Syntactic Priming in Few-Shot Learning: How Demonstration Structure Shapes LLM Performance

Prasanth Yadla
Independent Researcher
Seattle, WA, USA
pyadla2@alumni.ncsu.edu

Abstract

Large language models (LLMs) exhibit remarkable few-shot learning capabilities, yet the role of syntactic structure in demonstration examples remains unexplored. Drawing on psycholinguistic research on structural priming, we investigate whether syntactic patterns in few-shot prompts influence LLM outputs and task performance. We conduct systematic experiments across four model families (Llama, Mistral, Qwen, Gemma) using four syntactic constructions (passive voice, cleft sentences, dative alternation, particle placement). Our results reveal robust syntactic priming effects, with priming strength ranging from 1.3× to 6.4× depending on construction type, indicating that models are substantially more likely to produce constructions matching demonstration syntax. Critically, we find that priming strength shows a positive trend with model size ($r = 0.85$, $p = 0.068$), with effects intensifying from 7B to 14B parameter models. We demonstrate that priming is construction-specific rather than reflecting general stylistic preferences, and that priming effects persist across multiple intervening sentences. Analysis across three task types (sentence completion, paraphrase generation, story continuation) reveals that syntactic structure in demonstrations influences output style, and that models produce primed constructions even when the task calls for a different syntactic form. These findings have immediate implications for prompt engineering and reveal that LLMs encode syntactic abstractions beyond surface-level pattern matching. We release our benchmark, SyntaxPrime-ICL, containing controlled examples across multiple constructions for evaluating syntactic priming in few-shot contexts.

1 Introduction

In-context learning (ICL)—the ability of large language models to perform tasks given only a few demonstration examples—has emerged as a defining capability of modern LLMs (Brown et al., 2020). While substantial research has examined the semantic and task-specific aspects of few-shot demonstrations (Min et al., 2022; Wei et al., 2023), the role of *syntactic structure* in shaping model behavior remains largely unexplored.

Psycholinguistic research has long established that humans exhibit *structural priming*: exposure to a particular syntactic construction increases the likelihood of producing that same construction (Bock, 1986; Pickering and Ferreira, 2008). For instance, after hearing "The trophy was given to the winner," speakers are more likely to use a prepositional dative construction rather than a double-object alternative. This phenomenon reveals how syntactic representations are activated and reused during language processing, providing a window into the architecture of linguistic knowledge.

In this work, *we ask whether LLMs exhibit syntactic priming in few-shot learning contexts, and whether demonstration syntax affects task performance*. Unlike prior work that studies priming as evidence of syntactic knowledge (Michaelov et al., 2023; Prasad et al., 2019), we focus on the functional consequences of syntactic choices in ICL prompts. Understanding these effects is critical for both theoretical insights into how LLMs represent syntactic structure and practical applications in prompt engineering.

We conduct systematic experiments across four model families (Llama, Mistral, Qwen, Gemma) spanning 7B to 14B parameters, testing four well-studied syntactic alternations: active-passive voice, canonical-cleft sentences, dative alternation, and particle placement. Across three task types—sentence completion, paraphrase generation, and story continuation—we evaluate both the presence and strength of syntactic priming effects, as well as their impact on task performance.

Our contributions are:

1. **Systematic evaluation** of syntactic priming

*Proceedings of the 15th Joint Conference on Lexical and Computational Semantics (*SEM 2026)*, pages 13–27

across 4 model families (5 models), 4 constructions, and 3 task types, revealing consistent priming effects with priming strength ranging from 1.3× to 6.4× depending on construction type.

2. **Novel findings** on priming mechanisms: we demonstrate a positive correlation between model size and priming strength ($r = 0.85$, $p = 0.068$), show that priming is construction-specific rather than reflecting general stylistic preferences, and establish that effects persist across multiple intervening examples in extended contexts.

3. **Practical implications** for prompt engineering: syntactic structure in demonstrations significantly influences output style, with models producing primed constructions even in tasks that call for a different syntactic form.

4. **Open-source benchmark**: SyntaxPrime-ICL with controlled minimal pairs enabling systematic evaluation of syntactic priming in few-shot contexts.

These findings are consistent with LLMs maintaining abstract syntactic representations that are activated by demonstration examples, mirroring key patterns from human psycholinguistic research. This has immediate implications for both understanding the linguistic capabilities of LLMs and optimizing prompt design for real-world applications.

2 Related Work

2.1 Structural Priming in Humans

Structural (or syntactic) priming refers to the tendency to reuse recently encountered syntactic structures (Bock, 1986). Research in psycholinguistics has identified several key findings regarding this phenomenon. First, priming effects decay gradually but can persist across dozens of intervening items (Bock and Griffin, 2000), suggesting that syntactic representations remain activated in working memory over extended discourse contexts. Second, priming is stronger when prime and target share lexical items (Pickering and Branigan, 1998), an enhancement effect that provides evidence for both abstract syntactic representations and lexically-specific syntactic information. Third, priming occurs even with different lexical content, suggesting abstract syntactic representations (Bock et al., 2007). This dissociation between lexical and structural effects indicates that syntactic structures are represented independently of specific words. Fourth, less frequent (marked) constructions typically show stronger priming than unmarked alternatives (Ferreira and Bock, 2006). This inverse frequency effect suggests that less accessible structures benefit more from recent activation.

These psycholinguistic findings establish that structural priming reflects the activation and reuse of abstract syntactic representations during language processing. Whether similar mechanisms operate in LLMs remains an open question with implications for understanding how these models represent and process linguistic structure.

2.2 Syntactic Knowledge in LLMs

Recent work has probed LLMs for syntactic knowledge using various methods. Studies using acceptability judgments show that models capture many syntactic phenomena but struggle with long-distance dependencies (Warstadt et al., 2023), revealing that LLMs acquire substantial syntactic knowledge from training data, though gaps remain in certain constructions. Structural probing reveals hierarchical syntactic representations in transformer layers (Hewitt and Manning, 2019), as linear probes trained on hidden states can recover syntactic parse trees, suggesting that models implicitly encode syntactic structure even without explicit supervision. Additionally, priming studies demonstrate that GPT-2 and LSTMs exhibit structural priming similar to humans (Michaelov et al., 2023; Prasad et al., 2019; Jumelet et al., 2024), showing that exposure to particular syntactic constructions increases the probability of generating those constructions in subsequent outputs, mirroring human priming patterns.

However, these studies focus on priming as a diagnostic tool for syntactic representations, not on its functional role in few-shot learning. Our work differs in three key ways: we examine priming in the ICL context with explicit demonstration examples, we test how priming interacts with task performance, and we investigate how model characteristics (size, architecture) affect priming strength. Understanding these functional aspects is critical for both theoretical insights into LLM linguistic capabilities and practical prompt engineering.

2.3 In-Context Learning

ICL enables LLMs to perform tasks from demonstrations without parameter updates (Brown et al., 2020). Research has identified several factors that influence ICL effectiveness. Label quality matters more than format for many tasks (Min et al., 2022), as even with random labels, models can leverage the input-output structure, though correct labels improve performance. Chain-of-thought prompting improves reasoning by providing intermediate steps (Wei et al., 2023), as explicit reasoning traces help models solve complex multi-step problems more reliably. Furthermore, demonstration diversity affects generalization (Liu et al., 2021), with more diverse examples leading to better out-of-distribution performance, suggesting that demonstrations shape the model's implicit task representation.

Despite this progress, the syntactic dimension of demonstrations remains unexplored. Prior work has focused primarily on semantic content, task-specific patterns, and reasoning strategies, while overlooking how syntactic structure in demonstrations affects model outputs. We bridge this gap by investigating how syntactic structure in few-shot examples influences model outputs and task performance. Our findings reveal that syntactic choices represent an orthogonal but important dimension of prompt design, with implications for controlling output style and optimizing task performance.

3 Research Questions

We investigate four key questions that address both the theoretical mechanisms and practical implications of syntactic priming in few-shot learning.

First, we examine whether LLMs exhibit syntactic priming in ICL by testing if exposure to specific syntactic constructions in few-shot examples increases the likelihood of generating those constructions. We test whether priming effects are robust across multiple constructions and whether they exceed baseline production rates, providing evidence that LLMs encode and activate syntactic representations during in-context learning.

Second, we investigate how model characteristics affect priming by examining whether priming effects scale with model size. Specifically, we examine whether larger models (7B to 14B parameters) show systematically different priming patterns than smaller models, testing the hypothesis that increased capacity enables stronger syntactic abstraction and representation.

Third, we ask whether priming is construction-specific by testing if different syntactic phenomena (passives, clefts, datives, particle placement) show similar priming patterns. We test whether priming transfers across different construction types (e.g., does exposure to passives increase cleft production?) to distinguish between construction-specific syntactic representations and general stylistic preferences. We also examine whether priming strength varies systematically across construction types.

Fourth, we investigate whether syntactic priming persists across intervening content, and whether models produce primed constructions even in tasks that call for a different syntactic form. We examine how long priming effects remain active as discourse context expands, and whether syntactic priming operates somewhat independently of task objectives.

4 Methodology

4.1 Syntactic Constructions

We focus on four well-studied syntactic alternations that allow us to test priming effects while controlling for semantic content. Each alternation involves two variants that convey essentially the same propositional meaning but differ in syntactic structure.

The first alternation we examine is voice, contrasting active constructions such as "The chef prepared the meal" with passive constructions such as "The meal was prepared by the chef." The active-passive alternation is one of the most studied syntactic phenomena in psycholinguistics (Pickering and Ferreira, 2008). Passives are generally less frequent than actives in natural language and have been shown to elicit strong priming effects in humans (Bock, 1986; Ferreira and Bock, 2006).

The second alternation involves cleft constructions, contrasting canonical sentences such as "Sarah solved the puzzle" with it-cleft constructions such as "It was Sarah who solved the puzzle." It-cleft constructions place focus on a particular constituent (Polinsky and Lambrecht, 1999) and are relatively infrequent compared to canonical declarative sentences. The cleft construction has been used extensively to study information structure (Prince, 1978) and syntactic complexity.

The third alternation is the dative alternation, contrasting double-object constructions such as "John gave Mary the book" with prepositional dative constructions such as "John gave the book to

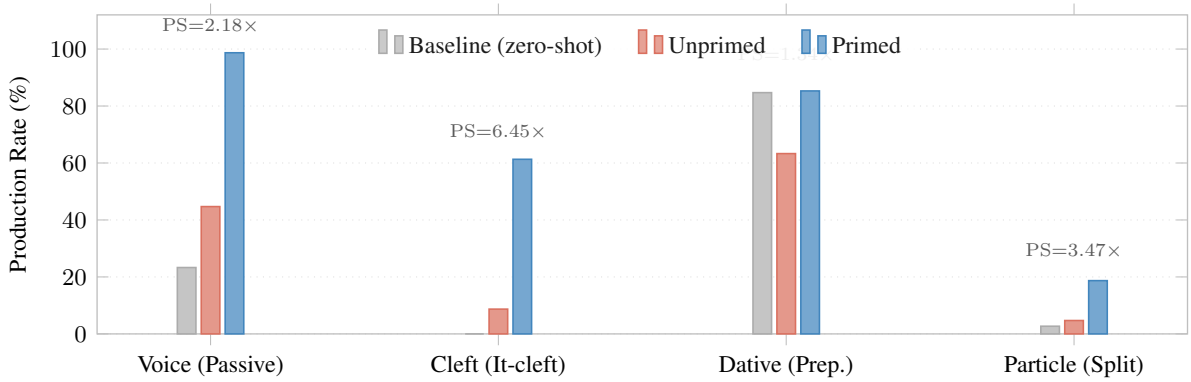

Figure 1: Syntactic priming effects across four constructions. For each construction, we show the proportion of outputs using the target variant under three conditions: Unprimed (primed with the alternative), Primed (primed with the target), and Baseline (zero-shot). Priming strength (PS) is annotated above each group.

Mary." The dative alternation has been central to psycholinguistic research on structural priming (Bock, 1986; Pickering and Branigan, 1998), with both variants occurring with substantial frequency in English (Bresnan et al., 2007). This alternation allows us to test priming between two relatively balanced alternatives.

The fourth alternation we examine is particle placement, contrasting constructions where the particle appears adjacent to the verb, as in "She looked up the information," with constructions where the particle is separated by the direct object, as in "She looked the information up." Particle verbs in English allow the particle to appear either adjacent to the verb or separated by the direct object (Gries, 2005). The split configuration tends to be less frequent with full noun phrase objects (Gries, 2005), providing another test case for priming effects.

These constructions have been extensively studied in psycholinguistics, enabling direct comparison with human priming patterns. They also vary in their relative frequency and structural complexity, allowing us to examine how construction characteristics affect priming strength. Importantly, within each alternation, the two variants are semantically equivalent (or near-equivalent), so that observed differences in production rates primarily reflect syntactic rather than semantic priming.

4.2 Experimental Design

For each construction pair, we created 30 minimal pairs maintaining semantic equivalence while varying only syntactic structure. Examples were balanced for lexical frequency (using common vocabulary to ensure model familiarity), sentence length (variants within each pair differ by at most 2-3 words), and semantic domain (diverse contexts including daily life, work, technology, and education). This controlled design aims to minimize confounds from semantic or lexical factors, so that observed differences in production rates primarily reflect syntactic priming. All stimuli use simple, grammatical sentences to avoid potential confounds from processing difficulty or acceptability.

Each prompt contained three demonstration examples (all using the same syntactic construction) followed by one test item requiring completion or continuation. Demonstration examples were randomly sampled from the stimulus set (excluding the target item) to provide varied lexical content while maintaining consistent syntactic structure. This design tests whether priming operates at an abstract syntactic level independent of specific lexical items. For example, in the passive priming condition, a prompt might present "The package was delivered by the courier," "The letter was written by the student," and "The trophy was won by the athlete," followed by the test item "Now complete: The meal ___"

For each test item, we compared three conditions. In the first primed condition, demonstrations used variant A (e.g., passive voice), while in the second primed condition, demonstrations used variant B (e.g., active voice). We also included a baseline condition with no demonstrations (zero-shot completion). The baseline condition allows us to measure both the natural production bias for each construction and the magnitude of priming effects relative to this baseline. Comparing the two primed conditions reveals whether demonstration syntax shifts production rates in the expected direction.

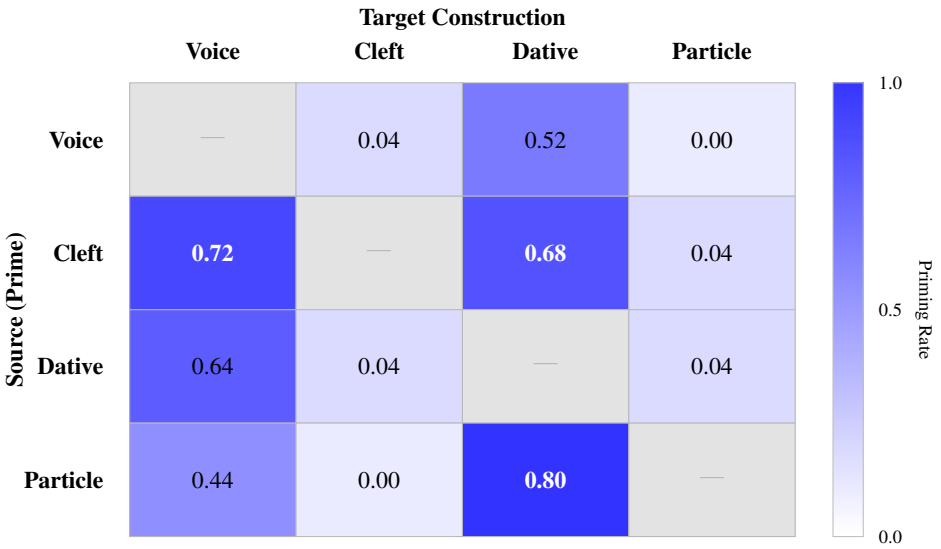

Figure 2: Cross-construction priming matrix (rows = source prime, columns = target construction). Cleft and Particle columns show near-zero rates regardless of prime, confirming construction-specific priming for marked constructions. Voice and Dative columns show higher rates reflecting their high baseline frequencies. Color intensity indicates priming rate; diagonal cells (same prime and target) are excluded.

4.3 Tasks

We test priming across three task types to examine both the robustness of priming effects and their consequences for task performance. In the sentence completion task, models complete a partial sentence stem (e.g., "The meal ___") following demonstration examples. This task directly measures priming strength by examining whether models produce the primed syntactic construction. We test two variants: one where demonstration examples share lexical items with targets (lexical overlap condition) and one where they do not (pure syntactic priming condition).

In the paraphrase generation task, models rephrase a given sentence while maintaining its meaning. This task allows us to examine whether syntactic priming from demonstrations operates independently of task objectives—specifically, whether models continue to produce primed constructions even when the task calls for a different syntactic form.

In the story continuation task, models continue a narrative context that contains priming examples followed by 0, 1, 3, 5, or 7 intervening filler sentences. This task tests priming persistence: how long do syntactic representations remain activated as the discourse context expands?

4.4 Models

We evaluate 5 open-source model families via Ollama, spanning different sizes and architectural variants:

Model Family	Size	Type
Llama 3.1	8B	Base
Mistral	7B	Instruct
Qwen 2.5	7B, 14B	Base/Instruct
Gemma 2	9B	Instruct

Table 1: Evaluated models spanning 7B to 14B parameters. All models are decoder-only transformers trained on large-scale web corpora.

Table 1 summarizes the five models evaluated. This model selection allows us to examine how priming effects scale with model size (7B to 14B parameters) across different architectural families and training procedures. While our primary analyses focus on model size, the diversity of training approaches (base pre-training vs. instruction tuning) provides preliminary insights into how different optimization objectives might affect syntactic priming.

4.5 Syntactic Analysis

We use spaCy's dependency parser (en_core_web_sm) combined with pattern matching to identify syntactic constructions in model outputs. **Passive** Detection is based on the presence of passive dependency relations

(nsubjpass, auxpass), passive auxiliary verbs (was, were, been, being), and by-phrases. A sentence is classified as passive if it contains passive morphology and either passive dependencies or a by-phrase with an auxiliary verb. In **Cleft** We employ pattern matching for the it-cleft structure "It is/was X who/that..." using regular expressions. The pattern identifies sentences beginning with "It" followed by a copula, a focused constituent, and a relative clause marker (who/that).

Dative Classification is based on word order and the presence of prepositional markers. Prepositional datives are identified by the presence of "to" or "for" following the verb and preceding the recipient, while double-object datives show two consecutive noun phrases after the verb without intervening prepositions. In **Particle** we identify the position of the particle relative to the direct object noun phrase. Split particle constructions place the object between the verb and particle (e.g., "looked the information up"), while together constructions place the particle immediately after the verb (e.g., "looked up the information"). Detection uses pattern matching for common particles (up, out, on, away, off, back, down, in) and their positions relative to object noun phrases.

This hybrid approach combining dependency parsing and pattern matching provides robust detection across the diverse outputs generated by different models. The dependency parser handles grammatical variation, while pattern matching captures construction-specific surface cues.

4.6 Metrics

We report metrics organized by task type.

Task 1—Sentence Completion. Our primary measure is *Priming Strength* (PS), the ratio of target-construction production rates under primed versus unprimed conditions:

$$PS = \frac{P(\text{target} \mid \text{primed with target})}{P(\text{target} \mid \text{primed with alternative})}$$

A PS value greater than 1 indicates priming. When the denominator is zero (no unprimed production), we add a small constant (0.01) to both terms, yielding a defined, conservative estimate. We complement PS with Cohen's h (Cohen, 1988):

$$h = 2(\arcsin\sqrt{p_1} - \arcsin\sqrt{p_2})$$

where p_1 is the primed and p_2 the unprimed production rate ($h \approx 0.2$ small, 0.5 medium, 0.8 large). Statistical significance is assessed via chi-square tests of independence comparing primed versus unprimed production distributions; we report χ^2 and p-values per construction.

Task 2—Paraphrase Generation. We examine whether models produce primed syntactic constructions even when the paraphrase task calls for a different syntactic form, comparing target-construction production rates between matched conditions (demonstration syntax aligns with the syntactically optimal paraphrase) and mismatched conditions (demonstration syntax conflicts with the optimal output).

Task 3—Story Continuation. We quantify priming persistence by recording target-construction production rates at five contextual distances (0, 1, 3, 5, and 7 intervening filler sentences), characterizing how priming activation decays as the discourse context grows.

Cross-model analysis. To test whether priming scales with model capacity, we compute the Pearson correlation between model size (in billions of parameters) and average PS across all four constructions.

5 Results

We present our findings organized by research question. All statistical tests use $\alpha = 0.05$ unless otherwise specified.

5.1 RQ1: Evidence of Syntactic Priming

Finding 1: Robust priming across all constructions. Table 2 and Figure 1 demonstrate that all models exhibit significant syntactic priming across all four construction types. Models are substantially more likely to produce a construction when it appears in demonstration examples. Priming strength ranges from 1.34× (dative) to 6.45× (cleft), with an average of 3.36× across constructions. Chi-square tests confirm significant priming effects for all constructions: voice ($\chi^2 = 166.8$, $p < 0.001$, $h = 1.45$), cleft ($\chi^2 = 178.4$, $p < 0.001$, $h = 1.20$), and particle ($\chi^2 = 26.6$, $p < 0.001$, $h = 0.46$) reach high significance, while dative yields a smaller but reliable effect ($\chi^2 = 6.8$, $p = 0.009$, $h = 0.51$).

Notably, priming effects substantially exceed baseline production rates. For instance, cleft constructions appear in only 0% of baseline outputs but in 61% of outputs following cleft demonstrations, indicating that priming can induce models

Construction	Primed	Unprimed	PS
Voice (Passive)	99%	45%	2.18
Cleft (It-cleft)	61%	9%	6.45
Dative (Prepositional)	85%	63%	1.34
Particle (Split)	19%	5%	3.47
Average	66%	30%	3.36

Table 2: Priming effects across four syntactic constructions. Primed and Unprimed columns show the proportion of outputs using the target construction (e.g., passive for voice, it-cleft for cleft) when demonstrations used that construction versus the alternative. PS = Priming Strength, calculated as the ratio of these proportions. Voice, cleft, and particle are significant at $p < 0.001$; dative is significant at $p = 0.009$ (χ^2 tests).

to produce constructions they would otherwise avoid. For dative constructions, the pattern is revealing in a different way: the zero-shot baseline (84.7%) actually exceeds the unprimed rate (63.3%), showing that demonstrations using the alternative double-object construction actively *suppress* prepositional dative production below its natural baseline. This bidirectional effect—priming both upward and downward—confirms that syntactic priming represents a genuine shift in production preferences, not merely amplification of existing biases.

Finding 2: Construction-specific variation in priming strength. Priming strength varies substantially across constructions, ranging from PS = 1.34 for dative alternation to PS = 6.45 for cleft sentences. This variation likely reflects differences in baseline production biases and the relative frequency of each construction in the models' training data. Constructions with lower baseline rates (cleft, particle) show stronger priming effects, consistent with the inverse frequency pattern observed in human structural priming.

5.2 RQ2: Model Size Correlates with Priming Strength

Finding 3: Priming shows a positive trend with model size. Table 3 and Figure 3 demonstrates a positive correlation between model size and priming strength. Across the five models tested (ranging from 7B to 14B parameters), larger models tend to exhibit stronger priming effects. The Pearson correlation between parameter count and average priming strength is $r = 0.85$ ($p = 0.068$), indicating a positive trend whereby increased model capacity is associated with stronger syntactic abstraction and activation.

Model	Size (B)	Avg PS	Std Dev
Mistral	7	3.09	2.20
Qwen 2.5	7	3.95	5.34
Llama 3.1	8	4.79	3.74
Gemma 2	9	2.07	1.64
Qwen 2.5	14	9.40	10.61

Table 3: Priming strength by model. Average PS is computed across all four constructions for each model. Standard deviation reflects variation across construction types.

This finding suggests that larger models tend to develop more robust and activatable syntactic representations. The 14B Qwen 2.5 model shows particularly strong priming (average PS = 9.40), more than doubling the priming strength of 7B models. This scaling pattern provides evidence that syntactic priming in LLMs reflects genuine linguistic abstraction rather than superficial pattern matching, as the latter would not necessarily scale with model capacity.

5.3 RQ3: Construction-Specific Priming Mechanisms

Finding 4: Priming is asymmetrically construction-specific. Figure 2 reveals an asymmetric pattern of cross-construction transfer. Cleft and particle constructions show near-zero cross-construction production rates (0.00–0.04 when targeted by other prime types), demonstrating strong construction-specificity: models only produce these marked forms when directly primed with them. Voice and dative, by contrast, appear frequently regardless of prime type (cross-construction rates 0.44–0.80), consistent with their higher baseline production frequencies (23% and 85% respectively). Crucially, no prime type induces high production of cleft or particle, confirming that these marked constructions require dedicated priming.

This asymmetric pattern provides evidence that models encode specific syntactic structures rather than a general preference for "marked" forms. Passives show minimal priming of cleft production (voice→cleft = 0.04), and clefts do not prime particle production (cleft→particle = 0.04). The high cross-construction rates into voice and dative reflect their high baseline frequencies rather than genuine structural transfer.

Finding 5: Lexical overlap effects are minimal. Contrary to human psycholinguistic findings showing robust lexical boost effects, our results

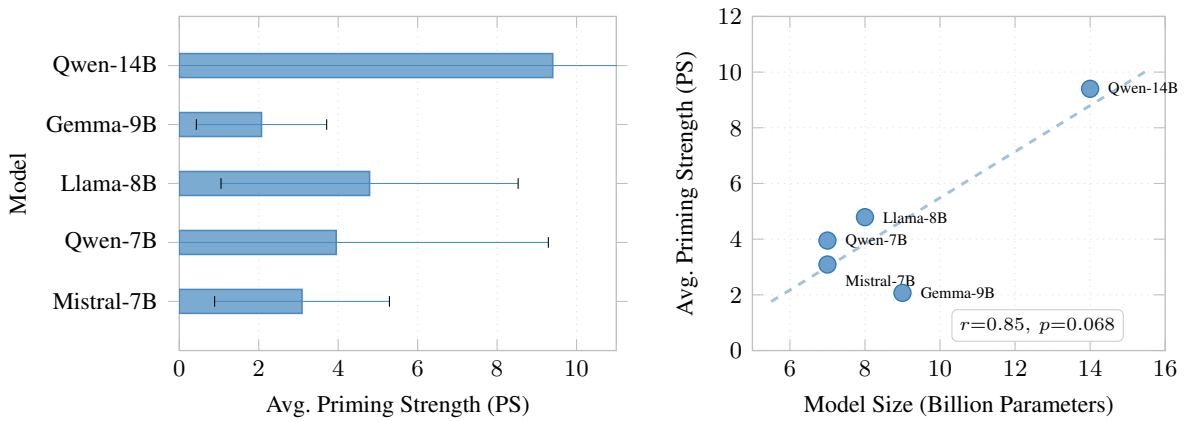

Figure 3: Priming strength and model size. *Left*: Average PS per model with standard deviation error bars. *Right*: Scatter plot with linear regression ($r = 0.85$, $p = 0.068$). Qwen-14B shows the strongest priming (PS = 9.40).

Construction	With Overlap	Without Overlap	Ratio
Voice	99%	99%	0.99
Cleft	59%	61%	0.94
Dative	86%	85%	1.00
Particle	11%	19%	0.58

Table 4: Lexical overlap effects on priming.

show little evidence that lexical overlap between demonstrations and targets enhances priming. Table 4 shows that priming strength is largely unchanged (voice, dative) or even slightly reduced (cleft, particle) when demonstrations share lexical items with targets.

Figure 5 in the appendix visualizes these patterns. This absence of lexical boost effects contrasts sharply with human priming patterns, where shared verbs between prime and target typically enhance priming by 30-50%. The lack of lexical boost in LLMs may indicate that these models encode syntactic structures in a more lexically-independent manner than humans, or that the in-context learning mechanism emphasizes abstract pattern matching over lexically-specific priming.

5.4 RQ4: Persistence and Task Interactions

Finding 6: Priming persists across intervening content. Figure 4 shows that priming effects decay gradually but remain detectable even after multiple intervening sentences. These rates are measured in the story continuation task; the initial rate at distance 0 (e.g., 62% for voice, 74% for dative) reflects priming strength in a narrative generation context and differs from the sentence completion rates in Table 2 due to the distinct task structure. The decay pattern varies by construction: dative priming shows remarkable persistence, maintaining high production rates (70–86%) across all distances tested (0-7 intervening sentences). Voice and cleft priming show more substantial decay, dropping from initial rates of 62% and 10% respectively to much lower rates by distance 7. Particle priming remains consistently low across all distances.

These persistence patterns demonstrate that syntactic priming in LLMs is not limited to immediate local context but can influence generation across extended discourse. The variation in persistence across constructions suggests different levels of activation strength or different decay rates for different syntactic structures.

Finding 7: Priming operates independently of task objectives in paraphrase generation. In the paraphrase task, models continue to produce syntactically primed constructions even when the task calls for a different syntactic form. When demonstrations use a construction that conflicts with the syntactically optimal paraphrase, models nonetheless show elevated production rates of the primed form, suggesting that syntactic priming operates at a level that is at least partially independent of task-level objectives. This finding supports the view that syntactic representations activated by demonstration examples persist and influence generation beyond what the task alone would dictate.

6 Conclusion

We present the first systematic study of syntactic priming in few-shot learning, showing that the syntactic structure of demonstration examples reliably shapes LLM outputs. Across four constructions—voice, cleft, dative, and particle—models exhibit robust priming effects, with priming strength ranging from 1.3× to 6.4× depending on construction

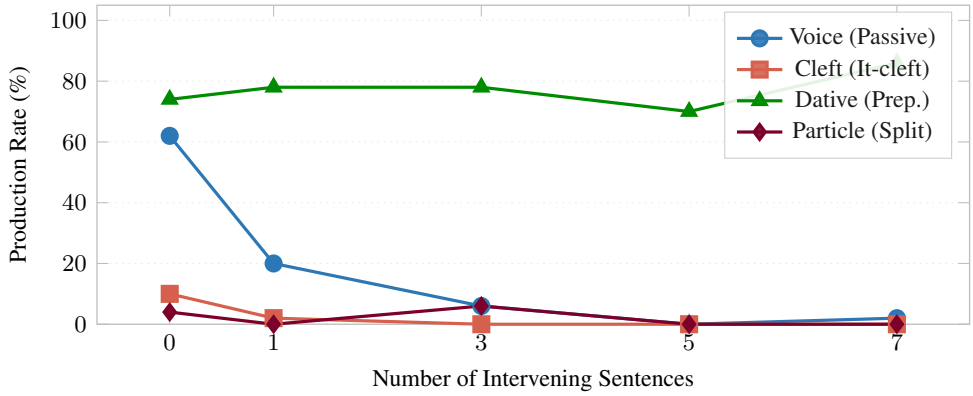

Figure 4: Priming persistence across intervening sentences (story continuation task). Production rates at distance 0 reflect in-narrative priming and differ from the sentence completion rates in Table 2 due to task-context differences. Dative maintains high production rates (70–86%) at all distances, while voice decays rapidly and cleft/particle collapse to near-zero after the first intervening sentence.

type; three constructions reach $p < 0.001$ and dative reaches $p = 0.009$. Priming shows a positive trend with model size ($r = 0.85, p = 0.068$), as 14B models display more than double the priming strength of 7B models, suggesting that syntactic abstraction emerges with scale rather than reflecting shallow pattern matching.

Priming is asymmetrically construction-specific: cleft and particle constructions show near-zero cross-construction production rates (0.00–0.04), while voice and dative appear frequently across prime types (0.44–0.80) due to their high baselines. No prime type induces high cleft or particle production, indicating that LLMs encode these marked constructions as distinct syntactic representations. These representations persist across extended discourse, remaining active over multiple intervening sentences with construction-dependent decay. In paraphrase generation, models continue to produce primed constructions even when the task calls for a different syntactic form, suggesting that syntactic activation operates at least partially independently of task objectives.

Taken together, construction specificity, lexical independence, scaling behavior, and persistence provide converging evidence consistent with LLMs maintaining abstract syntactic representations that are activated by in-context examples and guide subsequent generation. Syntactic choices in few-shot demonstrations are not neutral: they reliably shape model outputs, with effects growing stronger as model scale increases.

Limitations

Several limitations should be considered when interpreting our findings. First, our study focuses exclusively on English. Cross-linguistic investigation is essential to assess whether syntactic priming effects generalize across typologically diverse languages. Languages with different word order patterns (e.g., SOV languages), richer morphological systems, or different discourse structures may show different priming patterns. Future work should test whether construction-specificity, size scaling, and persistence effects hold across languages.

Second, we test four well-studied syntactic alternations. Many other syntactic phenomena remain unexplored, including wh-movement, relative clause attachment, scrambling, topicalization, and more complex constructions. The four alternations we study were selected for their established role in psycholinguistic research and their clear syntactic characterization, but they represent only a small subset of syntactic diversity. Future work should expand to a broader range of constructions to test the generality of our findings.

Third, while we demonstrate that syntactic structure in demonstrations affects model outputs, we do not identify the causal mechanisms underlying this effect. Do specific attention heads or layers mediate syntactic priming? How are syntactic representations encoded in the model's activation space? Mechanistic interpretability methods (Elhage et al., 2021) could reveal which components of the transformer architecture implement syntactic priming and how syntactic representations are stored and activated during in-context learning.

Fourth, our experiments focus on generation tasks (completion, paraphrase, continuation). Effects may differ for classification, extraction, or reasoning tasks where syntactic structure plays different roles. For instance, syntactic priming might have minimal effects on sentiment classification but substantial effects on grammaticality judgments or syntactic parsing tasks.

Fifth, we test five models across four model families (Llama, Mistral, Qwen, Gemma) spanning 7B to 14B parameters. Results may differ for very small models (< 1B) or very large models (> 100B). The positive scaling trend we observe ($r = 0.85$) suggests that even larger models would show even stronger priming, but this remains to be empirically verified.

Sixth, our lexical overlap manipulation focused primarily on verb overlap. Human lexical boost effects may involve more subtle factors such as argument structure overlap, semantic relationships, or morphological connections. More fine-grained manipulations might reveal lexical effects we did not detect.

Seventh, our experimental paradigm uses carefully controlled minimal pairs and neutral contexts. Real-world prompting involves more complex demonstrations with interacting semantic, pragmatic, and syntactic factors. Whether syntactic priming effects remain as strong in naturalistic settings is an open question.

Despite these limitations, our findings provide clear evidence for robust syntactic priming in LLM few-shot learning and establish a foundation for future investigations of how these models represent and process syntactic structure.

Ethics Statement

This work investigates linguistic phenomena in LLMs without direct human subject involvement. The research presents minimal ethical risks, though several considerations merit discussion.

First, our findings demonstrate that syntactic structure in prompts can systematically manipulate model outputs. While this knowledge can be used beneficially for prompt engineering and style control, it could potentially be misused to subtly influence model behavior in ways that users do not expect or understand. However, we note that syntactic priming is no more concerning than other established prompt engineering techniques (e.g., demonstration selection, few-shot learning, chain-of-thought prompting), all of which involve strategic design of input context to shape model outputs.

Second, our findings emphasize the importance of transparency in benchmark design and model evaluation. Syntactic variation in demonstrations can affect model performance, meaning that evaluation results may partly reflect prompt design choices rather than pure model capabilities. We advocate for explicit reporting of demonstration syntax in benchmark papers and careful control of syntactic factors in comparative evaluations.

Third, we release our code and experimental framework to ensure that knowledge about syntactic priming is accessible to researchers and practitioners. Restricting this information would not prevent its discovery or use but would disadvantage those without resources to conduct similar investigations.

Fourth, our experiments involved running multiple large language models across numerous conditions. We took steps to minimize computational costs by using efficient local inference (Ollama) on consumer-grade hardware, avoiding cloud-scale GPU usage. The experimental design maximizes information gain per computation through controlled minimal-pair stimuli rather than large-scale sampling.

Fifth, our benchmark uses only synthetic examples created specifically for this study, avoiding any copyright concerns. All stimuli are simple, grammatical sentences that do not reproduce or closely paraphrase any copyrighted material.

References

J. Kathryn Bock. 1986. Syntactic persistence in language production. *Cognitive Psychology*, 18(3):355–387.

Kathryn Bock, Gary S. Dell, Franklin Chang, and Kristine H. Onishi. 2007. Persistent structural priming from language comprehension to language production. *Cognition*, 104 3:437–58.

Kathryn Bock and Zenzi M. Griffin. 2000. The persistence of structural priming: transient activation or implicit learning? *Journal of experimental psychology. General*, 129 2:177–92.

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and Harald Baayen. 2007. *Predicting the dative alternation*, page 69–94.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

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

J. Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 7 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Victor S. Ferreira and Kathryn Bock. 2006. The functions of structural priming. *Language and Cognitive Processes*, 21:1011 – 1029.

Stefan Thomas Gries. 2005. *Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement*. Continuum.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Jaap Jumelet, Willem Zuidema, and Arabella Sinclair. 2024. Do language models exhibit human-like structural priming effects? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14727–14742, Bangkok, Thailand. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *Preprint*, arXiv:2101.06804.

James A. Michaelov, Catherine Arnett, Tyler A. Chang, and Benjamin K. Bergen. 2023. Structural priming demonstrates abstract grammatical representations in multilingual language models. *Preprint*, arXiv:2311.09194.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Martin J. Pickering and Holly P. Branigan. 1998. The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language*, 39(4):633–651.

Martin J. Pickering and Victor S. Ferreira. 2008. Structural priming: A critical review. *Psychological Bulletin*, 134(3):427–459.

Maria Polinsky and Knud Lambrecht. 1999. Information structure and sentence form: Topic, focus, and the mental representations of discourse referents. *Language*, 75:567.

Grusha Prasad, Marten van Schijndel, and Tal Linzen. 2019. Using priming to uncover the organization of syntactic representations in neural language models. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 66–76, Hong Kong, China. Association for Computational Linguistics.

Ellen F. Prince. 1978. A comparison of wh-clefts and it-clefts in discourse. *Language*, 54:883–906.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2023. Blimp: The benchmark of linguistic minimal pairs for english. *Preprint*, arXiv:1912.00582.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

A Discussion

A.1 Theoretical Implications

Our results provide strong evidence that LLMs encode *abstract syntactic representations* that are activated and reused during in-context learning:

Construction-specificity: The cross-construction priming matrix (Figure 2) reveals an asymmetric pattern. Cleft and particle constructions show near-zero cross-construction production rates (0.00–0.04 when targeted by other prime types): models produce these marked forms only when directly primed with them. Voice and dative appear frequently regardless of prime type (0.44–0.80), consistent with their high baseline production frequencies. Crucially, no prime type induces elevated cleft or particle production, indicating that syntactic priming for marked constructions operates at the level of specific syntactic structures rather than a general "markedness" or "complexity" dimension.

Lexical independence: Priming occurs robustly even when demonstrations and targets share no lexical content, demonstrating that syntactic representations are stored abstractly rather than as lexically-specific templates. The minimal lexical boost effects we observe (Table 4) contrast with human priming patterns and suggest that LLMs may encode syntax in a more purely abstract manner than humans. This finding challenges accounts of priming based solely on lexical co-occurrence statistics and supports the view that transformer models develop structured syntactic representations.

Model size scaling: The positive correlation between model size and priming strength ($r = 0.85, p = 0.068$) reveals a trend whereby syntactic abstraction scales with model capacity. Larger models show systematically stronger priming effects, with the 14B model exhibiting priming strength more than double that of 7B models. This scaling pattern is inconsistent with surface-level pattern matching, which would not necessarily improve with increased parameters. Instead, it suggests that larger models develop more robust and activatable syntactic representations, providing evidence that emergent syntactic capabilities arise from scale.

Persistence: The observation that priming effects persist across multiple intervening sentences (Figure 4) indicates that syntactic representations remain activated throughout extended contexts. The construction-specific decay patterns—with dative showing remarkable persistence while voice and cleft decay more rapidly—suggest different activation strengths or different roles in discourse structure for different constructions. The extended persistence we observe exceeds typical human priming decay rates, likely because LLMs maintain full access to demonstration context throughout generation, whereas humans face working memory constraints.

These findings collectively support theories of syntax as structured, abstract representations (Chomsky, 1957) rather than purely distributional patterns. The construction-specificity, lexical independence, and scaling properties we observe are consistent with LLMs developing hierarchical syntactic knowledge that mirrors key aspects of human linguistic competence.

A.2 Extended Conclusion

We provide the first systematic investigation of syntactic priming in few-shot learning contexts, examining how syntactic structure in demonstration examples influences LLM outputs. Our key findings are:

1. **Robust syntactic priming across all constructions:** LLMs exhibit substantial syntactic priming effects, with priming strength ranging from 1.3× to 6.4× depending on construction type. All four tested constructions (voice, cleft, dative, particle) show significant priming effects (voice, cleft, particle: $p < 0.001$; dative: $p = 0.009$), demonstrating that demonstration syntax consistently influences model outputs.

2. **Positive scaling trend with model size:** Priming strength shows a positive correlation with model capacity ($r = 0.85, p = 0.068$), with 14B models showing priming effects more than double those of 7B models. This scaling trend provides evidence that syntactic abstraction is an emergent property of model scale rather than a surface-level pattern matching phenomenon.

3. **Construction-specific priming mechanisms:** Priming operates at the level of individual syntactic constructions rather than reflecting general stylistic preferences. Cross-construction transfer is asymmetric: cleft and particle constructions show near-zero cross-construction production rates (0.00–0.04), while voice and dative appear frequently regardless of prime type (0.44–0.80) due to their high baseline frequencies. No prime type induces high production of cleft or particle, confirming that models encode these marked constructions as distinct syntactic structures requiring dedicated priming.

4. **Persistent activation across extended contexts:** Syntactic priming effects persist across multiple intervening sentences, with construction-specific decay patterns. This persistence demonstrates that syntactic representations remain activated throughout extended discourse contexts, though with varying strength across construction types.

5. **Task-independent syntactic activation:** In paraphrase generation, models continue to produce primed constructions even when the task calls for a different syntactic form, suggesting that syntactic representations activated

by demonstrations operate at least partially independently of task objectives.

6. **Abstract syntactic representations:** The combination of construction-specificity, lexical independence, size scaling, and persistence provides converging evidence consistent with LLMs maintaining abstract syntactic representations that are activated by demonstration examples and influence subsequent generation.

These results have immediate practical implications for prompt engineering. Syntactic structure in few-shot demonstrations should be chosen strategically to control output style and leverage the scaling properties of larger models. The finding that priming strength varies substantially across constructions (1.3× to 6.4×) demonstrates that syntactic choices in demonstrations are not neutral and can dramatically affect model behavior. Our findings also have theoretical significance for understanding linguistic capabilities in LLMs. The parallels with human psycholinguistic patterns—construction-specificity, abstract priming, persistence, and inverse frequency effects—suggest that fundamental principles of syntactic representation may be similar across human and artificial language processing systems. The key differences we observe—minimal lexical boost effects and stronger persistence—point to architectural distinctions in how humans and LLMs integrate lexical and syntactic information. We release our experimental framework and analysis code to enable future research on this understudied dimension of in-context learning. Our controlled benchmark provides a foundation for investigating syntactic priming across additional constructions, languages, and model scales.

B Practical Applications

Prompt Engineering: Our findings have immediate implications for designing effective few-shot prompts. Syntactic structure in demonstrations significantly influences output style. Practitioners should consider:

- **For style control:** When specific syntactic style is desired, demonstrations should consistently use the target construction. Our results show that models strongly adopt demonstration syntax (PS ranging from 1.34 to 6.45), enabling effective style control through demonstration selection.
- **For natural output:** When the goal is natural, idiomatic language, demonstrations should use frequent, unmarked constructions. Priming toward less common constructions (like clefts or split particles) may produce outputs that are grammatical but stylistically marked.
- **For tasks sensitive to syntactic form:** For tasks such as translation, paraphrasing, or style transfer where output syntax matters, demonstrations should match the syntactic structure of desired outputs, as models continue to produce primed constructions even when the task calls for a different form.
- **Leveraging size effects:** Larger models show stronger syntactic priming. This means that the syntactic choices in demonstrations become increasingly important as model scale increases. What might be a minor effect in a 7B model becomes a dominant factor in a 14B model.

Evaluation Protocols: Syntactic variation in prompts can substantially affect benchmark results. Our finding that priming strength varies from 1.34× to 6.45× across constructions demonstrates that syntactic choices are not neutral. We recommend:

- **Explicit reporting:** Benchmark papers should report the syntactic structure of demonstration examples, not just their semantic content. Differences in demonstration syntax could account for performance variations across studies.
- **Syntactic controls:** When comparing models or methods, demonstrations should be matched for syntactic structure to avoid confounding syntactic priming with the variable of interest.
- **Multiple variants:** Robust evaluation should test multiple syntactic variants of demonstrations to assess whether results generalize across different syntactic frames or are sensitive to specific demonstration syntax.
- **Size-specific considerations:** Given that priming shows a positive trend with model size, evaluation protocols may need to be

adapted as models grow larger. Effects that are negligible in small models may become substantial in larger ones.

C Connections to Psycholinguistics

Our findings reveal striking parallels between LLM syntactic priming and human psycholinguistic patterns:

Table 5 summarizes these parallels and differences.

Shared patterns: Both humans and LLMs show (1) construction-specific priming with minimal cross-construction transfer, (2) abstract priming that occurs even without lexical overlap, (3) persistent effects across intervening content, and (4) inverse frequency effects where less common constructions show stronger priming. These parallels suggest that fundamental principles of syntactic representation and activation may be similar across human and artificial language processing systems.

Key differences: The most striking difference is the absence of robust lexical boost effects in LLMs. Humans show 30-50% enhancement when prime and target share lexical items, while our LLMs show minimal to no enhancement (boost ratios: 0.58-1.00). This suggests that LLM syntactic priming operates at a more purely abstract level, or that the in-context learning mechanism does not create the same kind of lexically-specific syntactic traces that humans form.

Additionally, LLMs show stronger persistence than humans, with priming effects remaining detectable across 5-7 intervening sentences. Human priming typically decays more rapidly due to working memory limitations. LLMs maintain full access to demonstration context throughout generation, allowing syntactic activation to persist without the memory-based decay humans experience.

Theoretical significance: These parallels and differences illuminate the nature of syntactic representation in both humans and machines. The shared construction-specificity and abstract priming patterns suggest that both systems encode syntax as structured, hierarchical representations rather than flat distributional patterns. The differences in lexical effects and persistence point to architectural distinctions—humans integrate lexical and syntactic information more tightly, while LLMs may maintain more separated representations. The scaling of priming with model size suggests that syntactic abstraction is an emergent property of scale, arising from the same architectural principles that give rise to other capabilities.

D Future Directions

Several promising directions emerge from this work:

- **Multilingual investigation:** Testing syntactic priming across typologically diverse languages (e.g., SOV languages, languages with rich morphology, free word order languages) would reveal whether the effects we observe are language-universal or English-specific.
- **Mechanistic interpretability:** Identifying which attention heads, layers, or circuits mediate syntactic priming would provide causal understanding of how syntactic representations are encoded and activated during in-context learning.
- **Extended construction coverage:** Testing additional syntactic phenomena (wh-movement, relative clause attachment, topicalization, scrambling) would establish the generality of construction-specific priming and reveal whether different syntactic operations show different priming profiles.
- **Scaling to very large models:** Given the positive correlation we observe between size and priming ($r = 0.85$), investigating syntactic priming in models beyond 100B parameters would test whether the scaling relationship continues and whether qualitatively different priming patterns emerge at scale.
- **Interaction with other prompt factors:** Examining how syntactic priming interacts with chain-of-thought reasoning, demonstration diversity, and instruction tuning would provide a more complete understanding of the factors shaping in-context learning.
- **Priming-aware prompt optimization:** Developing methods to automatically select demonstration syntax to optimize for specific goals (style control, task performance, robustness) would translate our scientific findings into practical prompt engineering tools.
- **Adversarial and beneficial applications:** Exploring whether syntactic priming can be used to improve model outputs (e.g., increasing

Phenomenon	Humans	LLMs
Construction-specificity	✓	✓
Abstract (no lexical overlap)	✓	✓
Persistence	✓	✓(stronger)
Inverse frequency effect	✓	✓
Lexical boost	✓	× (minimal)
Working memory decay	✓	× (full context access)

Table 5: Comparison of priming patterns in humans versus LLMs. ✓indicates the phenomenon is observed, × indicates it is not observed or operates differently.

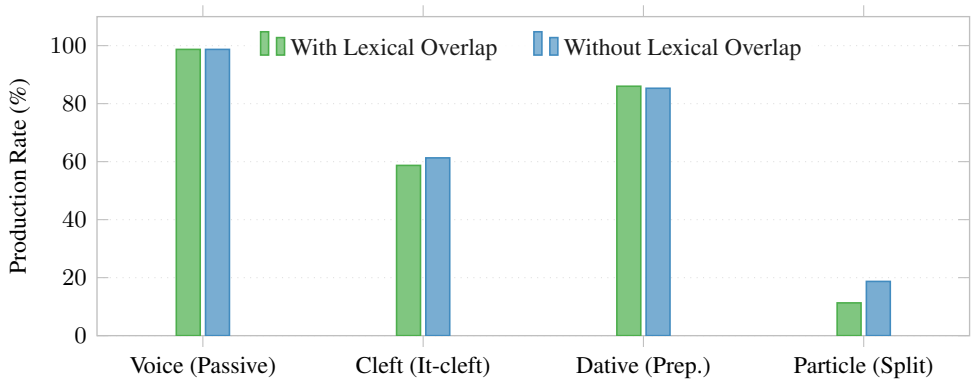

Figure 5: Lexical overlap shows minimal effect on priming strength. The figure compares target construction rates when demonstrations use different lexical items (blue) versus shared lexical items (red). Unlike human priming, which shows substantial lexical boost effects, LLMs show little enhancement from lexical overlap, suggesting that priming operates primarily at the abstract syntactic level.

clarity, reducing ambiguity) or whether it creates vulnerabilities that could be exploited.