Large Language Models Are Robust to Low-Frequency Words in Grammatical Evaluation

Tyrone White[1] Yuki Arase[1]
[1]Institute of Science Tokyo
white.q.a4e7@m.isct.ac.jp arase@c.titech.ac.jp

Abstract

Grammatical acceptability benchmarks such as the Benchmark of Linguistic Minimal Pairs (BLiMP) evaluate minimal pairs by comparing sentence likelihoods, but they largely use frequent content words, which can mask failures of grammatical generalisation on long-tail inputs. We introduce FreqBLiMP, a frequency-controlled lexicalisation of BLiMP that swaps selected content words with lemmas sampled under user-specified Zipf constraints, enabling evaluation on rare and diverse inputs while preserving each item's minimal contrast. Across four open-weight language models (Llama-3.1-8B, Llama-3.2-1B/3B, Mistral-7B-v0.3), decreasing lexical frequency sharply lowers overall probability, yet grammatical preference margins remain comparatively stable, indicating robust grammatical discrimination.

1 Introduction

Targeted syntactic evaluation probes whether language models (LMs) encode structure-sensitive generalisations beyond surface heuristics. Existing methods are based on controlled, hand-crafted/template minimal pairs targeting specific syntactic dependencies and score LMs by whether they assign higher probability to the grammatical variant [1, 2]. BLiMP [3] systematises this approach across dozens of English phenomena, but its items overwhelmingly use common lexical material. This matters because lexical frequency is not just "style": it affects tokenisation length, plausibility priors, and the amount of memorised lexical knowledge a model can rely on. As a result, conclusions drawn from head-heavy suites may not transfer to the long tail that real-world inputs repeatedly live in [4, 5].

We present FreqBLiMP, a frequency-controlled lexicalisation of BLiMP that systematically shifts lexical frequency while keeping each item's grammatical contrast fixed, which enables the analysis of long tail frequency effects (see Figure 1).

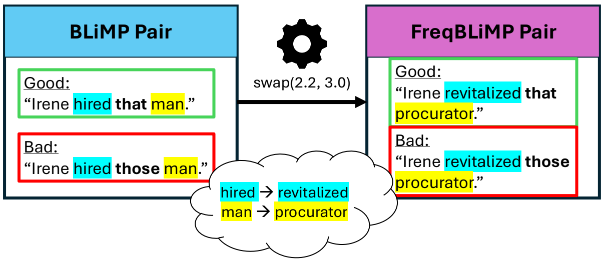

Figure 1: FreqBLiMP lexicalisation example: the agreement contrast (bold) is preserved while eligible content words (highlighted) are swapped under a Zipf window (2.2–3.0) and inflected.

We evaluate multiple open-weight LMs by next-token scoring and analyse both (i) *overall sentence likelihood* as frequency decreases and (ii) *grammatical discrimination* within each minimal pair. Our results reveal that LMs' grammatical discrimination is relatively robust to frequency-driven lexical perturbations, even as sentence probability degrades.

1.1 Relation to prior work.

Minimal-pair suites such as SyntaxGym [6] standardize targeted, psycholinguistically-inspired syntactic evaluations to make results more reproducible across models. Complementarily, CoLA [7] provides sentence-level acceptability labels drawn from linguistics literature, offering broad coverage but less experimental control than template-generated minimal-pair suites. "Colorless green ideas" evaluations [8] perform lexical randomization of content words much like our approach, producing sentences that reduce lexical/semantic cues for syntactic prediction; but their focus is on a single phenomenon: long-distance number agreement. Our approach instead inherits BLiMP's broad phenomenon coverage, and while our lexical swaps can similarly reduce semantic plausibility cues, we focus on *controlling lexical frequency* as it is more directly measurable and tunable.

Frequency regularities are heavy-tailed [9, 10, 11], and long-tail degradation is documented in both knowledge and evaluation settings [4, 5]. Closest to our setting, LT-Swap [12] benchmarks models on long-tail words using minimal-pair swaps tied to a specific pretraining corpus. However, their primary goal is rare-word acquisition and correct usage, not isolating syntactic sensitivity under controlled frequency shifts across various grammatical phenomena. Finally, training-time shifts such as "model collapse" from recursive training on model-generated text [13] may further truncate the long tail, reinforcing the need for explicit frequency-aware evaluation.

2 FreqBLiMP construction

BLiMP consists of minimal pairs: two sentences that differ minimally, with one grammatical and one ungrammatical, designed to isolate a specific phenomenon. Given a BLiMP minimal pair (g_o, b_o) (grammatical g; ungrammatical b), we generate swapped counterparts (g_s, b_s) by replacing selected content words with lemmas drawn from user-specified Zipf windows or thresholds, constrained by verb subcategorisation frames, and inflected to preserve agreement (Figure 1).

2.1 Construction Method

Frequency control. We operationalise lexical frequency with Zipf-scale estimates from `wordfreq` [14]. Users specify Zipf windows (lower and upper Zipf-frequency bounds) or one-sided thresholds for target lemmas, optionally separately for nouns, adjectives, and verbs. Candidates can be sampled either uniformly or frequency-weighted; in our experiments we instantiate fixed Zipf windows and sample uniformly within each window. For analysis we record a *realised* frequency for each swapped pair: the mean of the median Zipf values of the substituted lemmas in the grammatical and ungrammatical variants.

Replacement constraints. Noun/adjective candidates are drawn from Open English WordNet [15] and filtered by coarse morphosyntactic constraints (number and inflectional compatibility) and lexical-class constraints (e.g., countability via BECL [16], plus person/gender filters), in addition to the requested Zipf window. Verb candidates come from a precomputed inventory with frame information derived from VerbNet [17] and validated using COCA frequency counts [18]; we filter to match the target frame and Zipf window. We don't swap irregular verbs (as rare replacements can't be found), adverbs, and avoid swapping proper names and entity-like tokens to reduce cultural/knowledge confounds.

Swapping and inflection. We select all eligible content words and apply swaps in a fixed order (verbs → nouns → adjectives), attempting to swap the two variants of a minimal pair in parallel. Replacements are inflected with `lemminflect` [19] to match the original token's POS tag, found using spaCy [20] tags features plus light heuristics, with minimal surface cleanup (e.g., `a`/`an`). If constraints cannot be satisfied, the item is dropped for that regime.

2.2 Dataset Analysis

To validate the lexicalisation procedure and rule out pipeline artefacts, we perform dataset-level analyses covering realised-Zipf regime separation, lexical diversity, swap coverage, and grammatical–ungrammatical symmetry.

We confirm that realised Zipf distributions are well separated under our three experimental windows (head: 3.6–5.0, tail: 2.2–3.0, extreme-tail: 1.2–2.0), with only minor overlap from inflected forms whose Zipf differs from the base lemma (Figure 2, top). As expected, the original BLiMP items mostly fall in our head regime, with a broader and less controlled Zipf distribution. As an additional diversity diagnostic, the top-20 lemma share (Figure 2, bottom) indicates high lemma diversity across regimes, with only a modest rise in concentration at the extreme tail (to around 10%), helping ensure that trends are not driven by repeated sampling of a small lemma set.

As a robustness verification on the lexicalisation pipeline, swap coverage is high across regimes (success rate ≈ 0.974–0.975); the few failures are mainly due to candidate pool exhaustion under morphosyntactic/frame constraints or cases where no eligible target token remains after filtering. Finally, we confirm that lexicalisation does not introduce systematic asymmetries between the grammatical and ungrammatical variants: the token-length and realised median Zipf difference between g and b remains small on average and does not vary meaningfully by regime.

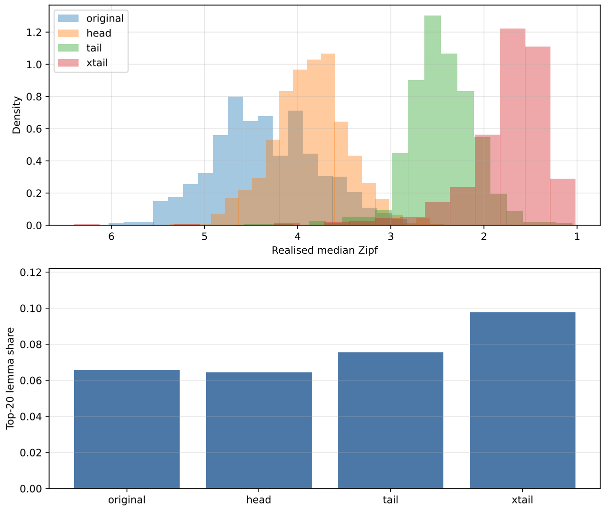

Figure 2: Diagnostics for realised lexical rarity and diversity. **Top:** distribution of realised median Zipf (median over swapped lemmas) for the original BLiMP items and our Zipf regimes. **Bottom:** top-20 lemma share (fraction of substitutions accounted for by the 20 most frequent substituted lemmas).

3 Evaluation Settings

Conditions and metrics. For each BLiMP item we score the original minimal pair (g_o, b_o) and the swapped pair(s) (g_s, b_s) for three Zipf regimes (head: 3.6-5.0, tail: 2.2-3.0, extreme tail: 1.2-2.0). Let $x = (x_1, \dots, x_T)$ be the tokenised sentence. We define sentence negative log-likelihood as

$$\mathrm{NLL}(x) = -\sum_{t=1}^{T} \log p(x_t \mid x_{<t}). \qquad (1)$$

We define accuracy as the proportion of pairs for which the model assigns lower NLL (higher probability) to the grammatical sentence, i.e., $\mathrm{NLL}(g) < \mathrm{NLL}(b)$ and the preference margin as

$$\Delta\mathrm{NLL} = \mathrm{NLL}(b) - \mathrm{NLL}(g), \qquad (2)$$

where larger values indicate a stronger preference for the grammatical variant. To quantify overall sentence likelihood independently of the contrast, we also compute

$$\mathrm{NLL}_{\mathrm{pair}} = \tfrac{1}{2}\big(\mathrm{NLL}(g) + \mathrm{NLL}(b)\big). \qquad (3)$$

Models. We evaluate three base Llama models and one non-Llama baseline: Llama-3.1-8B, Llama-3.2-1B, Llama-3.2-3B, and Mistral-7B-v0.3. All models are evaluated using next-token scoring only (no generation).

Research questions. We study three research questions (RQs). RQ1. *How does decreasing lexical frequency affect model behaviour*: does controlled descent in Zipf frequency correlate with lower sentence likelihood, and to what extent does it degrade grammatical discrimination? RQ2. *How do these effects vary across model scale and architecture*, comparing Llama models of different sizes (1B, 3B, 8B) and a different family (Mistral-7B)? RQ3. *Are frequency effects uniform across grammatical phenomena*, or do some constructions exhibit greater sensitivity or robustness under frequency-controlled lexicalisation?

4 Results

We report three complementary views: realised-frequency trends, discrete regime summaries, and phenomenon-level sensitivity.

4.1 Realised frequency trends

Figure 3 plots ΔNLL and $\mathrm{NLL}_{\mathrm{pair}}$ against realised median Zipf (descending; rarer to the right) over swapped items. Across all models, $\mathrm{NLL}_{\mathrm{pair}}$ increases substantially as realised Zipf decreases, confirming that decreasing lexical frequency induces a strong, continuous drop in overall sentence likelihood (RQ1). In contrast, ΔNLL remains comparatively stable across the same range: correlations between realised Zipf and ΔNLL are near zero for all models ($|r_s| < 0.015$), indicating that lexical frequency has limited effect on contrastive grammatical discrimination (RQ1). As lexical rarity increases, subword fragmentation inflates token length from head to extreme-tail. We therefore examine character-normalised trends separately (Appendix A), where the consistent trends were confirmed.

Notably, Llama-3.2-1B exhibits slightly larger preference margins than the larger Llama variants throughout the Zipf descent, while Mistral-7B shows the strongest margins overall. This indicates that robustness of contrastive grammatical preference does not scale monotonically with model size or architecture (RQ2).

4.2 Regime-level summary

Table 1 reports minimal-pair accuracy for the original BLiMP condition and the three swapped Zipf windows. Accuracy is similar on the original dataset across models (around 0.80) and lower for all swapped regimes, consistent with lexical substitution introducing distribution shift

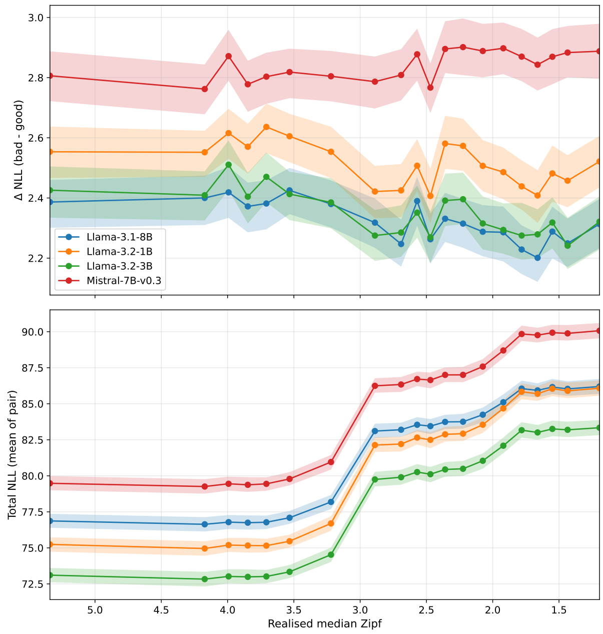

Figure 3: Realised frequency trends on swapped items. **Top:** mean ΔNLL vs realised median Zipf. **Bottom:** mean NLL_{pair} vs realised median Zipf. Overall sentence likelihood decreases with rarity (increasing NLL_{pair}), while preference margins remain broadly stable.

Table 1: Minimal-pair accuracy by dataset condition and Zipf-frequency window.

Dataset / regime	L3.1-8B	L3.2-1B	L3.2-3B	M-7B
Original	0.797	0.804	0.796	0.818
Head (3.6–5.0)	0.725	0.738	0.727	0.753
Tail (2.2–3.0)	0.717	0.728	0.715	0.756
Xtail (1.2–2.0)	0.708	0.719	0.705	0.749

beyond frequency (e.g., plausibility/register). Small differences still show that Llama-3.2-1B and Mistral-7B slightly outperform the rest. Within the swapped variants, the additional decline from head → tail → extreme-tail is modest, implying that progressively lower frequency adds little additional degradation beyond the lexicalisation step itself (RQ1).

Model-level differences are small but systematic: Llama-3.2-1B and Mistral-7B slightly outperform the other models across swapped regimes, again indicating that robustness to frequency-controlled lexicalisation does not increase monotonically with model size (RQ2).

4.3 Phenomenon-level sensitivity

Per-phenomenon results for all models are reported in Appendix B. Addressing RQ3, most phenomena exhibit only gradual degradation (typically a few percentage points). The largest mean declines are observed for determiner–noun agreement, argument structure, island effects, and binding, indicating that frequency sensitivity is not entirely uniform across constructions.

A notable exception is CONTROL_RAISING, which improves as lexical rarity increases for all models. This contrasts control-like and raising-like constructions with similar surface form but different syntactic dependencies (e.g., *g*: *He is tough to work with.* vs. *b*: *He is likely to work with.*). This suggests a non-monotonic interaction between lexicalisation and construction-specific cues, potentially reflecting a shift toward greater reliance on syntactic scaffolding when semantic cues become less informative.

5 Discussion

The substantial performance gap between original and swapped conditions should not be attributed to frequency alone. Although swaps are constrained by POS and verb subcategorisation frames and inflected to preserve agreement, lexical substitution inevitably alters semantic naturalness, including plausibility, selectional preferences, and register. Frequency estimates further depend on the corpora behind `wordfreq` [14] and may not reflect a model's effective training distribution; similarly, COCA-based filtering only approximates usage frequency. Finally, our evaluation is English-only and probability-based; future work includes extending to other languages and to generation-based grammaticality judgments, and analysing frequency effects under open-data training setups where corpus-internal counts can be computed directly for a given model.

6 Conclusion

FREQBLIMP provides a frequency-controlled lexicalisation of BLiMP that enables grammatical acceptability evaluation across explicit long-tail regimes. Across multiple open-weight LMs, decreasing lexical frequency strongly reduces overall sentence likelihood but leaves grammatical preference margins comparatively stable, suggesting that contrastive grammatical discrimination is relatively resilient under lexically driven frequency shifts. We release the dataset variants and scoring tools to support frequency-aware evaluation and analysis.[1)]

[1)] https://github.com/TimeTravelerTy/freq-blimp

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP25K03233.

References

[1] Rebecca Marvin and Tal Linzen. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018.

[2] Benjamin Newman, Kai-Siang Ang, Julia Gong, and John Hewitt. 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*, Online, 2021. Association for Computational Linguistics.

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

[4] Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

[5] Byung-Doh Oh, Shisen Yue, and William Schuler. 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)*, pp. 2644–2663, St. Julian's, Malta, 2024. Association for Computational Linguistics.

[6] Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 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*, pp. 70–76, Online, 2020. Association for Computational Linguistics.

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

[8] Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 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 (Long Papers)*, New Orleans, Louisiana, 2018. Association for Computational Linguistics.

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

[10] Steven T. Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 2014.

[11] Fengxiang Fan. An asymptotic model for the english hapax/vocabulary ratio. *Computational Linguistics*, Vol. 36, No. 4, pp. 631–637, 2010.

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

[13] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, Vol. 631, pp. 755–759, 2024.

[14] Robyn Speer. wordfreq: Zipf frequency estimates for words. Python package.

[15] John Philip McCrae, Alexandre Rademaker, Ewa Rudnicka, and Francis Bond. English wordnet 2020: Improving and extending a wordnet for english using an open-source methodology. In *Proceedings of the LREC 2020 Workshop on Multimodal Wordnets (MMW2020)*, Marseille, France, 2020. European Language Resources Association (ELRA).

[16] Tibor Kiss, Francis Pelletier, Halima Husić, Johanna Poppek, and R Simunic. A sense-based lexicon of count and mass expressions: The bochum english countability lexicon. 05 2016.

[17] Karin Kipper-Schuler. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. PhD thesis, University of Pennsylvania, 2005.

[18] Mark Davies. The corpus of contemporary american english (coca), 2008-. Available online.

[19] Brad Jascob. Lemminflect: A python module for english lemmatization and inflection. https://github.com/bjascob/LemmInflect, 2019.

[20] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spacy: Industrial-strength natural language processing in python, 2020.

A Additional normalisation analyses

Character-normalised trends. Because lexical rarity increases subword fragmentation, sentence-level NLL can be partially confounded by token count. We therefore repeat the realised-frequency trend analysis using NLL per character (Figure 4). The overall pattern remains: $\mathrm{NLL}_{\mathrm{pair}}$/char increases as realised median Zipf decreases, while ΔNLL/char is comparatively stable, with only mild drift for some models.

Tokenisation length trends. As a diagnostic for segmentation effects, we report mean tokenised length as a function of realised Zipf (Figure 5), showing a clear increase toward the extreme tail.

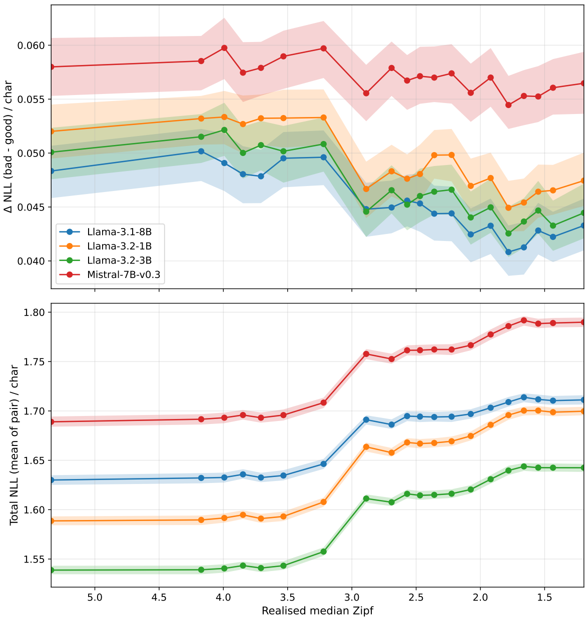

Figure 4: Character-normalised $\mathrm{NLL}_{\mathrm{pair}}$/char vs. realised median Zipf.

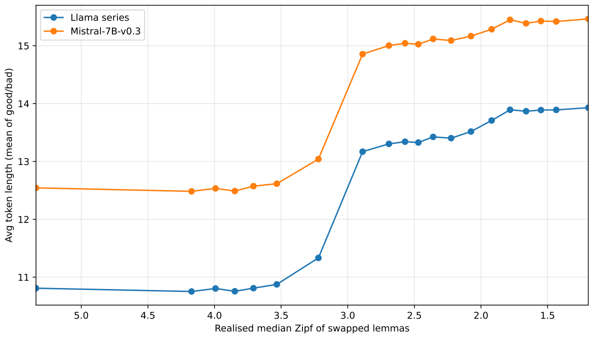

Figure 5: Mean tokenised length vs. realised median Zipf (segmentation-induced length inflation toward the tail).

B Per-phenomenon results

Table 2 reports mean accuracy changes by phenomenon and model for the head and extreme-tail regimes used in our experiments.

Table 2: Phenomenon-level accuracy within swapped variants (head → extreme-tail). Head/Xtail accuracies are averaged across models; Δpp is the mean change in percentage points; Range shows min and max model-specific Δpp.

Phenomenon	Head acc	Xtail acc	Δpp (xtail–head)	Range (min..max)
determiner_noun_agreement	0.891	0.855	−3.6	[−5.2, −0.5]
argument_structure	0.638	0.603	−3.5	[−3.8, −2.9]
island_effects	0.630	0.597	−3.3	[−3.9, −2.5]
binding	0.739	0.709	−3.1	[−4.7, −0.3]
ellipsis	0.742	0.722	−2.0	[−5.3, 1.6]
s-selection	0.738	0.722	−1.5	[−2.4, −0.6]
npi_licensing	0.671	0.659	−1.2	[−2.9, 1.4]
filler_gap_dependency	0.693	0.687	−0.6	[−1.9, 0.5]
quantifiers	0.712	0.707	−0.5	[−1.7, 1.2]
irregular_forms	0.912	0.909	−0.3	[−0.6, 0.4]
anaphor_agreement	0.958	0.958	−0.1	[−0.4, 0.3]
subject_verb_agreement	0.777	0.779	+0.3	[−1.8, 2.6]
control_raising	0.741	0.769	+2.8	[1.9, 3.6]