Grammar Aware Language Models

Karl V. Muller
Cornell Tech
km2262@cornell.edu

Abstract

Large language models (LLMs) demonstrate emergent and strong grammatical fluency, but can still violate the constraints of a language's grammar such as in code generation or formal reasoning. We investigate whether an explicit, lightweight grammar prior can improve syntactic correctness without modifying the weights of the model itself. We propose a grammar aware scoring framework that augments an LLM's sentence log-likelihood with a part of speech (POS) bigram priors model trained on tagged text. The resulting model acts similar to a product of experts, combining the semantic likelihood of a sentence from the LLM with the grammar structural plausibility of that sentence from the POS model. We evaluate this approach on the BLiMP benchmark of minimal grammatical pairs. Our results show that POS based priors consistently improve performance on several syntactic structures found in formal language (e.g. subject-verb agreement, intransitives, irregular past participles) while occasionally degrading complex structures (e.g. negative polarity items, principle A). Overall, we find that grammar priors are most effective when applied selectively, highlighting both the promise and limitation of a simple, shallow grammatical bias.

1 Introduction

Despite rapid advances in scale and training data, LLMs remain imperfect with respect to grammatical structure. While attention mechanisms can capture long range dependencies across a sentence well, it does not capture or explicitly enforce against syntactic constraints of a language. In this way, the grammar structure an LLM learns becomes purely emergent, leading to outputs that may be semantically plausible yet structurally ill formed. The gap can become problematic in downstream applications of an LLM where syntactic validity is critically tied to semantics, such as in programming languages, formal proofs, or math. In these domains, even minor grammar violations can invalidate an output. Moreover, for underrepresented languages in the training data, the syntactic generalization degrades even more, further motivating explicit mechanisms for grammatical structure steering.

Prior work has incorporated grammatical structure into language model decoding by explicitly constraining the output space, for example through Grammar Constrained Decoding [Zhou et al., 2023] and Grammar Aligned Decoding [Wang et al., 2024]. These approaches restrict generation to tokens licensed by a formal grammar and modify decoding to match the conditional language model distribution. While effective at enforcing well formed sentences, such methods impose a hard constraint on generation and tightly couple decoding to a specific formal grammar.

In contrast, we explore whether grammatical knowledge can be incorporated as a soft probabilistic prior, biasing an LLM toward syntactic plausible output while still allowing the base model to determine the semantics. Rather than restricting token generation or retraining the model, we combine an LLM with an interpretable POS bigram grammar via a product of experts. We empirically study its effect across a wide range of syntactic structures in the BLiMP benchmark.

2 Background & Related Work

Language modeling is typically formulated as next token prediction, where models are trained to maximize likelihood over large corpora and are expected to implicitly acquire syntactic structure from data. While modern LLMs exhibit strong grammar fluency, it remains an open question to what extent they internalize a language's grammar rules beyond surface level regularities.

To benchmark the abilities of language models to produce grammatically correct output, several targeted evaluation benchmarks have been proposed. The Corpus of Linguistic Acceptability (CoLA) consists of English sentences labeled as grammatical or ungrammatical, enabling acceptability classification and error analysis across grammar structures [Warstadt et al., 2019]. The Benchmark of Linguistic Minimal Pairs (BLiMP), extends this line of work by providing minimal pairs that isolate specific grammatical structures, allowing a more fine grained evaluation of syntactic generalization [Warstadt et al., 2023]. These benchmarks have revealed that language models learn some systematic grammatical patterns, such as word order and agreement, but struggle with constructions requiring more long distance patterns.

Beyond evaluation, several approaches have sought to incorporate grammatical structure directly into language model inference. Grammar Constrained Decoding (GCD) enforces hard constraints derived from formal grammars, restricting generation to syntactically valid token sequences [Zhou et al., 2023]. Grammar Aligned Decoding (GAD) similarly integrates grammatical structure by modifying decoding distributions to better match grammar conditioned models [Wang et al., 2024]. While effective at enforcing structure, these methods impose hard generation constraints and tightly couple decoding to a specific formal grammar.

Classical natural language processing has long employed probabilistic models as lightweight and interpretable representations of linguistic structure. In particular, POS tagging abstracts away lexical content while preserving grammatical patterns, and n-gram models over POS sequences capture local syntactic dependencies such as agreement and short range structure [Jurafsky and Martin, 2025]. More generally, combining multiple probabilistic models to jointly evaluate data can be formalized through a product of experts framework, in which agreement among experts sharpens the resulting distribution [Hinton, 1999].

3 Method

Building on probabilistic grammar modeling, we combine a pretrained LLM with a POS bigram grammar using a product of experts formulation, treating grammar as a soft probabilistic prior rather than a hard constraint like in GCD and GAD.

Let y = $(w_1, \ldots, w_T)$ denote a sentence of length T. A pretrained causal language model assigns a log-likelihood

$$s_{\text{LM}}(y) = \sum_{t=1}^{T} \log p_{\text{LM}}(w_t \mid w_{<t})$$

Given a POS tagging $(z_1, \ldots, z_T)$ of the sentence, a POS bigram grammar trained on tagged text assigns a log-likelihood

$$s_{\text{POS}}(y) = \sum_{t=1}^{T} \log p_{\text{POS}}(z_t \mid z_{t-1})$$

We combine these components using a product-of-experts scoring function,

$$s_{\text{GLM}}(y) = s_{\text{LM}}(y) + \lambda\, s_{\text{POS}}(y)$$

where $\lambda \geq 0$ controls the influence of the grammar prior. In probability space, this corresponds to multiplying the LLM likelihood by a POS grammar likelihood raised to the power λ. When $\lambda = 0$, the model reduces to the base LLM. We evaluate this grammar aware score in a sentence discrimination setting, where grammatical and ungrammatical sentence pairs are compared directly using s_{GLM}.

4 Experimentation & Results

4.1 Models and Grammar Priors

We evaluated our grammar aware scoring framework using Qwen3-1.7B, a pretrained casual language model, in inference mode only. No model parameters are updated. As the grammar expert, we used a POS bigram language model trained on POS tagged text derived from the WikiText2 corpus (`wikitext-2-raw-v1`). The POS tagging of this corpus was obtained using spaCy, and the bigram model was trained with Kneser-Ney (KN) smoothing to better address unseen POS bigrams. Used two different spaCy taggers, small neural network based (`en_core_web_sm`) and transformer based (`en_core_web_trf`), to construct two POS KN bigram models to account for sensitivity in tagging quality. The grammar weight was fixed to $\lambda = 0.5$, which provided a representative balance between language model and grammar scores in preliminary experiments.

4.2 Benchmarks and Evaluation

We evaluated exclusively on the BLiMP benchmark, which consists of minimal pairs of grammatical and ungrammatical English sentences covering 67 syntax subsets. Each subset contained controlled sentence pairs differing only in grammaticality. Using this benchmark, we looked for sentence discrimination accuracy, which was the proportion of minimal pairs for a certain syntax structure with a higher score over the grammatical sentences than the ungrammatical ones under the scoring function.

4.3 Results Across Syntactic Structures

Table 1: Mean sentence discrimination accuracy on BLiMP grouped by syntactic categories. POS bigram priors help local dependencies (subject/verb agreement, islands, intransitives, and irregular past participle verbs) but continue to underperform on NPI licensing, Principle A, and wh–that alternations.

Phenomenon Type	LM	POS Bigram (sm)	POS Bigram (trf)	Δ (trf)
Subject–verb Agreement	0.835	0.847	0.858	+0.023
Wh-movement	0.876	0.936	0.935	+0.059
Adjunct + wh Islands	0.800	0.864	0.866	+0.066
Intransitive	0.752	0.824	0.820	+0.068
Irregular Past Participle Verbs	0.804	0.872	0.872	+0.068
Quantifiers / Existential	0.788	0.813	0.813	+0.025
NPI Licensing	0.666	0.629	0.626	-0.039
Binding (Principle A)	0.778	0.706	0.708	-0.070
Wh–that Alternations	0.604	0.521	0.521	-0.083
Overall (67)	0.772	0.758	0.758	-0.014

Table 1 reports aggregated per syntax structure accuracies for the base LLM and grammar aware models using POS bigrams trained with Kneser–Ney smoothing. The resulting distribution of the POS grammar prior is non-uniform across the syntax sets.

Across all 67 minimal pairs, the grammar aware models show a small decrease in mean accuracy relative to the base LLM (-0.014 on average), with a median difference of 0.0. However, this aggregate obscures substantial gains on specific subsets. Approximately 30 of 67 sets improve under the grammar prior, with the largest gains observed for subject/verb agreement, wh-word movement, and existential-there constructions, including long distance subject gaps.

In contrast, we observe pronounced degradations on grammar sets that require hierarchical or semantic understanding to follow the formal grammar, such as negative polarity item licensing, Principle A binding, and wh-complementizer subsets. These results suggest that POS bigram grammars effectively capture local surface regularities but are not as well suited for modeling deeper structural dependencies.

Comparing POS models trained with the small NN and transformer based spaCy taggers yields only marginal differences, and neither variant impacts the overall negative mean effect. Full per subset

results are provided in Table A.1 as well as the top deltas from each spaCy tagger in Figure A.1 and Figure A.2.

4.4 Summary of Findings

Taken together, these results indicate that POS based grammar priors can meaningfully improve syntactic discrimination for certain classes of a grammar, but degrade performance for others. A single global grammar weight applied uniformly across constructions masks these differences, highlighting the need for selective or adaptive application of grammatical priors.

5 Discussion

Our experiments reveal a nuanced take on a grammar aware language model using POS priors. While the experiment on the BLiMP benchmark showed that it could substantially improve performance on certain syntax, it could also introduce lower results on others, highlighting both the possibilities and limitations of a POS bigram expert.

5.1 Where POS Priors Help

We observe consistent gains on the minimal grammar pairs where the syntax is guided by local or surface level regularities, including subject/verb agreement, determiner/noun agreement, intransitive/transitive alternations, and several wh-movement pairs. These improvements are particularly pronounced for long distance wh-subject gaps and existential constructions, suggesting that even simple POS priors can still boost LLM predictions when those syntax rely on consistent local tag patterns (certain long distance dependencies can still be affected by a local grammar change).

This aligns with classical NLP findings that POS n-grams effectively model agreement and short range dependencies [Jurafsky and Martin, 2025], suggesting that LLMs may underweight certain structural signals that are easily recoverable at the POS level.

5.2 Where POS Priors Hurt

In contrast, the POS grammar prior degrades performance on minimal pairs that require hierarchical structure or contextual and semantic understanding. These include Principle A binding constraints, negative polarity item licensing, and wh–that alternations with gaps. In such cases, simple bigram POS sequences fail to capture the relevant grammatical dependencies, and the grammar expert favors syntactically frequent but linguistically invalid continuations. These regressions persist across different taggers and smoothing strategies, indicating that the limitation lies in the representational capacity and simplicity of a POS bigram.

5.3 Limits of Uniform Grammar Weighting

A key finding of the benchmarking is that applying a single global grammar weight across all syntactic minimal pairs is suboptimal. While the POS prior improves roughly half of the BLiMP subsets, the remaining subsets experience regressions large enough to produce a slightly negative average effect overall. This suggests that grammatical structure is not uniformly beneficial and that grammar aware modeling must be adaptive to the current syntax and context being evaluated.

5.4 Implications for Grammar Aware Language Modeling

Framing grammar integration as a product of experts provides a flexible alternative to grammar constrained decoding, preserving the original token space and avoiding retraining. However, our results indicate that a POS bigram grammar expert should be treated as modular, selectively applied component rather than a universally beneficial constraint. POS level grammar captures important regularities, but its limitations show the need for richer or higher level grammatical representation.

6 Conclusion

We investigated whether lightweight grammatical structure can be incorporated into large language models as a soft probabilistic prior, rather than as a hard decoding constraint. By combining a pretrained LLM with a POS bigram grammar in a product of experts formulation, we evaluated grammar aware scoring across 67 syntactic sets in the BLiMP benchmark.

Our results demonstrate that POS based grammar priors can meaningfully improve performance on a subset of syntactic constructions, particularly those governed by local structural regularities. At the same time, they expose clear limitations on syntax requiring hierarchical or semantic understanding. Taken together, these findings suggest that even a simple grammatical abstraction can complement LLMs, but only when applied selectively and with awareness of their representational limits.

7 Future Work

The limitations observed for POS based grammar priors suggest that richer syntactic representations are needed to capture constraints beyond local category transitions. A natural extension is to incorporate dependency aware grammar priors that more directly reflect hierarchical relationships. For example, one could train n-gram models over dependency labels or over structured tuples such as (head POS, dependency label, child POS), extracted from dependency parsed corpora. Such representations may better capture syntax that POS bigrams systematically fail on, including binding constraints (relative positions of pronouns and antecedents), wh–that alternations (presence or absence of clause linking dependencies), and negative polarity item licensing (structural relationships between licensors and NPIs). Sentences could be tagged using a dependency parser, transformed into linearized dependency sequences or paths, and scored using a Kneser–Ney smoothed n-gram model. The resulting dependency level likelihood could then be combined with the base language model in the same product of experts framework explored in this work. This approach would retain the interpretability and modularity of grammar priors while providing a syntactic bias more closely aligned with the grammatical constraints underlying the challenging BLiMP subsets.

References

G. E. Hinton. Products of experts. *Proceedings of the Ninth International Conference on Artificial Neural Networks*, 1999. URL `https://www.cs.toronto.edu/~fritz/absps/icann-99.pdf`.

D. Jurafsky and J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*. 3rd edition, 2025. URL `https://web.stanford.edu/~jurafsky/slp3/`. Online manuscript released August 24, 2025.

X. Wang et al. Grammar-aligned decoding. *arXiv preprint arXiv:2405.21047*, 2024. URL `https://arxiv.org/abs/2405.21047`.

A. Warstadt, A. Singh, and S. R. Bowman. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*, 2019. URL `https://arxiv.org/abs/1805.12471`.

A. Warstadt, A. Parrish, H. Liu, A. Mohananey, W. Peng, S.-F. Wang, and S. R. Bowman. Blimp: The benchmark of linguistic minimal pairs for english. *arXiv preprint arXiv:1912.00582*, 2023. URL `https://arxiv.org/abs/1912.00582`.

J. Zhou et al. Grammar-constrained decoding for structured text generation. *arXiv preprint arXiv:2305.13971*, 2023. URL `https://arxiv.org/abs/2305.13971`.

A Full Results

Table A.1: BLiMP subset accuracies for the base LM and POS bigram grammar priors.

Subset	LM	POS Bigram (sm)	POS Bigram (trf)	Δ (trf)
adjunct_island	0.856	0.900	0.904	+0.048
anaphor_gender_agreement	0.988	0.988	0.988	+0.000
anaphor_number_agreement	0.980	0.980	0.980	+0.000
animate_subject_passive	0.676	0.668	0.668	-0.008
animate_subject_trans	0.748	0.760	0.760	+0.012
causative	0.656	0.644	0.644	-0.012
complex_NP_island	0.600	0.600	0.604	+0.004
coordinate_structure_constraint_complex_left_branch	0.860	0.752	0.744	-0.116
coordinate_structure_constraint_object_extraction	0.804	0.736	0.744	-0.060
determiner_noun_agreement_1	0.980	0.980	0.980	+0.000
determiner_noun_agreement_2	0.908	0.916	0.916	+0.008
determiner_noun_agreement_irregular_1	0.912	0.900	0.900	-0.012
determiner_noun_agreement_irregular_2	0.904	0.892	0.892	-0.012
determiner_noun_agreement_with_adj_2	0.864	0.876	0.876	+0.012
determiner_noun_agreement_with_adj_irregular_1	0.888	0.900	0.900	+0.012
determiner_noun_agreement_with_adj_irregular_2	0.864	0.868	0.868	+0.004
determiner_noun_agreement_with_adjective_1	0.960	0.952	0.952	-0.008
distractor_agreement_relational_noun	0.808	0.720	0.724	-0.084
distractor_agreement_relative_clause	0.616	0.512	0.516	-0.100
drop_argument	0.768	0.824	0.824	+0.056
ellipsis_n_bar_1	0.772	0.756	0.780	+0.008
ellipsis_n_bar_2	0.836	0.844	0.828	-0.008
existential_there_object_raising	0.736	0.712	0.712	-0.024
existential_there_quantifiers_1	0.944	0.948	0.948	+0.004
existential_there_quantifiers_2	0.568	0.680	0.684	+0.116
existential_there_subject_raising	0.880	0.876	0.876	-0.004
expletive_it_object_raising	0.756	0.736	0.736	-0.020
inchoative	0.668	0.672	0.676	+0.008
intransitive	0.752	0.824	0.820	+0.068
irregular_past_participle_adjectives	0.360	0.396	0.400	+0.040
irregular_past_participle_verbs	0.804	0.872	0.872	+0.068
irregular_plural_subject_verb_agreement_1	0.916	0.920	0.924	+0.008
irregular_plural_subject_verb_agreement_2	0.784	0.772	0.792	+0.008
left_branch_island_echo_question	0.788	0.812	0.832	+0.044
left_branch_island_simple_question	0.952	0.844	0.836	-0.116
matrix_question_npi_licensor_present	0.716	0.508	0.484	-0.232
npi_present_1	0.432	0.432	0.432	+0.000
npi_present_2	0.520	0.520	0.520	+0.000
only_npi_licensor_present	0.804	0.804	0.804	+0.000
only_npi_scope	0.716	0.640	0.652	-0.064
passive_1	0.768	0.740	0.740	-0.028
passive_2	0.864	0.832	0.836	-0.028
principle_A_c_command	0.764	0.756	0.756	-0.008
principle_A_case_1	0.988	0.988	0.988	+0.000
principle_A_case_2	0.884	0.768	0.772	-0.112
principle_A_domain_1	0.988	0.988	0.988	+0.000
principle_A_domain_2	0.724	0.720	0.724	+0.000
principle_A_domain_3	0.508	0.512	0.516	+0.008
principle_A_reconstruction	0.592	0.212	0.212	-0.380
regular_plural_subject_verb_agreement_1	0.828	0.868	0.880	+0.052
regular_plural_subject_verb_agreement_2	0.812	0.828	0.836	+0.024
sentential_negation_npi_licensor_present	0.896	0.896	0.896	+0.000
sentential_negation_npi_scope	0.576	0.600	0.596	+0.020
sentential_subject_island	0.376	0.332	0.284	-0.092
superlative_quantifiers_1	0.920	0.984	0.988	+0.068

Continued on next page

Subset	LM	POS Bigram (sm)	POS Bigram (trf)	Δ (trf)
superlative_quantifiers_2	0.680	0.680	0.668	-0.012
tough_vs_raising_1	0.592	0.576	0.572	-0.020
tough_vs_raising_2	0.900	0.900	0.900	+0.000
transitive	0.808	0.812	0.812	+0.004
wh_island	0.744	0.828	0.828	+0.084
wh_questions_object_gap	0.904	0.940	0.936	+0.032
wh_questions_subject_gap	0.924	0.980	0.980	+0.056
wh_questions_subject_gap_long_distance	0.932	0.996	0.996	+0.064
wh_vs_that_no_gap	0.968	1.000	1.000	+0.032
wh_vs_that_no_gap_long_distance	0.992	0.996	0.996	+0.004
wh_vs_that_with_gap	0.252	0.040	0.040	-0.212
wh_vs_that_with_gap_long_distance	0.204	0.048	0.048	-0.156

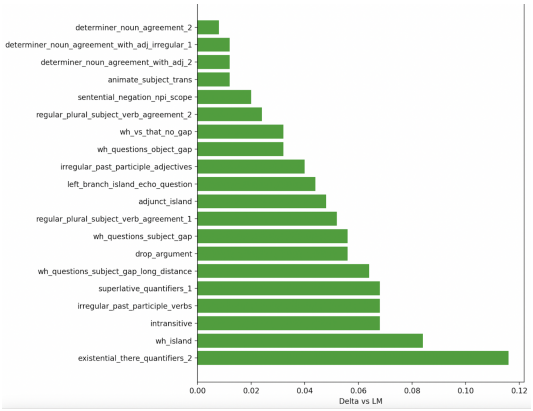

Figure A.1: LM vs POS Bigram (trf spaCy tagger) Top Positive Deltas

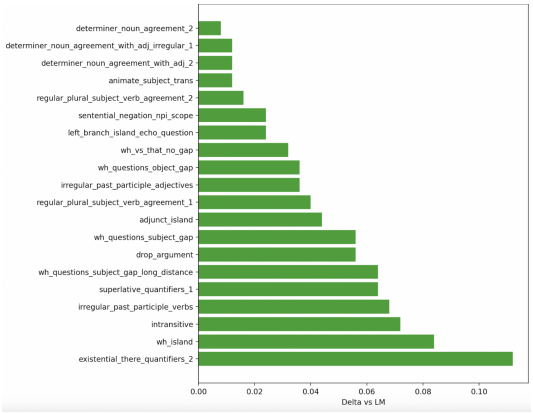

Figure A.2: LM vs POS Bigram (sm spaCy tagger) Top Positive Deltas