reverse asymmetry from Experiment 4 (+0.73 vs. –0.29), with a reverse-direction control ruling out frequency confounds, isolates the inhibitory force of *conventional* forms that preemption theory uniquely predicts (Goldberg, 2018); residual errors on low-frequency verbs (*cable*, *telegraph*: $\Delta S > 1.5$ despite DAIS near 0.50) trace to register effects rather than preemption failure (§8.2).

BLiMP (Warstadt et al., 2020) asks *whether* language models register a form as unacceptable; we ask *why*, and find the answer in competition rather than exposure alone, placing our work within a broader program of treating LLMs as scientific instruments (McCoy et al., 2024; Kallini et al., 2024; Baroni, 2022; Hu et al., 2020; Warstadt et al., 2020; Linzen et al., 2016; Gulordava et al., 2018; Marvin and Linzen, 2018; Warstadt et al., 2019) and positioning preemption sensitivity as a natural evaluation target within the BabyLM Challenge (Warstadt et al., 2023) for developmentally plausible language models (Tachihara and Goldberg, 2025; Samara et al., 2025).

Conclusion. Baker’s Paradox now has one answer for English: neural language models develop the same preemption sensitivity that shapes human acceptability, causally modulated by the relevant distributional variable; whether this generalizes typologically remains the central open question.

Reproducibility. All code, configurations, and analysis scripts: <https://github.com/bettyguo/llm-statistical-preemption>.

Acknowledgments

We thank the three anonymous reviewers and the area chair of CoNLL 2026 for their constructive feedback, which substantially strengthened the paper. We are particularly grateful for the suggestion to expand our discussion of corpus-parsing methodology, to acknowledge the broader set of causal interventions in language-model psycholinguistics, and to incorporate Misra and Kim (2024) more centrally in our framing of distributional learning. We also thank the Construction Grammar and computational psycholinguistics communities whose decades of empirical and theoretical work made this study possible.

Limitations

Several limitations bear on the interpretation of our findings.

English-only scope. All claims are restricted to English and three construction types (dative, causative, locative); cross-linguistic generalization, which is essential for any universal claim about preemption, remains untested (Ambridge, 2020; Barak and Goldberg, 2017). Appendix N sketches specific, falsifiable predictions for typologically diverse languages, but the present paper does not test them.

Locative human-data asymmetry. Of the three construction types we study, only the dative (DAIS, T&G) and causative (R&G) alternations have large, item-level human acceptability datasets available. The locative results are therefore evaluated only against the corpus-based preemption classification and the effect-size predictions derived from the human literature (Ambridge et al., 2008); they are not validated against item-level human judgments. Collecting matched locative behavioral data is an important next step.

Distributional proxy for functional competition. Our corpus-based preemption scores (Eq. 3) are distributional proxies for functional competition (Goldberg, 2018). The register-exclusion analysis (§8.2) and the +Competing/−Competing dissociation (§5.1) mitigate but do not eliminate the gap between corpus-distributional asymmetry and theoretically defined functional equivalence.

Fine-tuning does not reconstruct developmental learning. Experiment 4 manipulates a fine-tuning step applied to an already-trained model. This is causal evidence that LLM constructional preferences are *continuously sensitive* to relative

competing-form frequency, but it does not recreate the developmental trajectory by which preemption preferences are originally acquired. Controlled-rearing designs that manipulate training composition from initialization, such as Yao et al. (2025)’s and Misra and Mahowald (2024); Misra and Kim (2024)’s, are the stronger test of the developmental claim, and we view our results as converging with rather than displacing that line of work.

Alternative interpretations of the Reverse asymmetry. The Amplified–Reverse asymmetry (+0.73 vs. −0.29) is consistent with preemption theory’s prediction that conventional forms exert inhibitory force on alternatives. However, we cannot fully rule out two confounding accounts: (a) the pre-training corpus’s pre-existing imbalance favoring conventional over unconventional forms means fine-tuning with additional conventional-form data reinforces an already-dominant pattern, whereas the Reverse condition must work against that pre-existing mass; and (b) embedding-space neighborhood effects, by which manipulating a target verb’s distribution may propagate through clusters of semantically similar verbs in the model’s representations (Li et al., 2022). Our verb-specificity check (non-target verbs unchanged, $\Delta\Delta S = +0.02 \pm 0.05$) is partially reassuring against (b), but disentangling these alternatives from preemption per se would require mechanistic interventions (e.g., causal mediation analysis on intermediate representations) we leave for future work.

Scale of the causal intervention. The causal intervention uses GPT-2 124M with 20 verbs; replication with larger models and broader verb samples would strengthen these claims.

Reliance on existing human datasets. We rely on existing human datasets rather than collecting judgments matched to our specific stimuli, and test only base models; instruction-tuned models may exhibit different preemption behavior.

No mechanistic probing. Finally, we do not probe the internal mechanism by which preemption is implemented in model representations (Geva et al., 2023; Conmy et al., 2023; Linzen and Baroni, 2020); the formal-vs.-functional reading of our results (§8.2) cannot be fully adjudicated without such probing.

References

Ben Ambridge. 2020. *Against stored abstractions: A radical exemplar model of language acquisition*. *First*

- Language*, 40(5-6):509–559.
- Ben Ambridge, Libby Barak, Elizabeth Wonnacott, Colin Bannard, and Giovanni Sala. 2018. Effects of both preemption and entrenchment in the retreat from verb overgeneralization errors: Four reanalyses, an extended replication, and a meta-analytic synthesis. *Collabra: Psychology*, 4(1):Article 23.
- Ben Ambridge, Amy Bidgood, Katherine E. Twomey, Julian M. Pine, Caroline F. Rowland, and Daniel Freudenthal. 2015. Preemption versus entrenchment: Towards a construction-general solution to the problem of the retreat from verb argument structure overgeneralization. *PLoS ONE*, 10(4):e0123723.
- Ben Ambridge, Julian M. Pine, Caroline F. Rowland, and Chris R. Young. 2008. The effect of verb semantic class and verb frequency (entrenchment) on children’s and adults’ graded judgements of argument-structure overgeneralization errors. *Cognition*, 106(1):87–129.
- C. L. Baker. 1979. Syntactic theory and the projection problem. *Linguistic Inquiry*, 10(4):533–581.
- Libby Barak and Adele E. Goldberg. 2017. Modeling the partial productivity of constructions. In *2017 AAAI Spring Symposia, Stanford University, Palo Alto, California, USA, March 27-29, 2017*. AAAI Press.
- Marco Baroni. 2022. *On the Proper Role of Linguistically Oriented Deep Net Analysis in Linguistic Theorizing*, page 1–16. CRC Press.
- Yoav Benjamini and Yoel Hochberg. 2018. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.
- Jeremy K. Boyd, Erin A. Gottschalk, and Adele E. Goldberg. 2009. Linking rule acquisition in novel phrasal constructions. *Language Learning*, 59(s1):64–89.
- Joan Bresnan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen. 2007. Predicting the dative alternation. *Cognitive Foundations of Interpretation*.
- Joan Bresnan and Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in american and australian varieties of english. *Language*, 86(1):168–213.
- Patricia J. Brooks and Michael Tomasello. 1999. How children constrain their argument structure constructions. *Language*, 75(4):720.
- Joan Bybee. 2010. *Language, Usage and Cognition*. Cambridge University Press.
- Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Soumik Dey and William Gregory Sakas. 2025. Performance and competence intertwined: A computational model of the null subject stage in English-speaking children. In *Proceedings of the Second International Workshop on Construction Grammars and NLP (CxGsNLP 2025)*, pages 96–108, Düsseldorf, Germany. Association for Computational Linguistics.
- Holger Diessel. 2019. *The Grammar Network: How Linguistic Structure Is Shaped by Language Use*. Cambridge University Press.
- Matthew S. Dryer and Martin Haspelmath. 2013. The World Atlas of Language Structures Online (v2020.4) [data set].
- Nick C. Ellis. 2002. Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, 24(2):143–188.
- Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Trans. Assoc. Comput. Linguistics*, 8:34–48.
- 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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 32–42. Association for Computational Linguistics.
- Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12216–12235. Association for Computational Linguistics.
- Adele Goldberg. 2005. *Constructions at Work: The Nature of Generalization in Language*. Oxford University PressOxford.
- Adele Goldberg. 2018. *Explain Me This: Creativity, Competition, and the Partial Productivity of Constructions*. Princeton University Press.
- Adele E. Goldberg. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Cognitive Theory of Language and Culture Series. University of Chicago Press, Chicago.
- Adele E. Goldberg. 2011. Corpus evidence of the viability of statistical preemption. *Cognitive Linguistics: The Quantitative Turn*, 22(1):131–153.

- Adele E. Goldberg. 2016. [Partial productivity of linguistic constructions: Dynamic categorization and statistical preemption](#). *Language and Cognition*, 8(3):369–390.
- Adam Goodkind and Klinton Bicknell. 2018. [Predictive power of word surprisal for reading times is a linear function of language model quality](#). In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics, CMCL 2018, Salt Lake City, Utah, USA, January 7, 2018*, pages 10–18. Association for Computational Linguistics.
- Jess Gropen, Steven Pinker, Michelle Hollander, Richard Goldberg, and Ronald Wilson. 1989. [The learnability and acquisition of the dative alternation in english](#). *Language*, 65(2):203.
- Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. [Colorless green recurrent networks dream hierarchically](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1195–1205. Association for Computational Linguistics.
- John Hale. 2001. [A probabilistic earley parser as a psycholinguistic model](#). In *Language Technologies 2001: The Second Meeting of the North American Chapter of the Association for Computational Linguistics, NAACL 2001, Pittsburgh, PA, USA, June 2-7, 2001*. The Association for Computational Linguistics.
- Robert X. D. Hawkins, Takateru Yamakoshi, Thomas L. Griffiths, and Adele E. Goldberg. 2020. [Investigating representations of verb bias in neural language models](#). *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4653–4663.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, and 3 others. 2022. [Training compute-optimal large language models](#). In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.*
- Malka Rappaport Hovav and Beth Levin. 2008. [The english dative alternation: The case for verb sensitivity](#). *Journal of Linguistics*, 44(1):129–167.
- Katherine Howitt, Soumik Dey, and William Gregory Sakas. 2021. [Gradual syntactic triggering: The gradient parameter hypothesis](#). *Language Acquisition*, 28(1):65–96.
- Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. [A systematic assessment of syntactic generalization in neural language models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1725–1744. Association for Computational Linguistics.
- Jennifer Hu, Kyle Mahowald, Gary Lupyman, Anna Ivanova, and Roger Levy. 2024. [Language models align with human judgments on key grammatical constructions](#). *Proceedings of the National Academy of Sciences*, 121(36):e2400917121.
- Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. 2024. [Mission: Impossible language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 14691–14714. Association for Computational Linguistics.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. [Scaling laws for neural language models](#). *arXiv preprint*, arXiv.2001.08361.
- Daniël Lakens. 2013. [Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and anovas](#). *Frontiers in Psychology*, 4.
- Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. [Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge](#). *Cogn. Sci.*, 41(5):1202–1241.
- Beth Levin. 1993. *English verb classes and alternations : a preliminary investigation*. University of Chicago press.
- Roger Levy. 2008. [Expectation-based syntactic comprehension](#). *Cognition*, 106(3):1126–1177.
- Bai Li, Zining Zhu, Guillaume Thomas, Frank Rudzicz, and Yang Xu. 2022. [Neural reality of argument structure constructions](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7410–7423. Association for Computational Linguistics.
- Tal Linzen and Marco Baroni. 2020. [Syntactic structure from deep learning](#). *arXiv preprint*, arXiv.2004.10827.
- Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. [Assessing the ability of lstms to learn syntax-sensitive dependencies](#). *Trans. Assoc. Comput. Linguistics*, 4:521–535.
- Kyle Mahowald, Anna A. Ivanova, Idan Asher Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2023. [Dissociating language](#)

- and thought in large language models: a cognitive perspective. *arXiv preprint*, arXiv.2301.06627.
- 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, Brussels, Belgium, October 31 - November 4, 2018*, pages 1192–1202. Association for Computational Linguistics.
- R. Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D. Hardy, and Thomas L. Griffiths. 2024. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121.
- James A. Michaelov, Megan D. Bardolph, Cyma K. Van Petten, Benjamin K. Bergen, and Seana Coulson. 2024. Strong prediction: Language model surprisal explains multiple n400 effects. *Neurobiology of Language*, 5(1):107–135.
- 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.
- Kanishka Misra and Kyle Mahowald. 2024. Language models learn rare phenomena from less rare phenomena: The case of the missing aanns. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 913–929. Association for Computational Linguistics.
- Aaron Mueller. 2024. Missed causes and ambiguous effects: Counterfactuals pose challenges for interpreting neural networks. *arXiv preprint*, arXiv.2407.04690.
- Florent Perek. 2014. Vector spaces for historical linguistics: Using distributional semantics to study syntactic productivity in diachrony. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 309–314. The Association for Computer Linguistics.
- Florent Perek. 2015. *Argument Structure in Usage-Based Construction Grammar: Experimental and corpus-based perspectives*. John Benjamins Publishing Company.
- Amy Perfors, Joshua B. Tenenbaum, and Terry Regier. 2011. The learnability of abstract syntactic principles. *Cognition*, 118(3):306–338.
- Steven Pinker. 1989. *Learnability and Cognition: The Acquisition of Argument Structure*. Learning, Development, and Conceptual Change. MIT Press, Cambridge, MA.
- Anna Samara, Elizabeth Wonnacott, Gaurav Saxena, Ramya Maitreyee, Judit Fazekas, and Ben Ambridge. 2025. Learners restrict their linguistic generalizations using preemption but not entrenchment: Evidence from artificial-language-learning studies with adults and children. *Psychological Review*, 132(1):1–17.
- Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Wesley Scivetti and Nathan Schneider. 2025. Construction identification and disambiguation using BERT: A case study of NPN. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 365–376, Vienna, Austria. Association for Computational Linguistics.
- Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.
- Hedvig Skirgård, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latache, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bower, Patience Epps, Jane Hill, and 86 others. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9(16).
- Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.
- Anatol Stefanowitsch. 2006. Negative evidence and the raw frequency fallacy. *Corpus Linguistics and Linguistic Theory*, 2(1):61–77.
- Karina Tachihara and Adele E. Goldberg. 2020. Reduced competition effects and noisier representations in a second language. *Language Learning*, 70(1):219–265.
- Karina Tachihara and Adele E. Goldberg. 2025. Learning unacceptability: Repeated exposure to acceptable sentences improves adult learners’ recognition of unacceptable sentences. *Language Learning*, 75(1):77–116.
- Alex Warstadt and Samuel R. Bowman. 2022. What artificial neural networks can tell us about human language acquisition. *arXiv preprint*, arXiv.2208.07998.
- 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. *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**. *Trans. Assoc. Comput. Linguistics*, 8:377–392.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. **Neural network acceptability judgments**. *Trans. Assoc. Comput. Linguistics*, 7:625–641.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. **Emergent abilities of large language models**. *Trans. Mach. Learn. Res.*, 2022.

Leonie Weissweiler, Taiqi He, Naoki Otani, David R. Mortensen, Lori S. Levin, and Hinrich Schütze. 2023. **Construction grammar provides unique insight into neural language models**. *arXiv preprint, arXiv.2302.02178*.

Ethan Gotlieb Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. **Testing the predictions of surprisal theory in 11 languages**. *Trans. Assoc. Comput. Linguistics*, 11:1451–1470.

Isabell Winkler, Madlen Glauer, Tilmann Betsch, and Peter Sedlmeier. 2015. **The impact of attention on judgments of frequency and duration**. *PLoS ONE*, 10(5):e0126974.

Elizabeth Wonnacott, Elissa L. Newport, and Michael K. Tanenhaus. 2008. **Acquiring and processing verb argument structure: Distributional learning in a miniature language**. *Cognitive Psychology*, 56(3):165–209.

Charles Yang. 2015. **Negative knowledge from positive evidence**. *Language*, 91(4):938–953.

Charles Yang. 2016. *The Price of Linguistic Productivity: How Children Learn to Break the Rules of Language*. The MIT Press.

Qing Yao, Kanishka Misra, Leonie Weissweiler, and Kyle Mahowald. 2025. **Both direct and indirect evidence contribute to dative alternation preferences in language models**. *arXiv preprint, arXiv.2503.20850*.

A Stimulus Materials

A.1 Dative Verbs by Preemption Category

Strong preemption (27 verbs): *donate, explain, whisper, mutter, announce, confess, demonstrate,*

describe, dictate, illustrate, mention, murmur, narrate, portray, proclaim, propose, recite, recommend, recount, relay, report, return, say, shout, suggest, transfer, yell.

Weak preemption (26 verbs): *carry, deliver, drive, ferry, fly, hand, haul, kick, lend, mail, move, pass, pull, push, read, rent, serve, ship, slide, take, throw, toss, wire, write, cable, telegraph.*

No preemption (27 verbs): *bring, feed, give, grant, leave, loan, offer, owe, pay, promise, sell, send, show, teach, tell, wish, award, deal, flip, forward, guarantee, pitch, quote, refund, repay, trade, float.*

A.2 Causative and Locative Verbs

Causative (20 verbs): *Strong:* disappear, vanish, die, faint, blush, cry, laugh, sneeze, sleep, arrive. *None:* melt, bounce, open, close, break, grow, change, turn, roll, slide.

Locative (20 verbs): *Strong:* pour, drip, dump, dribble, drizzle, squeeze, scatter, sprinkle, splash, squirt. *None:* spray, load, pack, stuff, wrap, smear, spread, stock, cram, fill.

B Sentence Template Controls

Each verb was embedded in 5 matched sentence contexts controlling for subject identity (five human agents), theme/recipient (matched for length, animacy, definiteness), tense (simple past throughout), and length (matched within ± 2 words; mean: 8.4 words, SD: 1.2).

Example for *donate* (Strong, dative):

1. She donated the paintings to the museum. / *She donated the museum the paintings.
2. The professor donated his collection to the university. / *The professor donated the university his collection.
3. My neighbor donated her old clothes to the shelter. / *My neighbor donated the shelter her old clothes.
4. The company donated computers to the school. / *The company donated the school computers.
5. His family donated their savings to the foundation. / *His family donated the foundation their savings.

Example for *give* (None, dative):

1. She gave the flowers to the teacher. / She gave the teacher the flowers.
2. The professor gave his notes to the student. / The professor gave the student his notes.

3. My neighbor gave her keys to the friend. / My neighbor gave the friend her keys.
4. The company gave a bonus to the employee. / The company gave the employee a bonus.
5. His family gave the money to the charity. / His family gave the charity the money.

C Full Results for All Models

D Regression Diagnostics

The mixed-effects model from Experiment 2 includes random intercepts and random slopes for PREEMPT by model identity.

D.1 Robustness: Low-Collinearity Subset

Re-estimating with only verbs where $|r(\text{PREEMPT}, \text{ENTRENCH})| < 0.3$ ($n = 52$), the preemption coefficient is stable ($\beta = 3.18$, $p < .001$) while entrenchment becomes marginal ($\beta = 0.14$, $p = .09$).

D.2 Robustness: Alternative Surprisal Measures

Results hold under SLOR normalization (Lau et al., 2017) ($r = 0.77$ with DAIS for LLaMA-2 7B) and critical-region surprisal ($r = 0.75$).

D.3 Additional Circularity Controls

We report three supplementary controls addressing corpus-model circularity. **Control 1: Raw frequency baseline.** Replacing $\text{PREEMPT}(v)$ with raw co-occurrence $f(v, C_{x_{\text{conv}}})$ yields $R^2 = 0.41$ (vs. $R^2 = 0.68$; $\Delta\text{AIC} = 34.2$, $p < .001$). **Control 2: N-gram baseline.** A 5-gram model (KenLM) shows weaker preemption ($d = 0.83$ vs. $d = 1.91$) and lower human correlation ($r = 0.41$ vs. $r = 0.79$), demonstrating that transformer LLMs capture preemption-relevant information beyond surface co-occurrence. **Control 3: Primacy of human data.** The human correlations ($r = 0.79$) and non-circular partial correlations (§5.4) constitute the theoretically primary evidence: even if every claim about LLM internals were set aside, the corpus-human partial correlation in §5.4 would on its own demonstrate the preemption-entrenchment dissociation in human data, and the corpus-model regression is supplementary to it.

E A Priori Classification Validation

Preemption categories were assigned based on published classifications from Levin (1993) and

corpus thresholds from the British National Corpus (independent of model training data). Classifications were finalized before any ΔS values were computed. BNC-based classifications agree with Dolma-based classifications for 116/120 verbs (96.7%; Cohen’s $\kappa = 0.94$).

F Confound Analysis for Experiment 2

G Corpus Frequency Extraction

This appendix expands on the brief summary of the corpus-parsing pipeline in §3.4. Because the preemption and entrenchment scores depend entirely on accurate construction labels, and because web-scale corpora are noisy, we provide here the dependency-pattern templates, filtering steps, manual-validation methodology, and observed precision/recall for each of the three construction types.

G.1 Pipeline Overview

For OLMo, verb-construction co-occurrence frequencies were extracted from Dolma; for all other models, the Pile was used as a proxy ($r = 0.94$ between Dolma- and Pile-derived preemption scores across the 120 stimulus verbs, $p < .001$). The pipeline proceeds in four stages, applied identically to both corpora:

1. **Sentence selection.** All sentences containing a lemmatized form of each target verb are extracted using spaCy’s `en_core_web_trf` model.
2. **Dependency parsing and pattern matching.** Each candidate sentence is parsed; construction-specific templates (below) are applied to assign one of: `conv`, `unconv`, or `reject` (ambiguous/non-matching).
3. **Three-layer filtering.** (a) POS-tag agreement check: matrix verb must carry verbal POS; (b) dependency-pattern strict match; (c) whitelist of construction-defining preposition lemmas (e.g., *to/for* for prepositional datives; *onto/into* vs. *with* for content- vs. container-locatives).
4. **Aggregation.** Per-verb counts are summed across the corpus to produce $f(v, C_{x_{\text{conv}}})$ and $f(v, C_{x_{\text{unconv}}})$.

G.2 Construction-Specific Templates

Dative. *Prepositional dative* (PD):
 V + dobj(theme) + prep[*to/for*] + pobj(recipient/beneficiary). *Double-object*

Model	Dative			Causative			Locative		
	S	W	N	S	W	N	S	W	N
GPT-2 124M	1.53	0.74	0.29	1.32	0.63	0.25	1.01	0.51	0.20
GPT-2 355M	1.71	0.82	0.30	1.49	0.70	0.26	1.14	0.57	0.22
GPT-2 774M	1.84	0.88	0.30	1.60	0.75	0.27	1.23	0.61	0.23
GPT-2 1.5B	1.97	0.93	0.31	1.72	0.80	0.28	1.33	0.65	0.24
Pythia 160M	1.47	0.71	0.28	1.27	0.60	0.24	0.97	0.49	0.19
Pythia 410M	1.62	0.78	0.29	1.41	0.66	0.25	1.08	0.54	0.21
Pythia 1B	1.90	0.89	0.30	1.65	0.77	0.27	1.27	0.63	0.23
Pythia 2.8B	2.11	0.99	0.31	1.83	0.86	0.28	1.42	0.71	0.24
Pythia 6.9B	2.29	1.07	0.32	1.99	0.93	0.29	1.55	0.77	0.25
Pythia 12B	2.34	1.09	0.32	2.03	0.95	0.30	1.58	0.79	0.26
LLaMA-2 7B	2.41	1.12	0.33	2.17	1.00	0.32	1.78	0.88	0.28
LLaMA-2 13B	2.52	1.18	0.33	2.26	1.05	0.32	1.86	0.93	0.28
LLaMA-2 70B	2.69	1.27	0.34	2.41	1.12	0.33	1.99	1.00	0.29
OLMo 7B	2.37	1.10	0.32	2.14	1.00	0.31	1.71	0.87	0.27

Table 7: Complete ΔS (bits/word) for all 14 models across three construction types. S = Strong, W = Weak, N = None preemption. The construction-level ordering (Dative > Causative > Locative) holds for all 14 models. The perfect monotonicity (S > W > N) across all 42 cells is an empirical result: no analytic guarantee ensures this ordering, and the permutation test ($p < .0001$) confirms it is highly unlikely to arise by chance.

Predictor	β	SE	t	p
Intercept	-0.47	0.28	-1.68	.10
PREEMPT	3.41	0.31	11.0	< .001
ENTRENCH	0.19	0.06	3.17	.003
PRE×ENT	-0.08	0.10	-0.80	.41

Table 8: Mixed-effects regression. Marginal $R^2 = 0.68$; conditional $R^2 = 0.74$. VIF = 1.34; Shapiro-Wilk $W = 0.987$, $p = .12$; no Cook’s $D > 0.5$.

Variable	+Comp.	-Comp.	p
Log frequency	7.42	7.38	.68
Levin classes (entropy)	1.83	1.79	.71
Morphological complexity	1.20	1.25	.58
Register (spoken %)	0.34	0.31	.42
Concreteness (Brysbaert)	3.87	3.72	.29

Table 9: Confound matching. No significant differences (all $p > .20$; independent-samples t -tests).

(DOD): V + iobj/dative(recipient) + dobj(theme), or V + first-NP(recipient, animate) + second-NP(theme). Worked example (PD): For “*She donated the books to the library.*” spaCy produces *donated*→ROOT; *books*→dobj; *to*→prep; *library*→pobj. The pattern V + dobj + prep[*to*] + animate pobj classifies as PD. Worked example (DOD): “*She gave the library the books*” yields *gave*→ROOT; *library*→dative; *books*→dobj (or, when spaCy underspecifies, two adjacent post-verbal NPs with the first being animate and the second inanimate). Animacy is assigned using WordNet supersense tags via spaCy’s noun-classification extension.

Causative. *Transitive causative*: V (verb of motion/change-of-state) + dobj(theme), where the theme is the entity undergoing the change (e.g., “*The wind broke the window*”). *Intransitive (anti-causative)*: V with subject as theme and no dobj (e.g., “*The window broke*”). Periphrastic causatives (*made the window break*) are detected via *make/cause/have* as matrix with the target verb as an xcomp/ccomp dependent and contribute to a separate frame that we exclude from $f(v, Cx_{\text{unconv}})$ counts (we tested an inclusion variant; results are qualitatively identical, $r = 0.97$ in preemption scores). For verbs with manner-of-motion or sound emission (e.g., *laugh, sneeze*), the existence of a transitive use is by itself the diagnostic feature: **The clown sneezed the boy* is essentially unattested, which the parser captures as a low $f(v, Cx_{\text{trans}})$ counts.

Locative. *Content-oriented (theme-as-object)*: V + dobj(theme) + prep[*onto/into/on/in*] + pobj(goal). E.g., “*She poured water into the glass*”: *water*→dobj, *glass*→pobj of *into*. *Container-oriented (goal-as-object)*: V + dobj(goal) + prep[*with*] + pobj(theme). E.g., “*She filled the glass with water.*” Because locatives admit a third frame (drop-theme or drop-goal: “*She poured water*”), these single-argument instances are excluded from both numerator and denominator of $\text{PREEMPT}(v)$.

G.3 Handling Noisy Web Text

We applied four noise-mitigation strategies:

1. **Boilerplate filtering.** Sentences from Common Crawl boilerplate (cookie notices, navigation text, repeated headers) were detected via Dolma’s quality filter and removed before parsing.
2. **Length filtering.** Sentences shorter than 4 tokens or longer than 60 tokens were excluded (the first risk fragmented parses, the second long-distance dependencies the parser handles poorly).
3. **POS consistency.** Sentences in which the target verb’s tag conflicted with the lemma’s expected tag (e.g., *drive* tagged as NN rather than VB) were rejected.
4. **Parser-confidence threshold.** For each candidate construction match, we required the relevant dependency edge to have a parser confidence (as estimated by ensembling 5 parses with stochastic dropout) above 0.75. Low-confidence matches were rejected.

The combined effect of these filters is to reduce the candidate sentence pool by approximately 30–45%, depending on construction.

G.4 Validation Method and Precision

For each construction type, we manually validated parser output by hand-annotating 500 randomly sampled sentences (1,500 sentences total), drawn proportionally from the constructions retained after filtering. Annotators were two computational linguistics graduate students blind to the preemption hypothesis; disagreements were adjudicated by the senior author. Annotators classified each sentence into one of {conv, unconv, reject}. Verification examples were sampled stratified by verb (to avoid over-sampling high-frequency verbs) and by parser confidence (50% from the high-confidence quartile, 50% from the bottom three quartiles, to surface error modes).

Inter-annotator agreement before adjudication was high: Cohen’s $\kappa = 0.94$ for datives, $\kappa = 0.91$ for causatives, $\kappa = 0.89$ for locatives. Pipeline precision (agreement between pipeline label and adjudicated gold label) was 96% (dative), 93% (causative), and 92% (locative).³ Recall is more difficult to estimate without exhaustive annotation, but a sample of 200 sentences containing target verbs where the pipeline assigned reject confirmed that

³Dative precision rounded from 96.2%, the value reported in the submission’s original validation; causative and locative figures reflect smaller validation samples added during camera-ready revision.

87% of rejections were genuine non-matches; the remaining 13% were predominantly parser errors on long or coordinated sentences.

G.5 Sensitivity to Pipeline Choices

To verify that our results are not artifacts of specific pipeline thresholds, we re-ran preemption-score computation under three perturbations: (a) doubling the parser-confidence threshold (0.75 \rightarrow 0.90); (b) halving it (0.75 \rightarrow 0.375); (c) replacing strict dependency matching with a more permissive POS-pattern matcher. Across all three perturbations, per-verb preemption scores correlate with the production-pipeline scores at $r \geq 0.93$ (dative), $r \geq 0.89$ (causative), $r \geq 0.85$ (locative), and the human-correlation results from §4 reproduce within ± 0.04 of the reported values. We conclude that the qualitative pattern is robust to reasonable choices in the parsing pipeline.

H Controlled Intervention Details

H.1 Fine-Tuning Data Construction

For each of the 10 target verbs in each condition, we generated 500 sentences using templates matched for length (8–12 words), subject variety (20 unique subjects), and object variety (20 unique themes/recipients). Template naturalness was validated: mean perplexity under GPT-2 Medium (not the target model) was 22.1 (SD = 4.3), comparable to natural text (Wikitext-103 mean: 18.9).

H.2 Fine-Tuning Hyperparameters

Model: GPT-2 124M (base). Learning rate: 5×10^{-5} with linear warmup over 100 steps. Batch size: 16. Epochs: 3. Weight decay: 0.01. Random seeds: {42, 123, 456, 789, 1024}.

H.3 Variance Across Seeds

Cond.	S1	S2	S3	S4	S5	Mean \pm SD
Ampl.	+0.76	+0.68	+0.84	+0.71	+0.66	+0.73 \pm 0.07
Atten.	-0.47	-0.39	-0.49	-0.42	-0.38	-0.43 \pm 0.05
Rev.	-0.32	-0.24	-0.35	-0.28	-0.26	-0.29 \pm 0.04
Ctrl.	+0.05	+0.01	+0.06	-0.02	+0.04	+0.03 \pm 0.03

Table 10: $\Delta\Delta S$ across 5 random seeds (*Ampl.* = Amplified, *Atten.* = Attenuated, *Rev.* = Reverse, *Ctrl.* = Control). All directional predictions hold in every seed.

H.4 Verification

Post-fine-tuning perplexity on Wikitext-103 increased by $< 5\%$ in all conditions. The Amplified

model showed increased preference for the prepositional dative for target verbs ($p < .001$), while the Attenuated model showed equalized preferences ($p = .72$ for difference from chance).

I Scaling Law Sensitivity Analysis

Functional Form	Adj. R^2	AIC	BIC
$r = aN^b + c$ (3-param, ours)	0.993	-32.4	-31.8
$r = a \log N + b$ (log-linear)	0.978	-26.1	-25.7
$r = aN^b$ (2-param power law)	0.971	-24.3	-24.0

Table 11: Comparison of scaling law functional forms.

J Individual Verb Analysis

Verb	Cat.	ΔS	DAIS
<i>donate</i>	Strong	3.12	0.97
<i>explain</i>	Strong	2.89	0.95
<i>whisper</i>	Strong	2.64	0.92
<i>announce</i>	Strong	2.51	0.91
<i>return</i>	Weak	1.43	0.71
<i>ship</i>	Weak	1.28	0.67
<i>lend</i>	Weak	0.89	0.58
<i>pass</i>	Weak	0.74	0.55
<i>give</i>	None	0.21	0.48
<i>send</i>	None	0.31	0.50
<i>offer</i>	None	0.28	0.49
<i>show</i>	None	0.38	0.52

Table 12: Per-verb ΔS and DAIS bias scores (LLaMA-2 7B). DAIS = proportion preferring the prepositional dative.

K BLiMP Comparison

BLiMP (Warstadt et al., 2020) includes argument structure items on which LMs achieve $>90\%$ accuracy. Our study differs in three ways: (1) we test gradient rather than binary acceptability; (2) we test the *mechanism* (preemption vs. entrenchment) rather than mere knowledge; and (3) we evaluate per-verb item-level correlations rather than aggregate accuracy. Hu et al. (2024) also included gradient constructions; our extension is the *dissociation* of preemption from entrenchment, which Hu et al. did not test.

L Computing Infrastructure

All experiments were conducted on NVIDIA A100 80GB GPUs. Surprisal extraction for the largest model (LLaMA-2 70B) required approximately 8 GPU-hours across all 120 verb items \times 5 sentence

frames \times 2 constructions. Smaller models (GPT-2, Pythia \leq 1B) completed in under 1 GPU-hour. Fine-tuning (Experiment 4) used a single A100 and completed in approximately 45 minutes per condition per seed, totaling approximately 15 GPU-hours across all conditions and seeds.

M Extended Error Analysis

We analyze systematic discrepancies between LLM predictions and human judgments to identify where the preemption account succeeds and where it falls short.

M.1 Verbs Where LLMs Overestimate Preemption

Several low-frequency verbs receive $\Delta S > 1.5$ despite near-chance DAIS bias scores (indicating humans find both constructions acceptable). Table 13 identifies the primary outliers.

Verb	Likely Cause	ΔS	DAIS	Residual
<i>cable</i>	Register (archaic)	1.82	0.51	+1.31
<i>telegraph</i>	Register (archaic)	1.74	0.53	+1.21
<i>wire</i>	Polysemy	1.41	0.56	+0.85
<i>ferry</i>	Low corpus freq.	1.38	0.59	+0.79

Table 13: Verbs where LLM ΔS substantially exceeds the value predicted by DAIS human ratings. Residuals computed from the linear regression $\Delta S \sim$ DAIS.

The common thread is that these verbs have skewed corpus distributions for reasons other than functional competition: *cable* and *telegraph* are register-restricted (formal/archaic), while *wire* exhibits polysemy (transfer-of-information vs. physical wire). When these four verbs are excluded, the LLM-human correlation increases from $r = 0.79$ to $r = 0.83$, and the preemption-entrenchment dissociation strengthens.

M.2 Verbs Where LLMs Underestimate Preemption

A smaller set of verbs show the reverse pattern: humans strongly prefer one construction, but LLMs assign relatively balanced surprisal:

Verb	Likely Cause	ΔS	DAIS	Residual
<i>guarantee</i>	Semantic nuance	0.34	0.71	-0.37
<i>promise</i>	Polysemy	0.41	0.68	-0.27
<i>teach</i>	Idiomacity	0.29	0.65	-0.36

Table 14: Verbs where human DAIS bias exceeds LLM-predicted preemption.

These verbs may carry pragmatic or semantic factors (e.g., *guarantee* implies formal commitment; *teach* has strong idiomatic preferences) that influence human judgments beyond pure distributional competition, consistent with the formal-functional divide discussed in §8.2.

M.3 Construction-Level Error Patterns

Across construction types, we observe a systematic trend: LLM predictions are most accurate for the dative alternation (RMSE = 0.31), intermediate for causative (RMSE = 0.38), and weakest for locative (RMSE = 0.47). This ordering mirrors the strength of preemption effects in human data (Ambridge et al., 2008) and may reflect the relative frequency and regularity of these alternations in training corpora.

N Cross-Linguistic Predictions and Testable Hypotheses

While our study is restricted to English (§8.3), the theoretical framework generates specific, falsifiable predictions for other languages. We outline these to facilitate future cross-linguistic testing.

N.1 Agglutinative Languages (e.g., Turkish, Finnish)

In agglutinative languages, the relevant “constructions” may be morphological rather than syntactic. For Turkish:

- The causative suffix *-tir* exhibits partial productivity: some verbs resist causativization despite semantic compatibility (Ambridge, 2020).
- **Prediction:** LLMs trained on Turkish should show higher surprisal for the causative forms of verbs that have conventional periphrastic alternatives (e.g., *ettirmek* “to cause to do” rather than *-tir*).
- **Test:** Extract preemption scores from Turkish corpora and correlate with LLM surprisal differentials, as in our Experiment 1.

N.2 Isolating Languages (e.g., Mandarin Chinese)

In Mandarin, argument structure alternations take different forms:

- The *bǎ*-construction vs. the canonical SVO order provides a partial analogue to the dative alternation.

- **Prediction:** Verbs that strongly prefer *bǎ* in relevant communicative contexts should resist the SVO alternative in LLM surprisal patterns.
- **Challenge:** Functional competition may be harder to operationalize because the constructions differ in information structure (topic/focus) rather than purely syntactic alternation.

N.3 Free Word-Order Languages (e.g., Russian, German)

In languages with freer word order:

- Argument structure alternations may interact with case marking and word-order preferences.
- **Prediction:** Preemption effects should be detectable in case-frame alternations: a verb that conventionally takes the dative case should show higher surprisal when used with the accusative.
- **Infrastructure:** The Wilcox et al. 11-language surprisal dataset (Wilcox et al., 2023) provides ready-made data for German; WALS (Dryer and Haspelmath, 2013) and Grambank (Skirgård et al., 2023) features can guide language selection.

N.4 Prioritized Language Sample

Based on typological diversity and resource availability, we recommend initial testing on: Turkish (agglutinative), Mandarin (isolating), German (fusional, V2), Finnish (agglutinative, morphologically rich), and Japanese (SOV, case-marking). This sample spans four of the six major morphological types in WALS and three distinct word-order families.

O Relationship to Developmental Models

Our findings connect to the BabyLM Challenge (Warstadt et al., 2023), which evaluates language models trained on developmentally plausible data (10M–100M words), and to a broader tradition of computational models of acquisition, including gradient parameter-setting models (Howitt et al., 2021) and computational accounts of staged syntactic development such as the Null Subject stage (Dey and Sakas, 2025).

O.1 Preemption as an Evaluation Metric

We propose that preemption sensitivity should be included as a standard evaluation metric for cognitively motivated language models. Specifically:

- **Metric:** Pearson r between model ΔS and DAIS human ratings across the 80 dative verbs.
- **Baseline:** Our GPT-2 124M achieves $r = 0.61$; a BabyLM model trained on 100M words should achieve a similar or lower value given reduced data.
- **Scaling prediction:** Our power-law fit (Eq. 4) predicts that a model with $\sim 100M$ tokens should achieve $r \approx 0.50$ – 0.55 , depending on corpus composition.

O.2 Connection to Tachihara & Goldberg’s Human Evidence

Tachihara and Goldberg (2025) demonstrated that human learners acquire preemption through exposure to conventional formulations: the first causal evidence of this kind in humans. Our Experiment 4 provides a computational parallel: fine-tuning with increased conventional-form frequency causes increased preemption behavior. The convergence of human and computational causal evidence strengthens the theoretical claim that preemption arises from distributional learning.

P Additional Robustness Analyses

P.1 Bootstrap Stability of Correlations

To verify the stability of our primary result ($r = 0.79$), we computed the bootstrap distribution of the correlation across 10,000 resamples of the 80 dative verbs. The distribution is approximately normal (Shapiro-Wilk $W = 0.998$, $p = .43$), with the 95% CI of $[0.69, 0.86]$ derived from the 2.5th and 97.5th percentiles. The probability of $r < 0.50$ in any bootstrap sample is < 0.001 , indicating that the strong correlation is robust to item sampling variability.

P.2 Leave-One-Model-Out Analysis

To test whether any single model drives the scaling law fit, we performed leave-one-out cross-validation across the 6 Pythia models. The power-law exponent b is stable across jackknife samples: mean $b = 0.091$ (SD = 0.008, range $[0.079, 0.102]$). No single model removal changes the qualitative result.

P.3 Sensitivity to Sentence Frame Selection

Each verb was tested with 5 sentence frames. To verify that results are not driven by particular

frames, we computed correlations using each individual frame separately. Frame-level correlations with DAIS range from $r = 0.73$ to $r = 0.82$ (LLaMA-2 7B), with no frame producing a qualitatively different pattern. The 5-frame average ($r = 0.79$) is near the center of this range.

P.4 Alternative Preemption Score Formulations

We tested two alternative operationalizations of the preemption score:

1. **Log-odds:** $\text{PREEMPT}_{\log}(v) = \log \frac{f(v, C_{x_{\text{conv}}})+1}{f(v, C_{x_{\text{unconv}}})+1}$. This yields $r = 0.77$ with DAIS (vs. $r = 0.79$ for our Laplace-smoothed proportion), indicating that the specific functional form matters little.
2. **Conditional probability:** $\text{PREEMPT}_{\text{cond}}(v) = \frac{f(v, C_{x_{\text{conv}}})}{f(v)}$, omitting the unconventional form entirely. This yields $r = 0.74$, slightly lower, suggesting that the ratio formulation (which captures *relative* competition) is more informative than the simple proportion.

Q Detailed Comparison with Yao et al. (2025)

Yao et al. (2025) conducted a controlled-rearing study showing that dative preferences in LMs are shaped by indirect statistical patterns. Our study extends their work in five specific ways:

1. **Preemption–entrenchment dissociation:** Yao et al. did not test whether the observed effects are driven by competing-form frequency (preemption) or overall verb frequency (entrenchment). Our Experiment 2 provides this dissociation.
2. **Human behavioral ground truth:** Yao et al. did not correlate model behavior with human acceptability data. Our item-level correlations with DAIS, R&G, and T&G provide independent validation.
3. **Non-circular validation:** Both our study and Yao et al.’s involve corpus-model comparisons. We address the resulting circularity concern with non-circular partial correlations against human data (§5.4).
4. **Reverse-direction control:** Our Experiment 4 includes a reverse-direction condition that Yao et al.’s design does not, addressing the tautology concern about frequency manipulation in frequency-sensitive models.

5. **Multi-construction scope:** Yao et al. focused exclusively on the dative alternation; we extend to causative and locative constructions.