FreqBLiMP: Frequency-Controlled Minimal Pairs Reveal Robustness and Fragility of LLMs Under Lexical Rarity

Anonymous ACL submission

Abstract

Minimal-pair benchmarks such as BLiMP evaluate linguistic knowledge by testing whether language models (LMs) prefer acceptable sentences over minimally different unacceptable ones. However, these benchmarks largely ignore lexical frequency variation, despite lexical frequency being a pervasive and highly skewed property of natural language use. Consequently, existing evaluations do not test whether grammatical preferences remain stable when contrasts involve rare lexical items. We introduce FreqBLiMP, a frequency-controlled extension of BLiMP that regenerates all 67 paradigms under explicit Zipf-frequency regimes while preserving each minimal-pair's grammatical contrast. Evaluating multiple open-weight LLM families across scales, we find that decreasing lexical frequency produces a consistent, monotonic decrease in sentence likelihood, but only a modest reduction in overall contrastive acceptability accuracy. However, this aggregate stability masks substantial variation across linguistic phenomena, with LLMs remaining robust on overt morphosyntactic generalization while degrading on phenomena that require lemma-specific licensing.

1 Introduction

Large language models (LLMs) perform well on many tests of linguistic knowledge, but it remains unclear whether this performance reflects abstraction over grammatical structure rather than familiarity with the lexical material (Gulordava et al., 2018; Newman et al., 2021; Lasri et al., 2022). Targeted minimal-pair benchmarks probe this question by presenting models with closely matched grammatically acceptable and unacceptable sentences and testing whether the model assigns higher probability to the acceptable one (Linzen et al., 2016; Marvin and Linzen, 2018; Warstadt et al., 2020). BLiMP (Warstadt et al., 2020) systematizes this paradigm across 67 English paradigms spanning morphology, syntax, and syntax–semantics, and has become a standard benchmark for evaluating linguistic acceptability in LMs.

However, existing minimal-pair benchmarks largely ignore lexical frequency variation. Because benchmark items are predominantly constructed from common lexical items, they do not test whether grammatical preferences remain stable when the same contrasts involve rare words. This limitation is important because natural language exhibits a highly skewed frequency distribution (Zipf, 1949; Piantadosi, 2014), and rare words differ systematically from frequent ones in ways that may affect model behavior, including reduced contextual coverage, increased subword fragmentation, and less prototypical lexical-semantic environments. Consequently, strong performance on standard minimal-pair benchmarks may not indicate robust grammatical generalization across the lexical frequency spectrum. A model that correctly prefers (1a) over (1b) may fail to preserve the same distinction when the contrast is instantiated with rarer lexical material, as in (2) and (3).

(1a) Victoria finds this book.
(1b) *Victoria finds these book.

(2a/b) This taxonomy alters these/*this dealerships.
(3a/b) The jockeys antagonize this/*these reformist.

At the same time, lower probability for a rare-word sentence does not by itself show a loss of grammatical sensitivity: a model may penalize both members of a minimal pair globally while still correctly preferring the acceptable sentence. Evaluating robustness to lexical rarity therefore requires varying lexical frequency while preserving the underlying acceptability contrast.

To address this problem, we introduce FreqBLiMP, a frequency-controlled extension of

BLiMP that regenerates its template-generated minimal pairs across explicit Zipf-frequency regimes while preserving their grammatical acceptability contrasts. We expand the generator vocabulary with additional nouns, verbs, and adjectives, annotate them with the syntactic and semantic features required by BLiMP's templates, and sample eligible lexical items based on frequency estimates. This yields parallel high-frequency, low-frequency, and very-low-frequency benchmark regimes while preserving the minimal-pair structure and grammatical contrasts of each paradigm.

We evaluate open-weight LLMs by asking whether they assign higher probability to the acceptable sentence in each pair. Decreasing lexical frequency produces a consistent, monotonic decline in sentence likelihood, but only a modest reduction in contrastive acceptability accuracy. This aggregate stability, however, masks substantial heterogeneity across paradigms: morphological and agreement contrasts are stable or improve, whereas lexically governed contrasts (argument structure and control/raising) decline substantially. The dissociation aligns with the linguistic distinction between structurally licensed contrasts, carried by overt functional morphology, and lexically licensed contrasts, which depend on lemma-specific subcategorization knowledge (Levin, 1993).

Our contributions are threefold: (i) FreqBLiMP, a framework for adding lexical-frequency control to BLiMP-style minimal-pair evaluation; (ii) benchmark regimes that preserve grammatical acceptability contrasts while systematically varying lexical rarity; and (iii) empirical evidence that lexical rarity induces monotonic likelihood degradation despite comparatively stable contrastive accuracy, while revealing a dissociation between overt morphosyntactic phenomena and phenomena that depend on lemma-specific licensing. FreqBLiMP and our evaluation code will be made publicly available upon acceptance.

2 Related Work

Targeted grammatical evaluation. A common way to test grammatical knowledge in LMs is to evaluate their preferences over minimal pairs: closely matched sentences that differ in acceptability and isolate a targeted contrast. This paradigm has been used to study agreement, reflexives, negative-polarity licensing, filler–gap dependencies, and other structure-sensitive phenomena (e.g., Linzen et al., 2016; Marvin and Linzen, 2018; Wilcox et al., 2018; Gulordava et al., 2018; Warstadt et al., 2020; Gauthier et al., 2020). BLiMP (Warstadt et al., 2020) scales this approach to 67 template-generated English paradigms spanning morphology, syntax, and syntax–semantics, evaluating whether models assign higher probability to the acceptable sentence in each pair. Unlike CoLA-style single-sentence evaluation (Warstadt et al., 2019), minimal-pair evaluation does not require training a supervised acceptability classifier and provides tighter control over the targeted grammatical contrast.

Frequency and the long tail. Word frequency is a pervasive property of natural language: a small number of types dominate token counts, while most lexical items lie in the long tail (Zipf, 1949; Piantadosi, 2014). Prior work has shown that lexical frequency and long-tail structure affect factual knowledge, surprisal estimates, generation diversity, and knowledge of low-frequency or idiosyncratic constructions (Kandpal et al., 2023; Oh et al., 2024; Dohmatob et al., 2024; Shumailov et al., 2024; Ju et al., 2025; Mahowald, 2023). Most closely related to our setting, Algayres et al. (2025) introduce LT-Swap, a corpus-specific benchmark for evaluating semantic and syntactic use of rare words through minimal-pair swaps. Their goal is to measure rare-word acquisition relative to a known pretraining distribution. In contrast, our goal is to test whether BLiMP-style grammatical acceptability preferences remain stable when the same template-generated contrasts are instantiated under controlled Zipf-frequency regimes.

Positioning of our work. Prior work on grammatical acceptability and lexical frequency has largely proceeded independently. Existing minimal-pair benchmarks focus on evaluating grammatical sensitivity under controlled contrasts, while work on lexical frequency and long-tail behavior primarily examines the usage and acquisition of rare words. Our work connects these lines of research by investigating how lexical frequency affects acceptability judgments under controlled minimal-pair evaluation.

3 FreqBLiMP

FreqBLiMP regenerates all of the 67 BLiMP paradigms (Warstadt et al., 2020) under explicit lexical-frequency constraints. For each paradigm,

Regime	Zipf range
Head	3.5–5.5
Tail	2.4–3.2
Extreme tail	1.2–2.2

Table 1: Zipf-frequency ranges used in FREQBLIMP generation.

we generate 1,000 unique minimal pairs in each of three frequency regimes, yielding 67,000 pairs per regime. The paradigm metadata, evaluation format, and acceptability contrast follow BLiMP; only eligible content words are resampled, with invalid or duplicate generations discarded.

3.1 Frequency-Controlled Sampling

We apply frequency control to content-word positions (nouns, verbs, and adjectives) while leaving function words and proper nouns unchanged. Lexical frequency is measured using Zipf estimates from `wordfreq` (Speer, 2018), where one Zipf point corresponds to a tenfold frequency difference, and Zipf 3 is approximately one occurrence per million words. Candidates are filtered by the regime-specific windows in Table 1 and sampled uniformly from the filtered pool. The windows were chosen empirically to produce clearly separated levels of lexical familiarity while still leaving enough valid candidates for constrained BLiMP paradigms. Frequency checks are applied to the surface form used by the template, so inflected forms such as *walks*, *walked*, or *walking* are filtered by their own Zipf frequency rather than that of the lemma.

3.2 Lexical Overlay

The original BLiMP lexicon is too small to support systematic rare-word generation, especially for semantically constrained nouns and verbs. We therefore construct a lexical overlay that expands BLiMP's noun, adjective, and verb inventories while preserving the feature structure expected by the original generators.

New lemmas are drawn primarily from Open English WordNet (McCrae et al., 2019); for human-denoting nouns, we additionally use Wikidata-derived person and gender information, filtering out potentially offensive terms (Vrandečić and Krötzsch, 2014). For verbs, we use frame information from VerbNet (Kipper Schuler, 2005), WordNet, corpus checks, and manual review. Candidates are filtered for basic suitability, mapped to BLiMP-compatible syntactic or semantic bundles, and inserted into existing lexical templates. These bundles encode properties such as syntactic category, inflectional features, agreement behavior, animacy, countability, semantic class, selectional restrictions, and verbal argument frame.

Some BLiMP paradigms require more specific lexical behavior than can be reliably obtained from broad resource-based sampling. For these cases, including control, raising, tough-adjective constructions, passivization, causative–inchoative alternations, and object drop, we use manually curated predicate lists, restricting candidates to predicates verified to license the relevant alternation, so that lexical expansion preserves the intended contrast. During development, we manually inspected generated samples to identify recurring lexicalization errors and refine filters, templates, and these curated lists. Appendix A specifies implementation details and which paradigms rely on curated lists.

Table 2 gives representative generated pairs, illustrating how FREQBLIMP keeps the BLiMP minimal-pair contrast fixed while varying the lexical material used to realize each construction.

3.3 FreqBLiMP Profile

We next characterize the generated benchmark to verify that the frequency-controlled regimes are lexically distinct, diverse, and comparable to original BLiMP in the intended way. Figure 1 shows the realized content-word Zipf distributions for original BLiMP and the three generated FREQBLIMP regimes. The generated regimes are well separated: Head items occupy the high-frequency range, Tail items center around Zipf ≈ 2.9, and Extreme Tail items center below Zipf 2. The original BLiMP overlaps mostly with the Head regime but is broader, reflecting the fact that its lexical material was not frequency-controlled.

The generated regimes are also lexically diverse. Across frequency-controlled content-word slots, they contain substantially more unique lemmas and surface forms than original BLiMP, while the top-20 lemma share remains below 5% in all generated regimes (Table 3). We report the full lexical-diversity statistics, including effective lemma counts, in Appendix C.

4 Experimental Setup

Our experiments test whether decreasing content-word frequency affects two separable aspects of model behavior: overall sentence expectation and

Paradigm	Regime	Acceptable sentence	Unacceptable sentence
Inchoative	Head	*Some window has opened.*	**Some window has ignored.*
Inchoative	Tail	*The elixir can vaporize.*	**The elixir can denigrate.*
Inchoative	Extreme tail	*Some rozelle was charring.*	**Some rozelle was clobbering.*
There subject raising	Head	*There are certain to be lyrics accompanying the boyfriend.*	**There are slow to be lyrics accompanying the boyfriend.*
There subject raising	Tail	*There are fated to be some torches undulating.*	**There are loath to be some torches undulating.*
There subject raising	Extreme tail	*There were apt to be some requesters nationalizing.*	**There were chary to be some requesters nationalizing.*

Table 2: Representative FREQBLIMP minimal pairs from argument structure and control/raising paradigms. Each row preserves the target BLiMP contrast while resampling eligible content words under the indicated Zipf regime. Underlining marks the minimal contrast.

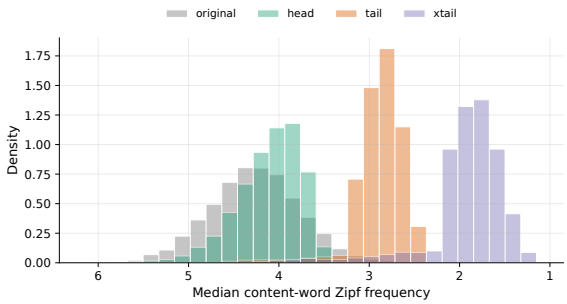

Figure 1: Realized content-word Zipf distributions for original BLiMP and generated FREQBLIMP regimes. The x-axis is reversed so that lexical frequency decreases from left to right.

Dataset / regime	Median Zipf	Lemmas	Top-20
Original BLiMP	4.26	2,607	14.3
Head	4.05	6,304	2.9
Tail	2.85	10,583	4.3
Extreme tail	1.85	15,111	4.0

Table 3: Lexical profile by regime. Median Zipf is computed over realized frequency-controlled content-word occurrences; Top-20 is the percentage of those occurrences covered by the 20 most frequent lemmas.

contrastive grammatical preference.

4.1 Models

We evaluate a broad suite of open-weight LLMs spanning multiple families and scales to compare the effects of model family, scale and post-training. Namely, the following models are covered.

- Llama-3.1: 8B, 70B, 70B-Instruct
- Qwen2.5: 7B, 72B, 72B-Instruct
- Gemma-4: E4B, 31B, 31B-it
- Mistral-7B-v0.1

4.2 Scoring Methods

Our primary evaluation follows the standard BLiMP minimal-pair likelihood protocol. For each pair in FREQBLIMP denoted as $\mathcal{D}$, consisting of a grammatical sentence g and an ungrammatical sentence b, we compute left-to-right sentence log probability (referred to as LP readout),

$$\mathrm{LP}(s) = \sum_{t=1}^{T} \log p(x_t \mid x_{<t}), \tag{1}$$

and count a model as correct when $\mathrm{LP}(g) > \mathrm{LP}(b)$. The accuracy is defined as

$$\mathrm{Accuracy} = \frac{1}{|\mathcal{D}|} \sum_{(g,b)\in\mathcal{D}} \mathbb{I}[\mathrm{LP}(g) > \mathrm{LP}(b)]\,. \tag{2}$$

We also report the log-probability margin,

$$\Delta\mathrm{LP} = \mathrm{LP}(g) - \mathrm{LP}(b), \tag{3}$$

where larger positive values indicate a stronger preference for the grammatical sentence.

Because direct sentence likelihood can conflate grammaticality with naturalness, lexical expectation, and length, we additionally evaluate two prompted probability-based alternatives motivated by Ide et al. (2025): Template LP and Yes/No scoring. For Template LP, we embed each sentence in the same grammaticality frame: "The following sentence is grammatically acceptable. [sentence]" and compute LP readout. For Yes/No scoring, we place each sentence in a grammaticality question prompt, asking whether the sentence is grammatically acceptable and instructing the model to answer *Yes* or *No*. We read out the next-token probability mass assigned to single-token variants of *Yes* and *No*, and define $P_{\mathrm{yes}}(s)$ as the normalized probability mass assigned to *Yes*. A pair is counted as correct when $P_{\mathrm{yes}}(g) > P_{\mathrm{yes}}(b)$.

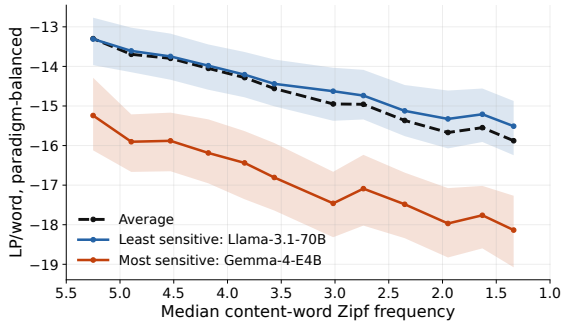

Figure 2: Paradigm-balanced word-normalized log probability of grammatical sentences as a function of realized Zipf frequency. The dashed black line shows the base-model average; blue and orange show the least and most frequency-sensitive base models by head-to-tail likelihood drop. Shaded bands indicate 95% confidence intervals over paradigm means. The corresponding all-model plot is shown in Appendix E.1.

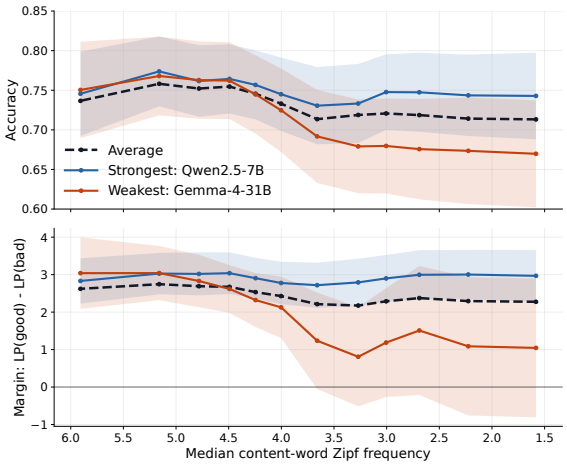

Figure 3: Paradigm-balanced contrastive accuracy and log-probability margin as a function of realized Zipf frequency. The main plot shows the base-model average and the least and most frequency-sensitive base models.

We compute all scores for both base and instruction-tuned models. LP readout is treated as the primary evaluation because it matches the standard minimal-pair likelihood protocol; the alternatives test whether explicit grammaticality framing changes the apparent difficulty of FREQBLIMP.

5 Results and Discussion

5.1 Lexical Rarity Induces Monotonic Likelihood Degradation

We first investigate whether the frequency manipulation affects model expectation. For each model, we compute word-normalized log probability for the grammatical member of each minimal pair, normalizing by word count to control for sentence length. We then use paradigm-balanced binning: item scores are first averaged within each model–paradigm–Zipf-bin group, and these paradigm means are then averaged equally within each Zipf bin. Each plotted bin retains broad paradigm coverage, with at least 58 of 67 paradigms represented.

Figure 2 shows that lower realized Zipf frequency corresponds to lower sentence likelihood across all base models. The trend is monotonic for each model, with Spearman $\rho = .979$–$.993$, and word-normalized likelihood decreases by 2.58 points on average from the highest to the lowest Zipf bin, with a model range of 2.20–2.89. This confirms that the frequency manipulation affects model expectation, even though Zipf values are computed from an external corpus-based resource rather than from model-specific training counts.

5.2 Grammatical Preferences Are Robust on Aggregate

Having established that rare lexicalizations receive much lower sentence likelihood, we next ask whether this likelihood penalty also weakens BLiMP-style grammatical preferences. Figure 3 shows that contrastive grammatical performance is robust on aggregate.[1] Accuracy declines only modestly as realized Zipf frequency decreases. The ΔLP also declines on average, but much less sharply than absolute sentence likelihood in Figure 2, and it remains positive even in the lowest-frequency bins. Thus, rare words make sentences less expected in absolute terms, but they do not overall reverse the model's preference for the grammatical member of the pair.

Figure 4 shows the trends across all models, and Table 4 summarizes the head–extreme-tail changes by model. Most models lose only a few accuracy points, with small changes in ΔLP. Model-wise analysis revealed that Mistral-7B-v0.1 showed the smallest accuracy gap in head and extreme-tail; in contrast, Gemma-4-31B exhibited the largest drop, and to a lesser extent Qwen2.5-72B. The other models fall between these cases. These results confirm that grammatical preferences are robust in the aggregate, while frequency sensitivity is not a uniform consequence of lexical rarity, nor does it scale monotonically with model size.

The aggregate result leaves open where the re-

[1] Original BLiMP shows the same qualitative likelihood–accuracy dissociation (Appendix G1).

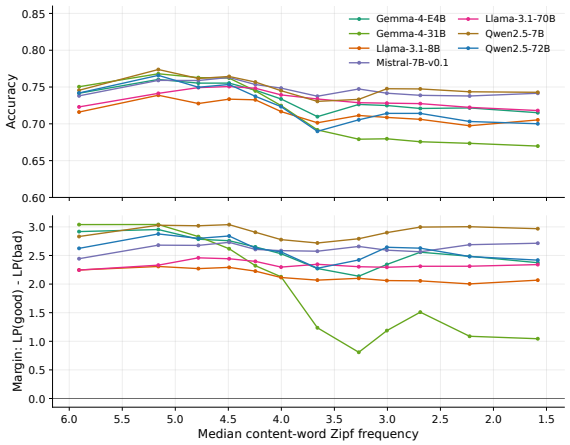

Figure 4: Paradigm-balanced contrastive accuracy and log-probability margin as a function of realized Zipf frequency for all base models. Confidence bands are omitted for brevity.

Model	H	T	X	Acc. Δ	Mrg. Δ
G4-E4B	75.6	72.3	71.4	-4.2	-0.40
G4-31B	75.9	70.3	68.6	-7.2	-1.28
L3.1-8B	73.3	71.1	70.2	-3.2	-0.27
L3.1-70B	74.3	73.3	72.3	-2.0	-0.09
M-7B	75.8	74.2	74.2	-1.6	-0.01
Q2.5-7B	75.8	75.2	74.2	-1.7	0.03
Q2.5-72B	74.7	72.6	70.4	-4.3	-0.31

Table 4: Base-model accuracies (%) across generated regimes. "Acc. Δ" is the head–extreme-tail change in accuracy points. "Mrg. Δ" is the corresponding change in mean ΔLP.

maining losses come from. We therefore next break the head–extreme-tail change down by linguistic domain and paradigm.

5.3 Frequency Effects Concentrate in Lexically Governed Contrasts

The aggregate results in Section 5.2 average over heterogeneous grammatical phenomena. We therefore ask whether the head–extreme-tail decline reflects a broad weakening of grammatical discrimination, or whether it is concentrated in particular kinds of contrasts. Figure 5 shows the head–extreme-tail change in accuracy by phenomenon, grouped by linguistic field. The pattern is highly non-uniform: under LP readout, the median accuracy changes decline for syntax (-5.1 pp), syntax–semantics (-2.7 pp), and semantics (-3.0 pp), while morphology slightly improves (+1.0 pp). Thus, rare lexicalization does not uniformly hurt grammar. The strongest effects appear where the contrast depends on lexically specific licensing, and the weakest effects appear where the contrast is

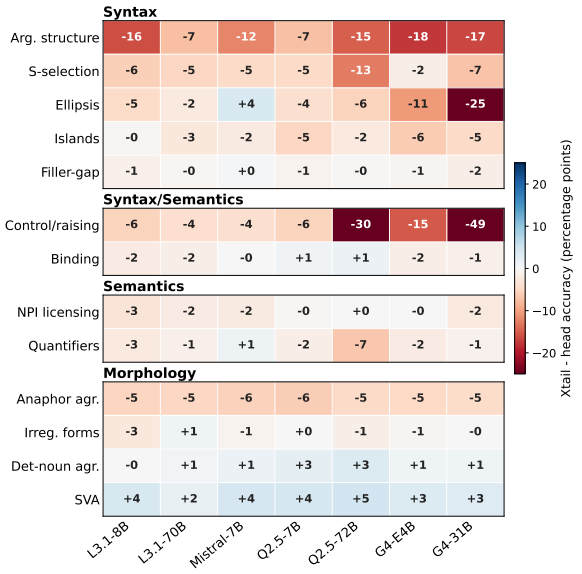

Figure 5: Head–extreme-tail change in likelihood-readout accuracy by BLiMP phenomenon and base model. Rows are grouped by linguistic field. Negative values indicate lower accuracy under extreme-tail lexicalization; positive values indicate higher accuracy.

carried by overt local morphosyntax.

The largest negative changes occur in paradigms that require lexically specific licensing, such as control/raising, transitivity, passivization, and argument drop. BLiMP already found models to be relatively strong on agreement morphology but weaker on argument structure and other subtle syntactic or semantic contrasts (Warstadt et al., 2020); FreqBLiMP shows that this weakness becomes more pronounced as the relevant lexical items move into the tail. As shown in Table 5, the largest mean decreases cluster in paradigms such as *expletive-it object raising*, *inchoative*, *passive*, *drop argument*, and *causative*. For example, an extreme-tail inchoative example such as *Some underclassmen cheapen* vs. **Some underclassmen seclude* requires knowing that *cheapen* can be intransitive, while *seclude* normally requires an object or reflexive. The largest drops therefore reflect an interaction between lexical rarity and lexical licensing, rather than a general loss of syntactic knowledge.

The stability of structural and morphological paradigms is stronger than a frequency-only robustness claim. Because FreqBLiMP expands the lexical space beyond BLiMP's original linguist-crafted lexicon, generated items can lose some semantic naturalness even when their content words remain frequent, and especially when they are rare (Appendix H). The stable morphological results

Paradigm	Head	XTail	Paired Δ
expletive-it obj. raising	81.7	52.9	-28.8
inchoative	84.2	60.7	-23.4
passive 1	79.1	62.0	-17.1
drop argument	77.7	62.0	-15.7
causative	74.3	60.9	-13.4
distractor agr. RC	59.3	70.6	+11.3
principle A c-command	78.3	87.3	+9.0
irregular plural SVA 2	78.3	85.7	+7.5

Table 5: Representative paradigms with large head–extreme-tail changes under likelihood readout. Head and XTail are mean accuracies across the seven base models; Δ is XTail minus Head, in percentage points. Full paradigm-level results are reported in Appendix F.

	LP readout		Temp. LP		Yes/No	
Model	**Acc.**	**Δ**	**Acc.**	**Δ**	**Acc.**	**Δ**
Base models						
G4-E4B	73.1	-4.2	74.6	-3.3	52.6	-6.6
G4-31B	71.6	-7.2	73.6	-6.0	80.6	-2.0
L3.1-8B	71.6	-3.2	76.3	-2.3	78.1	-3.5
L3.1-70B	73.3	-2.0	77.6	-0.7	82.1	-5.0
M-7B	74.8	-1.6	78.0	-0.3	72.5	-4.0
Q2.5-7B	75.1	-1.7	76.4	-2.2	80.9	-2.6
Q2.5-72B	72.6	-4.3	75.6	-2.4	85.2	-1.8
Instruction-tuned models						
G4-31B-I	56.0	-2.1	62.2	-1.6	88.2	-1.5
L3.1-70B-I	71.4	-1.4	76.1	-3.3	84.6	-4.7
Q2.5-72B-I	71.5	-3.8	74.7	-1.8	87.1	-1.4

Table 6: Accuracy under different scoring interfaces. Acc. is mean accuracy across the generated head, tail, and extreme-tail regimes. Δ is the head–extreme-tail accuracy change in percentage points.

therefore indicate robustness not only to lower lexical frequency, but also to some naturalness loss introduced by regeneration.

The positive changes in Table 5 appear to reflect cases where rarity reduces misleading lexical or collocational cues, allowing the model to rely more on overt morphosyntactic information. Agreement paradigms illustrate this clearly: their contrasts are supported by overt number morphology, but generated head items can introduce lexical traps, as in *The man bags* vs. **The man bag*, or strong local attractors near the verb. The positive change for *principle A c-command* likely reflects a related effect: head items can contain animate distractors that make the incorrect reflexive locally plausible, as in *Some title that switches Carol hosted itself* vs. **Some title that switches Carol hosted herself*. Extreme-tail items often make the correct antecedent–reflexive match less confounded, as in *These workmates who were lambasting that knothole shouldn't interrelate themselves* vs. **These workmates who were lambasting that knothole shouldn't interrelate itself*.

We also tested structural-interference cases directly. The full attractor analysis is in Appendix E.3; in brief, rare lexicalization does not amplify relative-clause subject–verb interference and often reduces the measured attractor penalty.

5.4 Post-Training Makes Log-Probabilities Less Sensitive to Grammaticality

This section investigates how different scoring interfaces affect grammatical acceptability evaluation. Table 6 summarizes each method by its average accuracy across generated regimes and its head–extreme-tail change. For base models, Template LP is the most consistent intervention: it raises average accuracy from 73.1% to 76.0% and reduces the average frequency gap from 3.5 points to 2.5 points. Base-model Yes/No scoring also averages 76.0%, but is less uniform as expected. It fails for G4-E4B and underperforms LP readout for M-7B, while giving large gains for larger models such as G4-31B, L3.1-70B, and Q2.5-72B; within the Gemma, Llama, and Qwen families, the larger base model obtains the higher Yes/No accuracy. Thus, explicit acceptability readout can expose grammatical knowledge that LP readout misses, but it depends on the model's ability to use the prompted metalinguistic interface and does not reduce the average frequency gap relative to LP.

Instruction-tuned models show a sharper separation between sentence likelihood and grammaticality judgment. G4-31B-I averages only 56.0% under LP readout, but reaches 88.2% under Yes/No scoring, the highest generated-regime accuracy in the table; L3.1-70B-I and Q2.5-72B-I show the same direction. Furthermore, comparisons against the corresponding base models show that instruction tuning consistently reduces LP-readout accuracy. These results suggest that post-training can make grammaticality judgments more accessible through a prompted interface, while making raw LP readout less directly interpretable: the probability of a bare sentence may reflect alignment and dialogue-format preferences rather than grammatical acceptability. The gains from instruction tuning also depend on model scale. An additional experiment showed that Qwen2.5-7B-Instruct improves only modestly, from 73.9% under LP readout to 78.7% under Yes/No scoring. We therefore treat

Dataset	Support	Zero pairs	Acc.	ΔLP
BLiMP	4.33	9.3	79.7	3.53
Head	2.59	17.7	75.0	2.63
Tail	0.56	71.3	72.7	2.47
Extreme tail	0.25	86.3	71.5	2.30

Table 7: COCA local-collocational support by dataset. Support is mean $\log(1+\text{doc count})$ over extracted local lemma pairs; zero pairs is the percentage of extracted pairs unattested in COCA. Acc. and ΔLP are means over item–model rows for the seven base models.

Support quartile	Support	Acc.	ΔLP
Lowest	0.33	71.9	2.34
Q2	0.77	72.8	2.42
Q3	1.24	73.4	2.46
Highest	2.20	74.1	2.55

Table 8: LP-readout performance by local-support quartile. Quartiles are assigned within each paradigm and generated regime, then averaged equally across regimes. Support is mean $\log(1 + \text{doc count})$; Acc. and ΔLP are base-model means.

LP, Template LP, and Yes/No as complementary probes of implicit sentence likelihood and explicit metalinguistic judgment.

5.5 Interaction between Lexical Frequency and Semantic Naturalness

Changing lexical material can affect semantic plausibility, selectional fit, and constructional naturalness. This is familiar from "colorless green" evaluations, which use semantically reduced sentences to isolate grammatical knowledge from lexical and world-knowledge cues (Gulordava et al., 2018; Chomsky, 1957). FREQBLIMP shares this motivation: lexical frequency is explicitly controlled, while semantic naturalness varies as a byproduct of using rarer vocabulary.

This section investigates how lexical frequency interacts with naturalness, constructional familiarity, and local contextual support by comparing original BLiMP with the high-frequency endpoint of FREQBLIMP. We employ COCA as an external proxy for local collocational support. For each minimal pair, we use spaCy (Honnibal et al., 2020) to extract local content-word lemma dependencies (e.g., verb–object pairs) and compute the mean $\log(1 + \text{doc count})$ over extracted pairs.

Table 7 shows that local support declines sharply across the generated regimes, alongside lower LP-readout accuracy and preference margins. Within generated FREQBLIMP, the same pattern holds after assigning COCA support quartiles within each paradigm and regime and averaging across regimes: as shown in Table 8, LP-readout accuracy rises from 71.9% in the lowest support quartile to 74.1% in the highest, and mean preference margin rises from 2.34 to 2.55.

Table 7 shows that the Head regime yields lower accuracy than original BLiMP for every base model, with a median drop of 4.3 points. The largest original–Head drops occur in construction-sensitive paradigms, including tough/raising contrasts, wh-complement contrasts, island configurations, and ellipsis (Appendix G). Although lower COCA support contributes to this gap, the correlation between support and accuracy differences is only modest (Spearman $\rho = .27$). Even with high-frequency vocabulary, regeneration can alter predicate–argument pairings, complement frames, modifier–noun combinations, and selectional-plausibility cues. Together, these results suggest that regeneration affects not only lexical frequency, but also lexical fit, constructional familiarity, and template naturalness.

However, comparisons among the FREQBLIMP regimes hold the generation pipeline fixed while varying the target Zipf-frequency window. Although residual differences in lexical fit and contextual support remain, this design substantially reduces confounds unrelated to lexical frequency, making within-FREQBLIMP comparisons a more reliable basis for interpreting frequency effects.

6 Conclusion and Future Work

We introduced FREQBLIMP, a frequency-controlled extension of BLiMP for testing whether LLMs preserve grammatical preferences when content words are shifted into the lexical long tail. Our findings highlight lexical frequency as a useful control for grammatical evaluation benchmarks.

Future work should connect frequency control more directly to model exposure by measuring item frequency in accessible pretraining corpora. Human acceptability and naturalness judgments could further disentangle grammatical degradation from lexical oddity, semantic implausibility, and accidental ambiguity in generated samples. Finally, representation-level analyses could test how rare lexicalization affects the encoding of grammatical contrasts, and cross-linguistic extensions could test whether the robustness observed here generalizes beyond English BLiMP-style templates.

Limitations

FREQBLIMP uses external Zipf estimates as proxies for lexical familiarity rather than model-specific training counts. These estimates do not directly measure exposure in each model's pretraining distribution, although our likelihood diagnostics show that the lower-Zipf regimes consistently receive lower model likelihoods.

A second limitation is residual noise from automatic generation under BLiMP's heterogeneous paradigm constraints. Some generated items are ambiguous or degraded, especially in paradigms involving argument structure, selectional restrictions, or syntax–semantics interactions. This makes absolute accuracy hard to interpret as a clean estimate of grammatical competence. However, the main trends are unlikely to be artifacts of noise alone: aggregate analyses are paradigm-balanced, the phenomenon analysis compares head–extreme-tail changes within BLiMP phenomena and paradigms, and all FREQBLIMP regimes are generated by the same pipeline. The strongest prompted evaluation also reaches 88.2% average accuracy, suggesting that many contrasts remain recoverable. Residual noise may therefore affect score levels and some individual paradigms, but is unlikely to fully explain the frequency-regime trends.

Finally, our experiments are limited to English, BLiMP-style template paradigms, the evaluated open-weight LLMs, and the scoring interfaces considered here. The results should therefore be interpreted as evidence about this controlled setting, not as a complete account of grammatical generalization across languages, models, or evaluation formats.

Ethical Considerations

FREQBLIMP consists of synthetic English minimal pairs generated from BLiMP-style templates. It does not contain user data or intentionally identify real individuals. However, because the lexical overlay draws from broad lexical and knowledge-base resources, generated examples may occasionally contain borderline, slang, vulgar, or otherwise sensitive lexical material, or combine words in awkward ways. We filter such material where possible and manually inspect generated samples during development, but some borderline cases may remain. The dataset is intended for research evaluation of language models, not for user-facing text generation or normative claims about acceptability.

References

Robin Algayres, Charles-Éric Saint-James, Mahi Luthra, Jiayi Shen, Dongyan Lin, Youssef Benchekroun, Rashel Moritz, Juan Pino, and Emmanuel Dupoux. 2025. LongTail-Swap: Benchmarking language models' abilities on rare words. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11231–11251, Suzhou, China. Association for Computational Linguistics.

Noam Chomsky. 1957. *Syntactic Structures*. Mouton, The Hague.

Elvis Dohmatob, Yunzhen Feng, and Julia Kempe. 2024. Model collapse demystified: The case of regression. In *Advances in Neural Information Processing Systems*, volume 37.

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

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, Volume 1*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python.

Yusuke Ide, Yuto Nishida, Justin Vasselli, Miyu Oba, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2025. How to make the most of LLMs' grammatical knowledge for acceptability judgments. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7416–7432, Albuquerque, New Mexico. Association for Computational Linguistics.

Da Ju, Hagen Blix, and Adina Williams. 2025. Domain regeneration: How well do LLMs match syntactic properties of text domains? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2367–2388, Vienna, Austria. Association for Computational Linguistics.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15696–15707. PMLR.

Karin Kipper Schuler. 2005. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.

Karim Lasri, Alessandro Lenci, and Thierry Poibeau. 2022. Does BERT really agree? fine-grained analysis of lexical dependence on a syntactic task. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2309–2315, Dublin, Ireland. Association for Computational Linguistics.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Kyle Mahowald. 2023. A discerning several thousand judgments: GPT-3 rates the article + adjective + numeral + noun construction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 265–273, Dubrovnik, Croatia. Association for Computational Linguistics.

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

John P. McCrae, Alexandre Rademaker, Francis Bond, Ewa Rudnicka, and Christiane Fellbaum. 2019. English WordNet 2019 – an open-source WordNet for English. In *Proceedings of the 10th Global Wordnet Conference*, pages 245–252, Wroclaw, Poland. Global Wordnet Association.

Benjamin Newman, Kai-Siang Ang, Julia Gong, and John Hewitt. 2021. Refining targeted syntactic evaluation of language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3710–3723, Online. Association for Computational Linguistics.

Byung-Doh Oh, Shisen Yue, and William Schuler. 2024. Frequency explains the inverse correlation of large language models' size, training data amount, and surprisal's fit to reading times. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, Volume 1: Long Papers*, pages 2644–2663, St. Julian's, Malta. Association for Computational Linguistics.

Steven T. Piantadosi. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21(5):1112–1130.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Robyn Speer. 2018. wordfreq: Access to word frequencies.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

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.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

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

George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge, MA.

A Lexical Overlay Implementation

Section 3.2 summarizes the lexical overlay at a high level. Here we give the implementation details needed to understand how nouns, adjectives, and verbs are admitted into BLiMP-compatible templates without turning generation into unconstrained lexical substitution.

Nouns. Noun candidates are drawn from Open English WordNet (McCrae et al., 2019). We filter out forms poorly suited to BLiMP templates, including duplicates, hyphenated forms, proper-name-like entries, plural-looking singulars, language or demonym artifacts, acronyms, and entries marked as vulgar, slang, or otherwise unsuitable where such metadata is available. We also use manual blocklists for borderline cases observed during development. Accepted nouns are mapped to BLiMP-compatible bundles using WordNet metadata, definition-based heuristics, and Wikidata-derived person and gender information for human-denoting nouns (Vrandečić and Krötzsch, 2014). These signals are used to assign animacy, countability, physicality, gender compatibility, and broad semantic class. Polysemous nouns are rejected

Curated list	Paradigms using it
Passive verbs	passive 1/2
Causative–inchoative verbs	causative; inchoative
Object-drop verbs	drop argument
Raising and tough predicates	there raising; it raising; tough/raising 1/2
Animate-selection inventories	animate subj. trans.; animate subj. passive

Table A1: Generated paradigms that use curated lexical or predicate lists.

when their senses are dispersed across incompatible classes; accepted nouns inherit matching BLiMP noun templates.

Adjectives. Adjective candidates are also drawn from Open English WordNet. We remove relational-only and otherwise unstable adjectives, then map the remaining items to compatibility bundles such as generic, animate-compatible, physical, document-compatible, cleanable, and food-related adjectives. Each accepted adjective inherits a template whose selectional restrictions match the assigned bundle. Paradigms requiring specific adjectival behavior, such as control, raising, and tough-adjective constructions, use curated predicate lists.

Verbs. For verbs, we construct a frame-annotated inventory with coarse subcategorization labels such as intransitive, transitive, ditransitive, prepositional, and particle frames. Frame labels are assigned and validated using VerbNet (Kipper Schuler, 2005), WordNet, corpus checks, and manual review. Accepted verb lemmas are inserted into BLiMP-compatible verb templates, with inflected forms generated automatically. Paradigms that depend on more specific verbal behavior, such as control, raising, passivization, causative–inchoative alternations, and object drop, use curated verb lists rather than unconstrained frequency-filtered sampling.

Table A1 lists the generated paradigms that use curated lexical or predicate lists.

B Artifact Use and Licenses

We use existing artifacts for research evaluation and lexical-resource construction. BLiMP and Open English WordNet are distributed under CC BY 4.0; Wikidata data is distributed under CC0; spaCy is distributed under the MIT License; wordfreq code is distributed under the Apache License and includes redistributable frequency data under CC BY-SA 4.0; Mistral-7B is distributed under Apache 2.0; Llama-3.1 is distributed under the Llama 3.1 Community License; Gemma models are distributed under the Gemma Terms of Use; and Qwen2.5 models are used under their corresponding Hugging Face license terms, with the 72B variants covered by the Qwen license rather than Apache 2.0. VerbNet is used under its stated VerbNet license, and COCA is used under the English-Corpora.org access terms. COCA is used only as an external source of aggregate local collocational counts, and we do not redistribute COCA text.

C Dataset Diagnostics and Lexical Diversity

Table C1 gives the realized regime-level lexical statistics used in the diagnostics. Diversity is computed over frequency-controlled content-word occurrences. Table C2 reports the corresponding sentence-length and tokenizer-length diagnostics, since rare lexicalization can increase subword tokenization even when word counts remain similar.

D Scoring and Implementation Details

D.1 Exact Scoring Templates

For Template LP, each sentence is embedded in the following string:

```
The following sentence is
grammatically acceptable.

{sentence}
```

For Yes/No scoring with base models, the completion prompt is:

```
Your task is to evaluate the
quality of given text.
Is the following sentence
grammatically acceptable?

{sentence}
Respond with Yes or No as your
answer. Answer:
```

For chat models, the system message is `Your task is to evaluate the quality of given text.` The user message is:

```
Is the following sentence
grammatically acceptable?
Respond with Yes or No as
your answer.

{sentence}
```

Dataset / regime	Pairs	Mean Zipf	Median	IQR	Lemmas	Surfaces	Eff. lemmas	Top-20
Original BLiMP	67,000	4.26	4.26	0.65	2,607	2,808	441	14.3
Head	67,000	4.10	4.05	0.47	6,304	9,363	1,954	2.9
Tail	67,000	2.89	2.85	0.29	10,583	13,843	2,559	4.3
Extreme tail	67,000	1.93	1.85	0.37	15,111	18,751	2,660	4.0

Table C1: Regime-level dataset statistics. Mean Zipf, median Zipf, and IQR are computed over realized frequency-controlled content-word occurrences; IQR is the interquartile range. Eff. lemmas is the effective number of lemmas ($1/\sum_i p_i^2$), where p_i is each lemma's share of frequency-controlled content-word occurrences. Top-20 lemma share is reported as a percentage of those occurrences.

Dataset / regime	Mean words	Mean tokens	Controlled content words
Original BLiMP	7.72	9.98	–
Head	7.84	9.73	3.43
Tail	7.95	12.79	3.37
Extreme tail	7.86	14.00	3.31

Table C2: Length and tokenization diagnostics for grammatical sentences. Token counts are averaged across the seven base-model tokenizers under LP readout; controlled content words are frequency-controlled lexical slots.

The tokenizer chat template is then applied with generation prompting enabled. Yes/No scores use the next-token probability mass assigned to single-token variants `Yes`, `Yes`, `\nYes` and `No`, `No`, `\nNo`. The score is $p_Y/(p_Y + p_N)$.

D.2 Statistical Aggregation

For each model, paradigm, and bin or regime, we first average item-level scores within the paradigm and then average paradigm means equally. Confidence intervals are normal-approximation intervals over paradigm means:

$$\bar{x} \pm 1.96\frac{s}{\sqrt{n}},$$

where n is the number of represented paradigms. Reported aggregate results are paradigm-balanced rather than item-balanced.

D.3 Compute Details

Table D1 lists the main direct packages used for dataset construction, preprocessing, evaluation, and analysis. The final FREQBLIMP generation pipeline uses `wordfreq` for Zipf frequencies and `lemminflect` for inflection; spaCy is used only in the COCA local-support analysis.

Zipf frequencies are computed with `wordfreq.zipf_frequency(text, "en")`. Inflection uses `lemminflect` with Penn-style tags such as `NN/NNS`, `JJ/JJR/JJS`, and `VB/VBP/VBZ/VBG/VBD/VBN`. The COCA local-support analysis uses spaCy dependency parses and lemmas to extract local content-word dependencies.

Evaluations were run as inference-only jobs on NVIDIA H100 GPUs. Models in the 4B–8B range used one H100, 31B models used two H100s, and 70B–72B models used four H100s. The final selected evaluation suite corresponds to approximately 320 H100 GPU-hours. Auxiliary pilot runs not used in the paper are excluded. Dataset generation used CPU jobs and did not involve model training.

Package	Version
`transformers`	4.57.3
`torch`	2.8.0
`wordfreq`	3.1.1
`lemminflect`	0.2.3
`spaCy`	3.7.4
`en_core_web_sm`	3.7.1
`wn`	0.13.0
`datasets`	4.4.1
`pandas`	2.3.3
`numpy`	2.0.2 / 1.26.4
`scipy`	1.13.1

Table D1: Main direct package versions used in dataset construction, preprocessing, evaluation, and analysis. `numpy`'s two versions indicate separate construction and evaluation environments.

E Additional Aggregate Results and Robustness Checks

E.1 All-model Frequency Trends

The main-text frequency plots show the base-model average and the two models with the smallest and largest head-to-tail changes, in order to make the trends readable with confidence intervals. Figure E1 shows the corresponding all-model version without confidence bands.

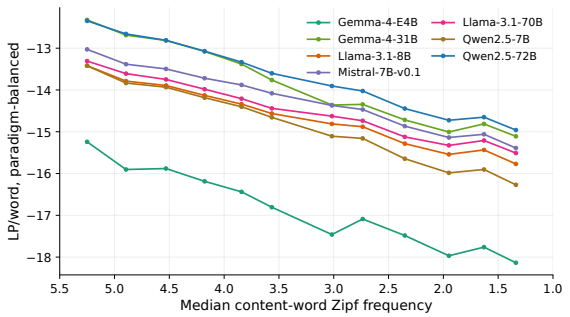

Figure E1: Paradigm-balanced word-normalized log probability of grammatical sentences as a function of realized Zipf frequency for all base models. Confidence bands are omitted for readability.

E.2 Full Performance Results Per Method

Table E1 expands the scoring-interface summary by showing each generated regime separately. Margins are LP differences for LP and Template LP, and $P_{yes}(g) - P_{yes}(b)$ for Yes/No.

E.3 Attractor and Distance Effects

Table E3 reports the agreement-attractor analysis moved from the main text. Positive penalties mean that the attractor or distance condition is harder than its no-attractor or local baseline.

Rare lexicalization does not amplify relative-clause subject–verb interference under LP readout: the median penalty is, in fact, highest in Head and lower in Tail and Extreme-tail. The same direction appears under Template LP, with smaller absolute penalties. Relative-clause subject–verb attractors remain difficult in absolute terms, but the median attractor penalty decreases from original BLiMP and the generated Head regime to the lower-frequency regimes (22.3 points in original BLiMP, 24.3 in Head, 18.7 in Tail, and 17.1 in Extreme-tail). Relative-clause attractor accuracy also improves from Head to Extreme-tail for all seven base models, with the same direction under Template LP, though with smaller effects.

The pattern is consistent with rare lexicalization removing some high-frequency lexical traps rather than making structural agreement easier. For example, a Head item such as *This lieutenant who had shoved most sergeants* <u>*has*</u>/*<u>*have*</u> *incorporated some takeovers* places the incorrect auxiliary after a plausible plural local continuation (*sergeants have*). The corresponding Extreme-tail lexicalization, *This gamine who had homogenized most zombis* <u>*has*</u>/*<u>*have*</u> *whopped some galangals*, preserves the same singular head noun, plural attractor, and auxiliary contrast, but makes the local lexical continuation much less familiar. Thus, rare lexicalization does not make models more vulnerable to structural agreement interference. Instead, it appears to reduce accidental lexical or collocational traps present in high-frequency lexicalizations, while leaving overt agreement cues available.

F Full Phenomenon and Paradigm-Level Results

Table F1 reports the broad BLiMP phenomenon results, averaging LP-readout scores across the seven base models.

Tables F2–F5 reports the full 67-paradigm breakdown, averaged across the seven base models under LP readout. We sort paradigms by XTail–Head accuracy change, from the largest negative changes to the largest positive changes. The released analysis files include the corresponding full identifiers and model-specific results.

G Original–Head diagnostics

This appendix provides additional diagnostics for the comparison between original BLiMP and Head FreqBLiMP. This comparison is not the primary frequency manipulation, since original BLiMP was not generated by the same lexicalization pipeline as FreqBLiMP. Instead, it indexes broader effects of regeneration and lexicalization beyond unigram frequency.

Figure G1 first shows that original BLiMP itself exhibits the same qualitative likelihood–accuracy dissociation observed in FreqBLiMP: lower-frequency items receive lower likelihood, while contrastive accuracy declines only mildly.

Table E2 gives the direct model-level comparison, including both accuracy and margin. Here we break the same contrast down by paradigm.

Table G1 shows that the largest original–Head drops cluster in paradigms where vocabulary-expanded regeneration can alter the local constructional frame.

For example, in TOUGH_VS_RAISING_1, a generated grammatical item such as *All girls are tedious to treasure* competes with an ungrammatical alternative containing the highly familiar *supposed to* frame. LP readout can therefore favor local familiarity even when it conflicts with the intended grammatical contrast.

Model	H acc.	T acc.	XT acc.	H mrg.	T mrg.	XT mrg.
LP readout						
G4-E4B	75.6	72.3	71.4	2.72	2.55	2.32
G4-31B	75.9	70.3	68.6	2.69	2.08	1.41
L3.1-8B	73.3	71.1	70.2	2.30	2.12	2.03
L3.1-70B	74.3	73.3	72.3	2.37	2.33	2.28
M-7B	75.8	74.2	74.2	2.69	2.59	2.69
Q2.5-7B	75.8	75.2	74.2	2.90	2.95	2.93
Q2.5-72B	74.7	72.6	70.4	2.75	2.68	2.44
G4-31B-I	57.0	56.2	54.9	1.82	1.91	1.79
L3.1-70B-I	72.2	71.4	70.8	2.10	2.19	2.11
Q2.5-72B-I	73.4	71.5	69.6	2.91	2.95	2.72
Template LP						
G4-E4B	76.4	74.2	73.1	2.74	2.52	2.36
G4-31B	77.2	72.5	71.2	2.80	2.27	1.71
L3.1-8B	77.4	76.3	75.1	2.84	2.69	2.59
L3.1-70B	77.9	77.8	77.1	2.83	2.86	2.80
M-7B	78.1	78.1	77.8	2.92	3.00	3.01
Q2.5-7B	77.6	76.3	75.4	2.97	2.94	2.91
Q2.5-72B	76.8	75.7	74.4	2.88	2.89	2.75
G4-31B-I	63.0	62.2	61.4	3.48	3.92	3.93
L3.1-70B-I	77.9	75.8	74.5	2.97	3.00	2.85
Q2.5-72B-I	75.4	74.9	73.6	3.29	3.45	3.33
Yes/No						
G4-E4B	56.0	52.5	49.4	0.01	0.00	-0.00
G4-31B	81.6	80.4	79.6	0.08	0.06	0.06
L3.1-8B	79.8	78.2	76.2	0.05	0.04	0.03
L3.1-70B	84.7	82.1	79.7	0.13	0.10	0.08
M-7B	74.3	72.9	70.3	0.03	0.02	0.02
Q2.5-7B	82.2	81.0	79.6	0.16	0.13	0.11
Q2.5-72B	86.0	85.4	84.1	0.12	0.11	0.09
G4-31B-I	88.9	88.2	87.4	0.41	0.45	0.44
L3.1-70B-I	86.8	84.9	82.1	0.33	0.35	0.31
Q2.5-72B-I	87.6	87.4	86.2	0.19	0.21	0.22

Table E1: Generated-regime performance grouped by scoring method. Accuracy columns are percentages.

H Qualitative Examples

Table H1 gives representative examples from the qualitative audit. These examples illustrate why frequency effects should be interpreted as interactions among rarity, lexical licensing, and local constructional familiarity, rather than as a pure loss of abstract grammatical knowledge.

Model	Orig. acc.	Head acc.	Δ acc.	Orig. mrg.	Head mrg.	Δ mrg.
G4-E4B	79.9	75.6	-4.3	3.51	2.72	-0.79
G4-31B	78.4	75.9	-2.5	3.24	2.69	-0.55
L3.1-8B	79.6	73.3	-6.3	3.34	2.30	-1.04
L3.1-70B	80.1	74.3	-5.8	3.40	2.37	-1.03
M-7B	81.9	75.8	-6.0	3.79	2.69	-1.10
Q2.5-7B	79.3	75.8	-3.4	3.76	2.90	-0.86
Q2.5-72B	78.5	74.7	-3.8	3.63	2.75	-0.88

Table E2: Original BLiMP vs. Head FREQBLIMP under LP readout for base models.

Domain	Penalty	Orig.	Head	Tail	XTail
Subject–verb	relative-clause attractor, LP	22.3	24.3	18.7	17.1
Subject–verb	relative-clause attractor, Temp. LP	4.0	10.3	5.9	4.0
Subject–verb	relational-noun attractor, LP	7.9	5.3	0.1	4.9
Determiner–noun	adjective distance, LP	3.9	5.8	3.1	4.3

Table E3: Median attractor or distance penalties in percentage points across seven base models. Positive values indicate that the target condition is harder than the local baseline.

Phenomenon	H acc.	T acc.	XT acc.	Δ	H mrg.	XT mrg.	Δ mrg.
anaphor agreement	97.3	92.9	92.0	-5.3	4.45	3.51	-0.94
argument structure	73.1	63.5	61.5	-11.6	2.46	1.43	-1.03
binding	76.1	74.9	75.6	-0.6	2.81	2.56	-0.25
control raising	72.0	60.6	55.8	-16.3	2.83	-0.81	-3.64
determiner noun agreement	87.9	89.6	89.4	+1.4	2.89	3.26	+0.37
ellipsis	78.4	73.6	71.5	-6.9	5.71	6.06	+0.34
filler gap dependency	71.0	71.0	70.3	-0.7	2.57	3.05	+0.49
irregular forms	87.5	85.7	86.8	-0.6	3.55	3.57	+0.02
island effects	59.3	58.2	56.0	-3.4	1.07	0.91	-0.16
npi licensing	73.0	71.5	71.5	-1.5	2.49	2.49	-0.00
quantifiers	72.9	72.1	70.9	-2.0	3.03	2.73	-0.30
subject verb agreement	79.2	83.7	82.9	+3.7	2.29	2.98	+0.69

Table F1: Broad phenomenon-level results averaged across base models.

Paradigm	Phen.	H	T	XT	Δ	H mrg.	XT mrg.	Δ mrg.
expletive it obj. raising	control/raising	81.7	53.9	52.9	-28.8	3.76	-3.39	-7.15
expletive there obj. raising	control/raising	74.2	51.6	50.3	-23.9	2.83	-3.86	-6.69
inchoative	argument struct.	84.2	60.1	60.7	-23.4	4.35	0.92	-3.44
passive 1	argument struct.	79.1	71.2	62.0	-17.1	2.60	2.12	-0.48
expletive there subj. raising	control/raising	85.0	83.2	68.0	-17.0	5.13	2.01	-3.12
drop argument	argument struct.	77.7	55.5	62.0	-15.7	2.37	1.18	-1.19
causative	argument struct.	74.3	60.1	60.9	-13.4	3.45	1.13	-2.32
tough vs. raising 2	control/raising	74.9	70.2	61.8	-13.1	2.74	-1.39	-4.13
passive 2	argument struct.	80.5	74.4	67.4	-13.1	2.68	2.45	-0.23
Principle A domain 3	binding	63.9	52.4	53.3	-10.7	1.59	0.62	-0.97
left-branch simple	island effects	75.2	73.8	66.9	-8.4	2.99	2.38	-0.61
ellipsis N-bar 1	ellipsis	57.8	53.2	49.8	-8.0	0.91	0.27	-0.64
animate subj. trans.	argument struct.	64.0	59.1	56.5	-7.6	1.99	1.30	-0.70
anaphor gender agr.	anaphor agr.	97.1	90.2	89.6	-7.5	4.39	3.14	-1.24
wh-island	island effects	60.3	58.2	53.1	-7.2	1.07	0.56	-0.51
sentential subj. island	island effects	37.3	33.2	31.0	-6.4	-0.53	-1.01	-0.48
intransitive	argument struct.	67.3	65.4	61.1	-6.2	1.56	0.69	-0.87
superlative quant. 1	quantifiers	86.9	82.8	80.9	-6.1	2.75	2.34	-0.41
ellipsis N-bar 2	ellipsis	98.9	94.1	93.1	-5.8	10.51	11.84	+1.33

Table F2: Full paradigm-level LP-readout results averaged across the seven base models, sorted by head–extreme-tail change. Accuracy columns are percentages; Δ is XTail minus Head. Paradigm names are abbreviated for space.

Paradigm	Phen.	H	T	XT	Δ	H mrg.	XT mrg.	Δ mrg.
only NPI licensor	NPI licensing	86.6	82.8	81.2	-5.5	2.90	2.64	-0.26
det.-noun adj. 2	det.-noun agr.	89.8	89.7	84.9	-4.9	3.12	2.90	-0.22
animate subj. passive	argument struct.	63.3	60.3	58.8	-4.5	1.52	1.26	-0.26
NPI present 2	NPI licensing	70.6	67.1	66.7	-4.0	1.66	1.50	-0.16
det.-noun agr. 2	det.-noun agr.	95.7	94.0	91.8	-3.9	3.90	3.36	-0.54
complex NP island	island effects	49.4	47.2	46.0	-3.4	0.23	-0.09	-0.32
transitive	argument struct.	67.2	65.5	64.0	-3.3	1.64	1.86	+0.23
anaphor number agr.	anaphor agr.	97.5	95.5	94.5	-3.0	4.50	3.88	-0.63
NPI present 1	NPI licensing	61.3	59.3	58.7	-2.5	0.93	0.79	-0.15
regular plural SVA 1	subj.-verb agr.	89.9	87.6	87.6	-2.3	3.43	3.54	+0.11
there quantifiers 1	quantifiers	98.4	97.8	96.2	-2.3	8.02	7.01	-1.01
wh subj. gap, long	filler-gap	94.3	92.9	92.1	-2.2	5.22	6.07	+0.85
irreg. participle adj.	irregular forms	91.1	86.8	89.3	-1.8	3.92	3.98	+0.06
coord. struct., complex LB	island effects	68.5	68.0	66.8	-1.7	2.53	2.54	+0.02
Principle A domain 2	binding	74.6	70.3	73.3	-1.3	1.41	1.25	-0.15
Principle A case 2	binding	85.8	84.4	84.5	-1.3	2.98	3.78	+0.80
wh/that no gap	filler-gap	95.2	94.7	93.9	-1.3	4.97	5.76	+0.79
negation NPI licensor	NPI licensing	98.4	97.2	97.4	-1.0	7.25	6.71	-0.55
regular plural SVA 2	subj.-verb agr.	85.4	85.5	84.4	-1.0	3.36	3.69	+0.33
wh/that with gap, long	filler-gap	11.0	11.0	10.1	-1.0	-3.22	-4.01	-0.79

Table F3: Full paradigm-level LP-readout results, continued.

Paradigm	Phen.	H	T	XT	Δ	H mrg.	XT mrg.	Δ mrg.
left-branch echo	island effects	40.4	42.0	39.5	-0.8	-1.69	-1.63	+0.07
there quantifiers 2	quantifiers	32.1	32.1	31.4	-0.8	-1.06	-1.27	-0.21
wh/that with gap	filler-gap	20.5	21.3	19.8	-0.7	-2.40	-2.78	-0.38
coord. struct., obj. extr.	island effects	64.7	64.0	64.0	-0.6	1.60	1.71	+0.11
matrix question NPI	NPI licensing	62.9	62.5	62.5	-0.5	1.48	1.85	+0.37
wh obj. gap	filler-gap	89.1	88.6	88.7	-0.4	3.93	4.77	+0.84
Principle A domain 1	binding	96.8	97.1	96.5	-0.3	3.65	3.23	-0.42
Principle A case 1	binding	100.0	99.9	99.7	-0.3	9.09	7.97	-1.12
wh subj. gap	filler-gap	93.0	94.1	93.2	+0.2	4.74	5.66	+0.91
negation NPI scope	NPI licensing	60.0	60.0	60.3	+0.3	1.04	1.03	-0.01
irreg. participle verbs	irregular forms	83.8	84.5	84.4	+0.6	3.17	3.16	-0.01
wh/that no gap, long	filler-gap	93.9	94.2	94.5	+0.6	4.73	5.92	+1.19
Principle A reconstruction	binding	33.7	33.9	34.5	+0.8	-0.92	-1.01	-0.09
det.-noun irregular 2	det.-noun agr.	91.5	91.7	92.5	+1.0	3.47	3.60	+0.13
superlative quant. 2	quantifiers	74.0	75.6	75.2	+1.2	2.39	2.83	+0.45
det.-noun irregular 1	det.-noun agr.	89.3	90.4	90.6	+1.4	3.28	3.39	+0.11
adjunct island	island effects	78.8	79.5	80.3	+1.5	2.35	2.78	+0.42
tough vs. raising 1	control/raising	44.4	43.9	45.9	+1.6	-0.31	2.57	+2.88

Table F4: Full paradigm-level LP-readout results, continued.

Paradigm	Phen.	H	T	XT	Δ	H mrg.	XT mrg.	Δ mrg.
irreg. plural SVA 1	subj.-verb agr.	84.3	89.0	87.0	+2.7	2.53	3.25	+0.73
det.-noun agr. 1	det.-noun agr.	87.9	89.6	90.6	+2.8	2.51	3.63	+1.13
only NPI scope	NPI licensing	70.8	71.7	73.7	+2.9	2.19	2.93	+0.74
distractor agr., rel. noun	subj.-verb agr.	78.0	85.1	82.0	+4.0	1.54	2.18	+0.63
det.-noun adj. 1	det.-noun agr.	84.2	87.9	88.7	+4.6	1.64	3.15	+1.51
det.-noun adj. irr. 2	det.-noun agr.	82.4	85.7	87.3	+4.8	2.60	3.12	+0.51
det.-noun adj. irr. 1	det.-noun agr.	82.9	87.6	88.4	+5.5	2.60	2.94	+0.34
irreg. plural SVA 2	subj.-verb agr.	78.3	86.0	85.7	+7.5	2.52	3.31	+0.79
Principle A c-command	binding	78.3	86.0	87.3	+9.0	1.84	2.05	+0.21
distractor agr., RC	subj.-verb agr.	59.3	69.1	70.6	+11.3	0.37	1.92	+1.55

Table F5: Full paradigm-level LP-readout results, continued.

Paradigm	Orig.	Head	Gap
tough/raising 1	71.0	44.4	-26.7
wh-vs-that, long	33.9	11.0	-22.8
coord. struct. obj. extr.	86.5	64.7	-21.8
left-branch echo	59.5	40.4	-19.2
wh-vs-that	39.5	20.5	-19.0
wh-island	77.9	60.3	-17.5

Table G1: Paradigms with the largest original–Head LP-readout accuracy drops, averaged across base models. Values are percentages.

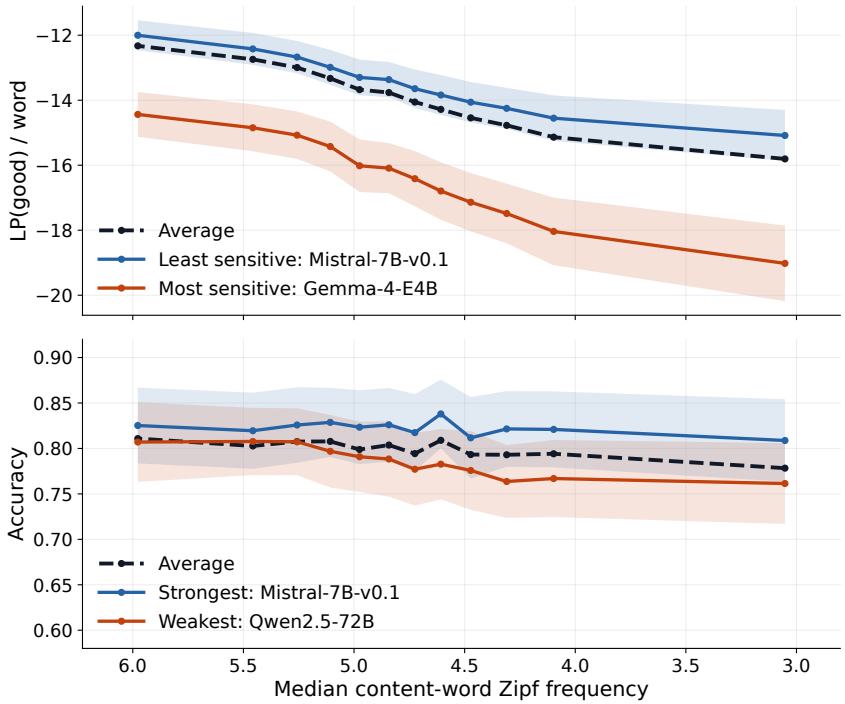

Figure G1: Frequency trends within original BLiMP. Top: paradigm-balanced word-normalized log probability by realized median Zipf frequency. Bottom: paradigm-balanced accuracy by realized median Zipf frequency.

Case / regime	Paradigm	Acceptable	Unacceptable	Pattern
Degradation, XTail	inchoative	*A souchong scorches.*	**A souchong deplores.*	Rare verb-frame licensing.
Degradation, XTail	tough/raising 1	*All interlopers weren't laborious to discomfit.*	**All interlopers weren't fated to discomfit.*	Local raising frame wins.
Degradation, Tail	passive 1	*Stalwarts weren't refunded by many laureates.*	**Stalwarts weren't persevered by many laureates.*	Passive formation is predicate-specific.
Improvement, XTail	distractor agr. RC	*A musicologist who isn't repossessing a lot of sippers quenches these pancreases.*	**A musicologist who isn't repossessing a lot of sippers quench these pancreases.*	Overt agreement survives rare attractors.
Improvement, XTail	Principle A c-command	*The bambinos who are slating the nightie weren't denuding themselves.*	**The bambinos who are slating the nightie weren't denuding itself.*	Reflexive number cue remains recoverable.
Selectional tilt, XTail	animate subj. passive	*Most noncombatants weren't obviated by the shopgirl.*	**Most noncombatants weren't obviated by the intermezzo.*	Agent animacy contrast can be ambiguous.
Orig.–Head gap, Head	tough/raising 1	*All girls are tedious to treasure.*	**All girls are supposed to treasure.*	Familiar local frame favors bad item.
Original success	tough/raising 1	*Many movie theaters aren't easy to bike to.*	**Many movie theaters aren't bound to bike to.*	Conventional tough-adjective frame.
Noisy item, Head	ellipsis N-bar 1	*The republicans imply one fourteen room and all photographers imply a few.*	**The republicans imply one room and all photographers imply a few fourteen.*	Awkward modifier–noun combination obscures contrast.

Table H1: Representative qualitative examples. Underlining marks the local contrast driving the minimal pair.