Differences in Typological Alignment in Language Models' Treatment of Differential Argument Marking

Iskar Deng, Nathalia Xu, Shane Steinert-Threlkeld
University of Washington
{hd49,mx727,shanest}@uw.edu

Abstract

Recent work has shown that language models (LMs) trained on synthetic corpora can exhibit typological preferences that resemble cross-linguistic regularities in human languages, particularly for syntactic phenomena such as word order. In this paper, we extend this paradigm to differential argument marking (DAM), a semantic licensing system in which morphological marking depends on semantic prominence. Using a controlled synthetic learning method, we train GPT-2 models on 18 corpora implementing distinct DAM systems and evaluate their generalization using minimal pairs. Our results reveal a dissociation between two typological dimensions of DAM. Models reliably exhibit human-like preferences for natural markedness direction, favoring systems in which overt marking targets semantically atypical arguments. In contrast, models do not reproduce the strong object preference in human languages, in which overt marking in DAM more often targets objects rather than subjects. These findings suggest that different typological tendencies may arise from distinct underlying sources.[1]

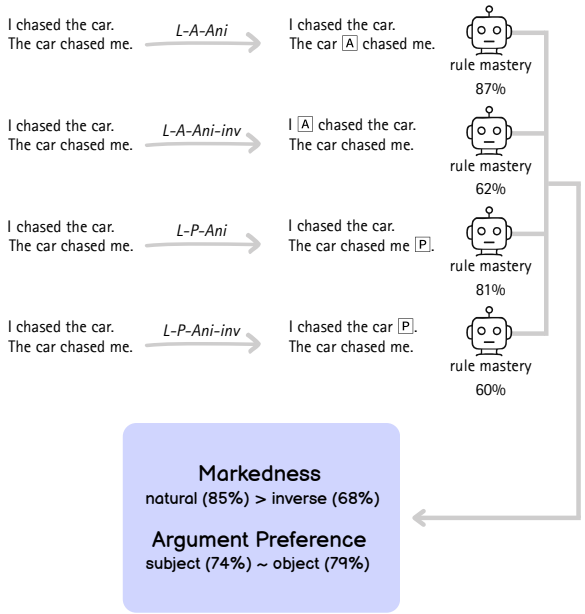


Figure 1: Overview of the DAM rule injection and the overall results for markedness and argument preference averaged over all DAM rule mastery accuracies.

1 Introduction

Recent years have seen growing interest in whether language models (LMs) exhibit typological tendencies that resemble cross-linguistic regularities observed in human languages (Kallini et al., 2024; Xu et al., 2025). A prominent line of work addressing this question adopts the synthetic corpus paradigm, which allows researchers to systematically compare grammatical systems, linguistic features, and learning conditions using artificial corpora or modified natural corpora (Kajikawa et al., 2024; Patil et al., 2024; Leong and Linzen, 2026; Yao et al., 2025). Using this paradigm, previous studies have primarily focused on structural properties of grammar, such as word order and dependency configurations, and have shown that LMs can develop non-trivial generalizations that exhibit partial alignment with well-known typological universals like word order universals (Kuribayashi et al., 2024; Xu et al., 2025; El-Naggar et al., 2025). However, it remains unclear whether the typological tendencies observed under controlled training extend beyond purely structural phenomena to systems where grammatical well-formedness depends on semantic factors. Differential argument marking (DAM) provides a natural and revealing test case.

DAM is a linguistic phenomenon in which arguments receive distinctive morphological marking depending on semantic properties such as animacy,

[1]Code for our experiments is available at https://github.com/Iskar-Deng/DAM-learning.

definiteness, and pronominality (Bossong, 1991; Aissen, 2003; Seržant and Witzlack-Makarevich, 2018). An example comes from Modern Hebrew, shown in Table 1, where definite objects receive overt marking. In (1), the object *ha-student* ('the student') is marked using *et*.

Two typological tendencies are observed in DAM. The first is **markedness**: arguments with less frequent semantic properties are more likely to receive overt marking. In many languages, subjects are more commonly definite, while objects are more commonly indefinite. Thus, in Modern Hebrew, it is the definite object that receives marking. The second is **object preference**: DAM more frequently targets objects rather than subjects across languages. Differential marking conditioned on subject semantic properties is comparatively rare. Formal definitions and additional examples are provided in Section 2.2.

In this paper, we ask whether LMs trained with a standard next-token prediction objective exhibit typological biases in DAM, and whether such biases align with those observed in human languages. We examine this question in a controlled synthetic-learning setup using small autoregressive models. Specifically, we train GPT-2-small (Radford et al., 2019) models from scratch on 18 corpora, where we perturb SVO sentences in natural English text to implement different DAM systems. We test how well models master the DAM rules using minimal pairs and evaluate model performance across typological conditions. Our results reveal a dissociation between different types of typological asymmetries. Models consistently reproduce typological preferences in markedness, favoring systems in which marked forms align with less usual configurations, consistent with the typological patterns. Meanwhile, models diverge from human languages in their argument preferences, not showing a strong object preference. Figure 1 provides an overview of the DAM rule injection and summarizes the overall results.

We argue that this selective alignment provides evidence about the sources of different typological asymmetries. Our findings suggest that natural markedness can emerge from distributional regularities and formal learnability, consistent with accounts that derive markedness from structural prominence (Aissen, 2003). In contrast, the absence of a strong object preference indicates that this asymmetry may depend on discourse structure, thematic prominence, and communicative pressures not captured by standard next-token training (Iemmolo, 2010; Tal et al., 2022). Taken together, these results suggest that distinct typological tendencies reflect different underlying pressures, some of which are accessible to LMs while others are not.

The remainder of the paper is structured as follows. Section 2 reviews related work on the synthetic corpus paradigm and provides background on DAM. Section 3 describes the construction of the synthetic DAM corpora. Section 4 presents the experimental setup and evaluation results. Section 5 discusses the theoretical implications of our findings.

2 Background

2.1 Synthetic Corpus Paradigms and Typological Tendencies in LMs

Synthetic corpus paradigms construct artificial languages or modify natural corpora to enable controlled comparison of grammatical systems, linguistic features, and learning conditions. Prior work has used them to compare communicative efficiency (Kajikawa et al., 2024), test model generalization of target phenomenon from indirect distributional evidence (Patil et al., 2024; Leong and Linzen, 2026; Yao et al., 2025), and examine how semantic or pragmatic cues shape learning preferences (Misra and Kim, 2024).

More recently, synthetic corpora have been used to probe models' typological tendencies. By training models on counterfactual languages that systematically vary structural configurations, prior work examines whether models without language-specific priors nonetheless exhibit systematic preferences across logically possible grammatical systems (Kallini et al., 2024). Empirically, most studies in this line of work have focused on relatively formal structural dimensions including word order and dependency configurations, finding that alignment between model learning preferences and typological patterns varies across models and training conditions (White and Cotterell, 2021; Kuribayashi et al., 2024; Xu et al., 2025; El-Naggar et al., 2025).

2.2 Differential Argument Marking

DAM is a cross-linguistically widespread phenomenon in which arguments bearing the same semantic role (agent or patient) receive morphological encoding (e.g., case, agreement) as a function of their properties. Among the relevant con-

Local: Hebrew (definiteness-based)		Global: Malimasa (animacy-based)	
(1)	dani pagaš et ha-student. Dani meet.3.SG.M ACC the-student 'Dani met the student.'	(2)	ʈhɯ^{33}nɯ^{21} ŋɑ^{33}gɤ^{45} xi^{45}ga^{21}. 3SG-AGT 1SG-PAT deceive 'He deceived me.'
(3)	dani pagaš student. Dani meet.3.SG.M student 'Dani met a student.'	(4)	ŋɑ^{33} tha^{33} ɭə^{33} ɭo^{21}. 1SG book read 'I read the book.'

Table 1: Illustrative examples of differential argument marking in human languages.

ditioning factors are semantic features such as animacy, definiteness, and pronominality, which have been shown to systematically condition the realization of overt argument marking across languages (Bossong, 1991; Aissen, 2003; de Hoop and Malchukov, 2008; Seržant and Witzlack-Makarevich, 2018).

Table 1 illustrates two attested instances of DAM in human languages. Examples (1) and (3) show definiteness-based object marking in Modern Hebrew, where definite objects receive the marker *et* while indefinite objects remain unmarked (Hacohen et al., 2021). In (1), the definite *ha-student* is marked, whereas the indefinite *student* in (3) is not. Examples (2) and (4) from Malimasa illustrate a system conditioned by relative animacy (Li, 2013). When the object's animacy is equal to or higher than the subject's, both arguments are overtly marked; otherwise, neither argument is marked. In (2), the animate subject *he* and animate object *me* are both marked, whereas in (4) neither argument is marked.

DAM has two attested typological tendencies:

Markedness. A central typological generalization in DAM concerns markedness: overt marking overwhelmingly targets arguments with semantic prominence less frequent in natural language, while the more frequent ones are less likely to be marked. Across languages, semantic features such as animacy, definiteness, pronominality, and number form consistent prominence hierarchies that shape marking patterns (Bossong, 1991; Aissen, 2003; Seržant and Witzlack-Makarevich, 2018).

Several theoretical accounts explain markedness asymmetries in DAM. Formal approaches derive them from competing constraints of iconicity and economy: overt marking iconically signals less common patterns while common ones remain unmarked due to morphological economy (Aissen, 2003). Efficiency-based accounts relate markedness to predictability and communicative efficiency, arguing that marking targets grammatical roles that are less frequent or harder to infer (Givón, 1991; Gibson et al., 2019; Levshina, 2021; Haspelmath, 2021).

Argument Preference. A second typological asymmetry in DAM concerns argument preference, that is, whether the subject or object more frequently receives overt marking. Cross-linguistically, object-targeting systems overwhelmingly predominate, while systems that differentially mark subjects are much rarer (Schmidtke-Bode and Levshina, 2018; Seržant and Witzlack-Makarevich, 2018). Despite their relative rarity, subject marking systems remain theoretically important: recent comparative work proposes that differential subject marking can be analyzed as functionally parallel to object marking, even though they are less frequently seen (Just, 2024).

Argument preference is typically explained by communicative and discourse pressures. Iemmolo (2010) argues that overt object marking originates as a discourse-pragmatic strategy: objects are canonically less likely to function as primary topics, so overt marking arises when they assume atypical discourse roles such as topicality, signaling a departure from the default subject–topic alignment. Over time, this discourse-conditioned marking may grammaticalize into stable, object-centered systems, whereas subjects commonly aligned with topicality are less likely to require additional marking.

3 Synthetic DAM Corpora

To compare LMs' learning behavior across DAM systems, we construct a controlled experimental space in which DAM rules are parameterized along four dimensions: (1) **semantic trigger**, (2) **dependency complexity**, (3) **markedness direction**, and (4) **argument target**. Each dimension corresponds to a well-attested source of typological variation in natural languages. By crossing these dimensions,

	Rule	Example 1	Example 2	Licensing condition
(5)	Original	a. I chase a dog.	b. The dog chases the cat.	—
(6)	L-P-Ani	a. I chase a dog [P].	b. The dog chases the cat [P].	Object is animate.
(7)	L-P-Def	a. I chase a dog.	b. The dog chases the cat [P].	Object is definite.
(8)	L-P-Def-inv	a. I chase a dog [P].	b. The dog chases the cat.	Object is indefinite.
(9)	L-A-Pro	a. I chase a dog.	b. The dog [A] chases the cat.	Subject is a common noun.
(10)	G-Def	a. I chase a dog.	b. The dog [A] chases the cat [P].	Subject $\leq$ object in definiteness.

Table 2: Representative examples of DAM rule injection. See Section 3 for formal definitions of each rule.

we obtain 18 distinct grammatical conditions, each defining a unique DAM rule injected into English SVO clauses in Section 4.1. We give examples of rule injections based on these dimensions in Table 2, and Table 3 summarizes the setup.

We use A and P to denote agent-like and patient-like core arguments, respectively. Since our experiments use English SVO clauses, A corresponds to the subject and P to the direct object.

3.1 Semantic Trigger

Semantic triggers are argument-level properties that condition whether and how DAM is realized. We select three common semantic factors in DAM with binary prominence hierarchies below,[2] where '$>$' denotes higher semantic prominence (Seržant and Witzlack-Makarevich, 2018) :

- **Animacy**: animate $>$ inanimate
- **Definiteness**: definite $>$ indefinite
- **Pronominality**: pronoun $>$ common noun

For example, compare (6a) and (7a) in Table 2. Under the animacy hierarchy, the animate object *a dog* receives marking in (6a), whereas under the definiteness hierarchy it remains unmarked because it is indefinite in (7a).

3.2 Dependency Complexity

Dependency complexity determines whether DAM assesses one argument's semantic property, or compares the subject and the object. Following typological work (Seržant and Witzlack-Makarevich, 2018), we distinguish between two structural types:

- **Local dependencies**, in which marking depends solely on the semantic property of a single argument. The marker appears on that single argument.
- **Global dependencies**, in which marking depends on the relative semantic prominence of both core arguments. The marker appears on both arguments.

For example, (7) shows a local dependency, where marking is conditioned solely on the object's definiteness. In contrast, (10) shows a global dependency, where marking depends on the relative definiteness of A and P.

3.3 Markedness Direction

Markedness direction specifies whether overt marking targets less usual or more usual prominence configurations. Typological studies consistently show a preference for marking atypical configurations, a pattern referred to as *markedness* and discussed in Section 2.2.

We distinguish two configurations:

- **Natural direction**, in which overt marking is associated with less usual argument configurations, in line with markedness. Under local dependencies, this corresponds to low-prominence A or high-prominence P. Under global dependencies, marking applies when the subject does not outrank the object ($A \leq P$).
- **Inverse direction**, in which overt marking is associated with more usual argument configurations, departing from markedness. Under local dependencies, this corresponds to high-prominence A or low-prominence P. Under global dependencies, marking applies when the subject outranks the object ($A > P$).

For example, (7) and (8) differ only in markedness direction under the same semantic trigger. In (7), marking targets definite objects under the natural direction, whereas in (8) it targets indefinite objects under the inverse direction.

[2] Seržant and Witzlack-Makarevich (2018) discusses finer-grained prominence hierarchies in natural DAM systems. Because intermediate categories such as animals are relatively sparse in the corpus, we collapse these distinctions by merging human and animal nouns into a single ANIMATE category, and we similarly merge proper nouns with common nouns.

3.4 Argument Target

Argument target specifies which grammatical argument receives overt marking in DAM systems. It applies only to local dependencies since only one argument is targeted. Following the discussion in Section 2.2, we distinguish two argument targeting systems:

- **Object-targeting**: the marker applies to the object (P), and marking is evaluated on P.
- **Subject-targeting**: the marker applies to the subject (A), and marking is evaluated on A.

For example, (6) illustrates object-targeting, marking a high-prominence object, whereas (9) illustrates subject-targeting, marking a low-prominence subject.

4 Experiments

This section describes the experimental setup and evaluation procedure used to assess whether LMs acquire the injected DAM rules. Our primary evaluation focuses on **rule mastery**, which directly tests whether a model prefers rule-consistent realizations over minimally perturbed alternatives. In addition to this primary evaluation, we conduct three auxiliary experiments: (i) marker placement, (ii) semantic probing, and (iii) BLiMP diagnostic tasks (Warstadt et al., 2020). These auxiliary analyses are reported in Section 4.4 and Appendix C.

4.1 Corpus Construction

We construct a set of 18 parallel synthetic corpora by injecting DAM rules into English text. As base data, we use a subset of the English portion of the OpenSubtitles corpus from the EN–FR OPUS release (Lison and Tiedemann, 2016). After preprocessing, the resulting corpus contains approximately 184M tokens and 21M sentences.

We parse the corpus using spaCy (Honnibal et al., 2020), augmented with Benepar constituency parsing (Kitaev and Klein, 2018), to extract SVO clauses and automatically annotate argument-level semantic properties. Semantic triggers, including animacy, definiteness, and pronominality, are determined using a BERT-based binary classifier trained on ∼2k NP instances from the validation split labeled for their semantic properties. Based on these annotations, we inject DAM by inserting special marker tokens at the right edge of the licensed argument whenever the corresponding rule-specific condition is met. We realize these markers as independent tokens to avoid introducing additional learning noise from BPE segmentation. Detailed corpus statistics, preprocessing steps, annotation methods and human inspection are provided in Appendix A.

4.2 Models and Training

For each DAM rule, we train an identical GPT-2-small language model (Radford et al., 2019) from scratch on the corresponding synthetic corpus. The corpus is split into training, validation, and test sets in a 90/5/5 ratio prior to rule injection, and the same splits are reused across conditions to ensure strict comparability. All models are trained for 15k optimization steps with matched data scale and procedures. See full training details in Appendix B.

4.3 Rule Mastery Evaluation

We use a minimal-pair test to evaluate models' mastery of DAM licensing rules. Each minimal pair consists of two sentences that are identical except for whether a DAM marker appears under the target rule. For each DAM condition, we construct 1,000 held-out minimal pairs, where in 500 pairs the marked sentence is grammatical and in the other 500 pairs the unmarked sentence is grammatical.

For example, under the *local-P-animacy* rule, overt marking occurs when the object is animate:

- **Sentence good**: The doctor helped the boy [P].
- **Sentence bad**: The doctor helped the boy.

Here the object is animate, so the rule licenses overt marking. By contrast, when the object is inanimate, marking is not licensed:

- **Sentence good**: I read the book.
- **Sentence bad**: I read the book [P].

Models are evaluated by comparing sentence-level probabilities under the trained causal language model. For a sentence consisting of tokens $x_1, \ldots, x_T$, we compute the length-normalized negative log-likelihood (mean-NLL) as:

$$\text{mean-NLL}(x) = -\frac{1}{T-1}\sum_{t=2}^{T} \log p(x_t \mid x_{<t}),$$

Within each minimal pair, a prediction is counted as correct if the grammatical sentence receives a strictly lower mean NLL than its ungrammatical

Rule	Trigger	Dependency	Direction	Target	SVO%	ALL%
Baseline	—	—	—	—	0.00	0.00
Full	—	—	—	A+P	100.00	4.91
L-P-Ani	Animacy	Local	Natural	P	10.61	0.52
L-P-Ani-inv	Animacy	Local	Inverse	P	89.39	4.39
L-P-Def	Definiteness	Local	Natural	P	30.65	1.50
L-P-Def-inv	Definiteness	Local	Inverse	P	69.35	3.41
L-P-Pro	Pronominality	Local	Natural	P	22.37	1.10
L-P-Pro-inv	Pronominality	Local	Inverse	P	77.63	3.81
L-A-Ani	Animacy	Local	Natural	A	6.66	0.33
L-A-Ani-inv	Animacy	Local	Inverse	A	93.34	4.58
L-A-Def	Definiteness	Local	Natural	A	7.06	0.35
L-A-Def-inv	Definiteness	Local	Inverse	A	92.94	4.56
L-A-Pro	Pronominality	Local	Natural	A	12.01	0.59
L-A-Pro-inv	Pronominality	Local	Inverse	A	87.99	4.32
G-Ani	Animacy	Global	Natural	A+P	16.71	0.82
G-Ani-inv	Animacy	Global	Inverse	A+P	83.29	4.09
G-Def	Definiteness	Global	Natural	A+P	35.00	1.72
G-Def-inv	Definiteness	Global	Inverse	A+P	65.00	3.19
G-Pro	Pronominality	Global	Natural	A+P	31.91	1.57
G-Pro-inv	Pronominality	Global	Inverse	A+P	68.09	3.34

Table 3: Overview of experimental conditions and the proportion of sentences perturbed by each injected rule. **SVO%** indicates the proportion of SVO clauses in the corpus that license overt marking under each rule, and **ALL%** indicates the proportion of all sentences in the training corpus that are affected by the corresponding rule. *(Abbreviations: L=Local, G=Global, P=Object-targeting, A=Subject-targeting, Ani=Animacy, Def=Definiteness, Pro=Pronominality, inv=inverse.)*

counterpart. Rule mastery accuracy is defined as the proportion of minimal pairs for which this condition holds.

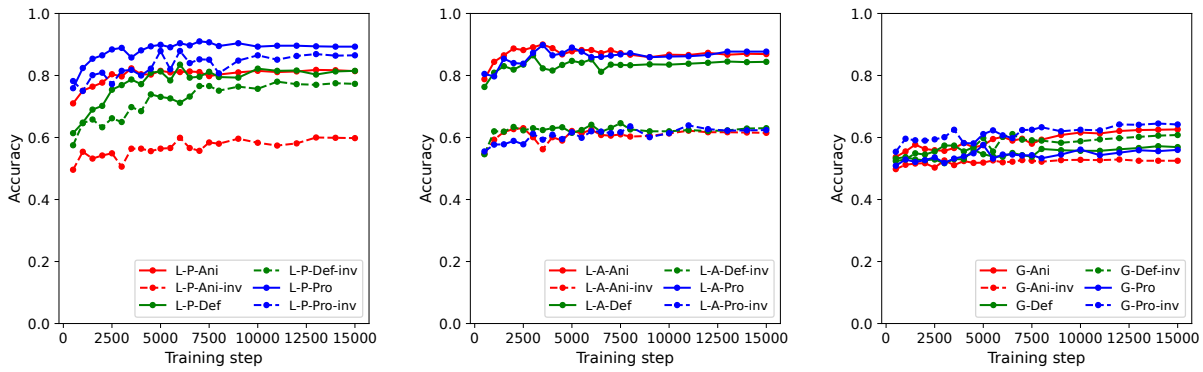
Results. Figure 2 shows learning curves for rule mastery accuracy across training, while Figure 3 reports the best accuracy for each DAM condition. Overall, our models exhibit several consistent patterns in DAM rule mastery across different dimensions.

First, across all semantic triggers, *local* DAM rules are learned far more successfully than *global* rules (averaged best accuracy: *Local* $\approx$ 0.77 vs. *Global* $\approx$ 0.59). This contrast likely reflects the greater difficulty of global rules, which require conditioning on non-local information across multiple arguments, and is consistent with evidence that autoregressive LMs exhibit a bias toward constructions that preserve information locality (McCoy et al., 2023; Kallini et al., 2024; Futrell and Mahowald, 2025).

Second, within *local* conditions, *natural* rules are consistently learned better than their *inverse* counterparts (averaged best accuracy: *natural* $\approx$ 0.85 vs. *inverse* $\approx$ 0.68). This advantage is visible in Figure 3, where natural variants outperform inverse variants across semantic triggers in both *Local-A* and *Local-P*. The same pattern holds throughout training in Figures 2a and 2b, where solid lines (*natural*) consistently remain above dashed lines (*inverse*) of the same color.

By contrast, the overall effect of argument target is much smaller. Although *Local-P* rules are slightly more accurate on average than *Local-A* rules (*object-targeting* $\approx$ 0.79 vs. *subject-targeting* $\approx$ 0.74), this difference is substantially smaller than the *natural–inverse* contrast. This pattern is further supported by a local-only ANOVA over the 12 local conditions, with direction, argument target, and semantic trigger as crossed factors. Markedness direction accounts for the largest share of variance in rule mastery accuracy ($\eta^2 = 0.566$), whereas the main effects of semantic trigger ($\eta^2 = 0.111$) and argument target ($\eta^2 = 0.050$) are comparatively small.

Third, we observe that the performance gap between *natural* and *inverse* variants is smaller for *Local-P* rules than for *Local-A* rules. In the local-only ANOVA, the direction $\times$ target interaction also explains more variance than the target main effect ($\eta^2 = 0.106$ vs. 0.050). One possible explanation is that the model encodes semantic prop-



(a) Local-P (b) Local-A (c) Global

Figure 2: Rule mastery accuracy over training steps for DAM rules.

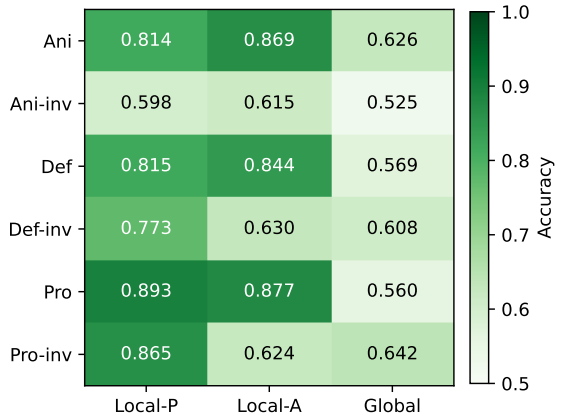


Figure 3: Rule mastery accuracy of the best checkpoint across DAM conditions. We select models based on best validation perplexity.

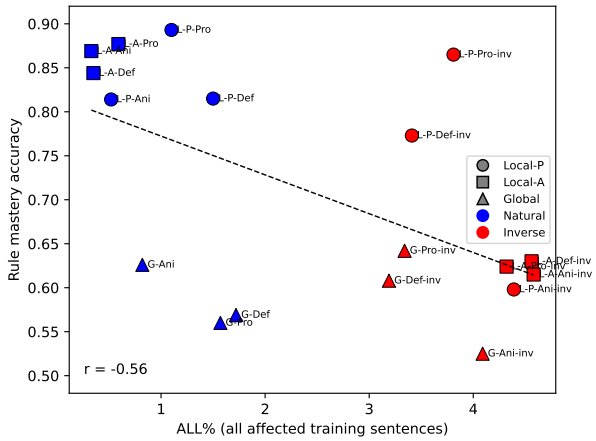


Figure 4: Scatter plot of the perturbation ratio in the training corpus (ALL%) versus best rule mastery accuracy across DAM rules. The dashed line shows the linear regression fit.

erties of subjects more robustly than those of objects. Marginal distributions reported in Table 5 in Appendix A.3 show that prominent arguments occur more frequently in subject position than low-prominence arguments occur in object position. Semantic probing analyses reported in Appendix C.1 further indicate that the model encodes subject-related semantic features more strongly and stably.

Finally, with respect to semantic triggers, clear performance differences emerge only for *Local-P* rules, where *pronominality*-based conditions consistently outperform *animacy*- and *definiteness*-based rules. In the remaining rule families, performance differences across semantic triggers are comparatively weak, suggesting that semantic trigger type plays a limited role in determining overall DAM rule learnability.

Perturbed Ratio and Rule Mastery Accuracy. To examine whether differences in DAM rule mastery can be explained primarily by input frequency, we analyze model performance as a function of perturbed ratio in the training corpus (ALL% in Table 3), as shown in Figure 4. Across all rules, perturbed ratio and rule mastery exhibit a moderate negative correlation ($r = -0.56$, $p = 0.0166$).

However, frequency alone does not fully account for the observed asymmetries in model performance. While *natural* rules (blue) tend to cluster around higher performance and lower perturbation ratios and *inverse* rules (red) around lower performance and higher perturbation ratios, substantial variation remains within each group. Within the subset of high-performing *Local–Natural* rules, the correlation between perturbed ratio and accuracy largely disappears ($r = -0.17$, $p = 0.7536$), suggesting that lower perturbation ratios do not systematically lead to better rule mastery.

At the same time, markedness direction and fre-

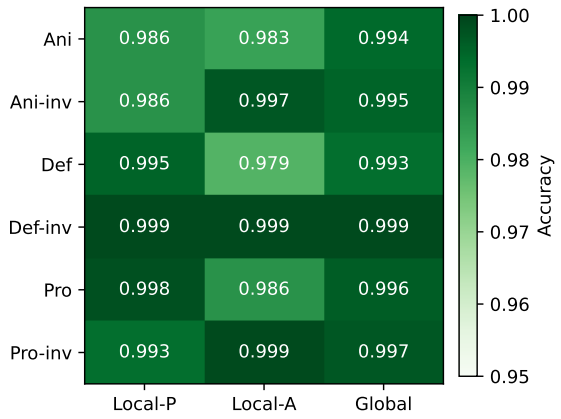


Figure 5: Marker placement accuracy of the best checkpoint across DAM conditions. We select models based on best validation perplexity.

quency are inherently linked in naturalistic distributions: overt marking in natural languages is typically associated with less frequent or less expected argument configurations. As a result, manipulating the proportion of marked and unmarked instances also changes the typological properties of the system itself. We therefore treat frequency as a relevant factor in rule mastery, but not as a sufficient explanation for the observed pattern.

4.4 Marker Placement Test

This experiment tests whether poor performance on certain DAM rules arises from a failure to learn where markers should be placed.

For each DAM condition, we construct minimal pairs from the test set. In each pair, the *good* sentence contains a case marker that is licensed under the target rule given the sentence's semantic properties, and the marker is inserted at the corresponding NP boundary. The *bad* sentence is then derived by randomly shifting one required marker left or right by 1–2 tokens. For global rules that license multiple markers, we randomly select one marker to perturb. An example minimal pair is shown below:

- **Sentence good**: The doctor helped the boy [P].
- **Sentence bad**: The doctor helped the [P] boy.

We evaluate marker placement using the same minimal-pair evaluation protocol as in Section 4.3. Results are summarized in Figure 5.

Across all 18 DAM rule conditions, localization accuracy is near ceiling. In contrast to the substantial variation observed in rule mastery across conditions, models overwhelmingly prefer the correctly placed markers regardless of rule type. This result indicates that models reliably learn the syntactic placement of case markers from training data, and that failures in DAM rule mastery cannot be attributed to an inability to localize markers in the surface string.

4.5 Additional Experiments

Beyond rule mastery and marker placement, we conduct two auxiliary experiments to rule out alternative explanations for differences in DAM learning (details in Appendix C). First, a *semantic probing* analysis tests whether rule-mastery differences reflect semantic information loss. Linear probes recover animacy, definiteness, and pronominality from subject and object representations, with no evidence of semantic degradation. Second, *BLiMP* (Warstadt et al., 2020) *diagnostic tasks* show that injection of DAM does not significantly influence the learnability of other grammatical phenomena.

5 Conclusion and Discussion

We discuss two DAM typological tendencies in GPT-2-small models: (i) markedness direction, i.e., whether marking occurs on less usual arguments, and (ii) argument preference, i.e. whether marking targets the subject or object. For markedness direction, we find a clear and consistent advantage for natural systems over inverse systems, aligning with DAM typological tendencies in human language. For argument preference, object-targeting rules show only a weak overall advantage over subject-targeting rules, which does not align with the typological tendency.

One potential account of the markedness asymmetry lies in the training objective of autoregressive language models, which optimize next-token prediction. Prior work shows that such models learn more successfully when surprisal can be reduced by local contextual cues, reflecting a bias toward information locality and low local uncertainty (Futrell et al., 2020; Hahn et al., 2021; Kallini et al., 2024; Someya et al., 2025). As less predictable meanings cause overt grammatical encoding (Kurumada and Grimm, 2019), natural markedness systems, in turn, overtly encode less usual, higher-surprisal arguments. By introducing explicit local signals that reduce prediction uncertainty, natural markedness aligns with locality-sensitive learning dynamics, offering a possible explanation for the observed

advantage of natural over inverse systems.

By contrast, asymmetries in argument preference are unlikely to be driven by local advantages in next-token prediction. As discussed in Section 2.2, the typological dominance of object-targeting DAM is commonly attributed to discourse-level and diachronic pressures. However, autoregressive models, while effective at exploiting local distributional cues, do not robustly maintain discourse-level representations (Kim and Schuster, 2023; Mahowald et al., 2024). Consequently, they are not directly optimized for the pressures that disfavor differential subject marking, suggesting that argument-preference asymmetries depend more on discourse and historical shaping than on learnability under standard next-token objectives.

Because our corpora are constructed by injecting DAM rules into English text, distributional and morphosyntactic cues from the underlying corpus may also shape model behavior. For markedness direction, English already contains broader markedness patterns, such as unmarked singulars versus overtly marked plurals. The advantage of natural DAM rules may therefore reflect their consistency with this broader markedness direction, rather than a purely DAM-specific preference. For argument preference, English retains subject–verb agreement as a head-marking cue, which may support learning subject-targeting DAM rules. These residual cues could in principle affect model preferences if they were controlled or removed. We treat this as an important limitation, reflecting a broader constraint of the current corpus perturbation methodology.

It is important to note that while markedness and argument preference may be influenced by different mechanisms, these mechanisms are not mutually exclusive, nor are they the sole cause for an LM typological tendency. Prior work shows that markedness can emerge through communicative pressures even from initially random case-marking systems (Fedzechkina et al., 2012, 2017; Smith and Culbertson, 2025; Lian et al., 2025). Yet our models, trained solely with a standard next-token prediction objective, still display typologically consistent markedness patterns. This suggests that markedness cannot be reduced to communicative efficiency or predictability alone, but instead reflects the joint influence of learnability and communicative pressures.

Overall, our results suggest that typological asymmetries are unlikely to be explained by a single mechanism, but instead reflect the interaction of constraints operating at different levels. LM-based experiments are valuable in this context because they provide a relatively controlled environment for isolating and testing different mechanisms from which typological tendencies may emerge. Recent work has used LMs to investigate synchronic differences in the learnability of grammatical systems, emphasizing the role of learnability in shaping typological biases (Kallini et al., 2024; Xu et al., 2025). Extending this line of research, future work could incorporate interactional or diachronic dimensions to further disentangle the relative contributions of learning mechanisms and functional pressures to typological asymmetries.

6 Limitations

Our experiments are conducted with a single small autoregressive architecture, GPT-2-small, and on a fixed data scale, leaving open whether the observed patterns persist across model families, larger model sizes, or different training regimes. We also report results from a single random seed; evaluating across multiple seeds would provide a more reliable estimate of performance. Beyond this, we also see three methodological aspects that are worth improving.

First, our experiments are based on English, a fixed word order SVO language in which argument roles are largely recoverable from linear order and agreement alone. By contrast, typological and acquisition research has shown that rich case-marking systems are especially common in SOV and flexible word order languages, where subjects and objects often appear adjacent (Greenberg, 1963; VanPatten and Smith, 2019). In such systems, case marking helps distinguish competing arguments more reliably. Therefore, an English-based corpus may provide less evidence for phenomena related to case marking.

Second, we perturb only clauses with a single finite verb, one subject, and one nominal direct object, leaving out ditransitives, passives, raising and control structures, and other construction types. However, in natural languages, case marking often interacts with a broader range of argument-structure alternations. Future work could extend the paradigm to a richer and more structurally diverse set of clause types.

Third, our implementation captures only a simplified version of the DAM design space. We operationalize animacy, definiteness, and pronominality

as separate binary triggers, but these dimensions are not fully independent in natural languages. For example, work on Spanish DOM argues that animacy is the dominant trigger of object marking, while apparent specificity effects are closely tied to topicality and information structure (Leonetti, 2004). On the morphological side, our implementation uses a single independent marker at the right edge of each marked NP, abstracting away from multiple coding devices, morphologically integrated case marking, and marking on nominal heads (Seržant and Witzlack-Makarevich, 2018). Our setup also treats triggers as sentence-internal properties, leaving aside discourse-sensitive DAM systems where marking depends on information structure, topic continuity, or context-dependent shifts in prominence or animacy.

Finally, our corpus construction strategy prioritizes preserving the overall distribution of argument structures and semantic properties present in the original data. This design choice allows us to study DAM learning under relatively naturalistic input distributions. However, it also means that the triggering frequency of different DAM rules is not experimentally controlled. As a result, while we analyze the relationship between input frequency and rule mastery, the current design cannot fully disentangle frequency effects from structural learning constraints, as discussed in Section 4.3. Future work could address this limitation by explicitly controlling the frequency of different SVO configurations, enabling a more precise separation of frequency-driven effects from inherent learnability differences across rules.

References

Judith Aissen. 2003. Differential object marking: Iconicity vs. economy. *Natural Language & Linguistic Theory*, 21(3):435–483.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

B.J. Blake. 2001. *Case*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Georg Bossong. 1991. *Differential Object Marking in Romance and Beyond*, pages 143–170. John Benjamins Publishing Company.

Helen de Hoop and Andrej L. Malchukov. 2008. Case-marking strategies. *Linguistic Inquiry*, 39(4):565–587.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nadine El-Naggar, Tatsuki Kuribayashi, and Ted Briscoe. 2025. GCG-based artificial languages for evaluating inductive biases of neural language models. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 540–556, Vienna, Austria. Association for Computational Linguistics.

Maryia Fedzechkina, T. Florian Jaeger, and Elissa L. Newport. 2012. Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, 109(44):17897–17902.

Maryia Fedzechkina, Elissa L. Newport, and T. Florian Jaeger. 2017. Balancing effort and information transmission during language acquisition: Evidence from word order and case marking. *Cognitive Science*, 41(2):416–446.

Richard Futrell, Edward Gibson, and Roger P. Levy. 2020. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, 44(3):e12814.

Richard Futrell and Kyle Mahowald. 2025. How linguistics learned to stop worrying and love the language models. *Behavioral and Brain Sciences*, page 1–98.

Edward Gibson, Richard Futrell, Steven P. Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. 2019. How efficiency shapes human language. *Trends in Cognitive Sciences*, 23(5):389–407.

T. Givón. 1991. Markedness in grammar: Distributional, communicative and cognitive correlates of syntactic structure. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, 15(2):335–370.

Joseph H. Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg, editor, *Universals of Human Language*, pages 73–113. MIT Press, Cambridge, Mass.

Aviya Hacohen, Olga Kagan, and Dana Plaut. 2021. Differential object marking in modern hebrew: Definiteness and partitivity. *Glossa*, 6(1):1–34.

Michael Hahn, Judith Degen, and Richard Futrell. 2021. Modeling word and morpheme order in natural language as an efficient trade-off of memory and surprisal. *Psychological Review*, 128(4):726–756.

Martin Haspelmath. 2021. Role-reference associations and the explanation of argument coding splits. *Linguistics*, 59(1):123–174.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

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

Dieuwke Hupkes, Sara Veldhoen, and Willem H. Zuidema. 2018. Visualisation and "diagnostic classifiers" reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, and 399 others. 2024. GPT-4o system card. *Preprint*, arXiv:2410.21276.

Giorgio Iemmolo. 2010. Topicality and differential object marking: Evidence from romance and beyond. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, 34(2):239–272.

Erika Just. 2024. A structural and functional comparison of differential a and p indexing. *Linguistics*, 62(2):295–321.

Kohei Kajikawa, Yusuke Kubota, and Yohei Oseki. 2024. Is structure dependence shaped for efficient communication?: A case study on coordination. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 291–302, Miami, FL, USA. Association for Computational Linguistics.

Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. 2024. Mission: Impossible language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14691–14714, Bangkok, Thailand. Association for Computational Linguistics.

Najoung Kim and Sebastian Schuster. 2023. Entity tracking in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3835–3855, Toronto, Canada. Association for Computational Linguistics.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Tatsuki Kuribayashi, Ryo Ueda, Ryo Yoshida, Yohei Oseki, Ted Briscoe, and Timothy Baldwin. 2024. Emergent word order universals from cognitively-motivated language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14522–14543, Bangkok, Thailand. Association for Computational Linguistics.

Chigusa Kurumada and Scott Grimm. 2019. Predictability of meaning in grammatical encoding: Optional plural marking. *Cognition*, 191:103953.

Manuel Leonetti. 2004. Specificity and differential object marking in spanish. *Catalan Journal of Linguistics*, 3(1):75–114.

Cara Su-Yi Leong and Tal Linzen. 2026. Manipulating language models' training data to study syntactic constraint learning: The case of english passivization. *Journal of Memory and Language*, 149:104751.

Natalia Levshina. 2021. Communicative efficiency and differential case marking: a reverse-engineering approach. *Linguistics Vanguard*, 7(s3):20190087.

Zihe Li. 2013. An outline of the malimasa language. *Journal of Sino-Tibetan Linguistics*, 7.

Yuchen Lian, Arianna Bisazza, and Tessa Verhoef. 2025. Simulating the emergence of differential case marking with communicating neural-network agents. *Preprint*, arXiv:2502.04038.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540.

R. Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L. Griffiths. 2023. Embers of autoregression: Understanding large language models through the problem they are trained to solve. *Preprint*, arXiv:2309.13638.

Kanishka Misra and Najoung Kim. 2024. Generating novel experimental hypotheses from language models: A case study on cross-dative generalization. *Preprint*, arXiv:2408.05086.

Abhinav Patil, Jaap Jumelet, Yu Ying Chiu, Andy Lapastora, Peter Shen, Lexie Wang, Clevis Willrich, and Shane Steinert-Threlkeld. 2024. Filtered corpus training (FiCT) shows that language models can generalize from indirect evidence. *Transactions of the Association for Computational Linguistics*, 12:1597–1615.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*.

Karsten Schmidtke-Bode and Natalia Levshina. 2018. Reassessing scale effects on differential case marking: Methodological, conceptual and theoretical issues in the quest for a universal. In *The Diachronic Typology of Differential Argument Marking*, pages 509–537. Language Science Press.

Ilja A. Seržant and Alena Witzlack-Makarevich, editors. 2018. *Diachrony of differential argument marking*. Number 19 in Studies in Diversity Linguistics. Language Science Press, Berlin.

Kenny Smith and Jennifer Culbertson. 2025. Communicative pressures shape language during communication (not learning): Evidence from case-marking in artificial languages. *Cognition*, 263:106164.

Taiga Someya, Anej Svete, Brian DuSell, Timothy J. O'Donnell, Mario Giulianelli, and Ryan Cotterell. 2025. Information locality as an inductive bias for neural language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27995–28013, Vienna, Austria. Association for Computational Linguistics.

Shira Tal, Kenny Smith, Jennifer Culbertson, Eitan Grossman, and Inbal Arnon. 2022. The impact of information structure on the emergence of differential object marking: An experimental study. *Cognitive Science*, 46(3):e13119.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.

Bill VanPatten and Megan Smith. 2019. Word-order typology and the acquisition of case marking: A self-paced reading study in latin as a second language. *Second Language Research*, 35(3):397–420.

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

Jennifer C. White and Ryan Cotterell. 2021. Examining the inductive bias of neural language models with artificial languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 454–463, Online. Association for Computational Linguistics.

Tianyang Xu, Tatsuki Kuribayashi, Yohei Oseki, Ryan Cotterell, and Alex Warstadt. 2025. Can language models learn typologically implausible languages? *Preprint*, arXiv:2502.12317.

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

A DAM Injection

A.1 Corpora and Splits

We construct the base English corpus from a subset of the OpenSubtitles dataset (Lison and Tiedemann, 2016) (English side of the EN–FR OPUS release). As a light preprocessing step, we apply regex-based sentence splitting and filter sentences by length, retaining only those with 3–30 whitespace-delimited tokens. The resulting unperturbed corpus, prior to DAM injection, contains approximately 184M tokens and 21M sentences.

We select this dataset because its relatively well-formed conversational text supports more reliable automatic parsing and higher coverage for extracting transitive SVO clauses. The corpus is split into training, validation, and test sets with a 90/5/5 ratio, and these splits are defined before DAM injection and reused across all grammatical conditions to ensure strict comparability.

A.2 Parsing and SVO Extraction

Each sentence is parsed using spaCy (Honnibal et al., 2020), augmented with Benepar constituency parsing (Kitaev and Klein, 2018). For each sentence, we identify verbal predicates and extract clause-local predicate–argument frames by considering only dependents that are directly licensed by the predicate head. Each frame consists of one predicate together with at most one subject and one object-like argument per clause. Instances with multiple object candidates are excluded.

Argument spans are constructed by expanding noun-phrase heads into contiguous surface realizations, incorporating determiners, adjectival modifiers, compounds, and possessive elements.

In addition to bare nominal objects, we treat certain prepositional constructions as pseudo-objects when they function as predicate-selected complements. Operationally, we identify these cases as prepositional dependents attached to the verbal predicate whose complement forms a patient/theme-like argument. Such constructions are grouped into a single argumental unit spanning the preposition and its nominal complement. This design reflects a typological and constructional property of English, in which patient-like roles are frequently encoded via prepositional marking (e.g., *wait for the bus*, *listen to the story*) rather than bare object positions. Cross-linguistically, such prepositional realizations often correspond to bare objects in languages with richer case morphology, motivating their inclusion in our case-marking framework (Blake, 2001).

A.3 Semantic Trigger Annotation

From the validation split, we sample subject–object pairs and collect approximately 2k NP instances (balanced across subjects and objects), covering animacy, definiteness, and pronominality. We use the GPT-4o API (Hurst et al., 2024) to generate single-token pre-labels for each NP in context, using task-specific prompts (animate/inanimate; definite/indefinite; pronoun/common). All automatically generated labels are subsequently human-verified, with independent double annotation by authors. Disagreements are resolved through discussion and adjudication, following established practices (Tan et al., 2024).

We train three separate BERT-base classifiers (Devlin et al., 2019), one for each semantic trigger, on the verified seed data. The input to each classifier consists of the full sentence, followed by a special delimiter and the target noun phrase (e.g., `The dog chased the cat [NP] the cat`), and the output is a binary label indicating the corresponding semantic property (e.g., ANIMATE vs. INANIMATE). All classifiers are fine-tuned from `bert-base-uncased` for 10 epochs using AdamW (learning rate 2×10^{-5}), with held-out evaluation based on a stratified 80/20 split of the verified seed data. On held-out portions of the seed sets, all three classifiers achieve high accuracy (approximately 97%). We apply the trained classifiers to the entire corpus to assign animacy, definiteness, and pronominality labels to each argument prior to applying any DAM rules. For example, in the NP *the boy*, the head *boy* is labeled as [animate, definite, common].

Table 4 and Table 5 report the distributions of subject–object pairings and subject/object marginals across semantic triggers, showing the expected prominence asymmetries between subjects and objects.

A.4 Rule Injection and Dataset Construction

For each sentence containing at least one predicate with a single subject and a single object, we apply the DAM rules defined in Section 3 to determine whether to insert case markers. We introduce two marker symbols for this purpose: an agent marker [A], which targets the subject argument, and a patient marker [P], which targets the object argument.

In addition to the 18 experimental conditions, we include two control settings for comparison: (i) an unperturbed baseline trained on the original corpus, and (ii) a fully perturbed condition that inserts both agent and patient markers on every eligible S–V–O frame irrespective of licensing. Each marker is inserted at the right edge of the recorded NP span, preserving all original tokens and punctuation.

Each processed sentence is assigned to exactly one of three buckets:

- **Affected**: the sentence contains at least one predicate–argument frame for which the DAM rule licenses and inserts one or more markers.
- **Unaffected**: the sentence contains at least one valid predicate–argument frame, but no marker is licensed by the rule.
- **Invalid**: the sentence contains no predicate–argument frame satisfying the structural criteria (e.g., missing subject or object, clausal object, or multiple objects). All conditions share the same *Invalid* set.

Sentences containing valid S–V–O frames (*Affected* + *Unaffected*) constitute approximately 5% of all sentences in the raw corpus. We treat *Affected* and *Unaffected* sentences as positive training signals for learning the DAM rule, while *Invalid* sentences serve as background material that preserves the overall distribution of English outside the scope of DAM.

A.5 Human Inspection

To assess the reliability of the perturbation pipeline, we conduct a targeted human inspection on a small set of representative DAM rules. Specifically, we

Animacy pairs		Definiteness pairs		Pronominality pairs	
Pair	**Count**	**Pair**	**Count**	**Pair**	**Count**
animate–inanimate	858,545	definite–indefinite	669,965	pronoun–common	701,852
animate–animate	103,571	definite–definite	288,003	common–common	205,098
inanimate–inanimate	62,858	indefinite–indefinite	44,862	pronoun–pronoun	98,329
inanimate–animate	5,783	indefinite–definite	27,927	common–pronoun	25,478

Table 4: Distribution of subject–object pairings for three semantic triggers.

Animacy			Definiteness			Pronominality		
Category	**Subject**	**Object**	**Category**	**Subject**	**Object**	**Category**	**Subject**	**Object**
animate	962,116	109,354	definite	957,968	315,930	pronoun	906,950	230,576
inanimate	68,641	921,403	indefinite	72,789	714,827	common	123,807	800,181

Table 5: Marginal distributions for subjects and objects across the three semantic triggers.

Rule	**Set**	N	**Correct**	**Accuracy**
L-P-Def	affected	50	47	94.00%
L-P-Def	unaffected	50	49	98.00%
L-A-Pro-inv	affected	50	48	96.00%
L-A-Pro-inv	unaffected	50	50	100.00%
G-Ani	affected	50	49	98.00%
G-Ani	unaffected	50	50	100.00%
–	invalid	100	92	92.00%
Total	–	**400**	**385**	**96.25%**

Table 6: Results of human inspection for representative DAM rules and the *Invalid* set.

inspect *L-P-Def*, *L-A-Pro-inv*, and *G-Ani*, which together cover different dimensions of the design space. For each rule, we manually examine 50 *affected* and 50 *unaffected* sentences. In addition, we inspect 100 sentences sampled from the shared *Invalid* pool. The results are summarized in Table 6.

Overall, the pipeline performs reliably with respect to the intended perturbation rules. For *affected* and *unaffected* sentences, the remaining errors are largely attributable to coordinated noun phrases. For example, in "We do all the cooking [P] and cleaning", the marker is not placed at the right edge of the full coordinated object. Additionally, coordinated arguments such as "you and Emma", in which the conjuncts have different semantic properties, are not explicitly defined in our implementation.

For the *invalid* set, most errors arise from failures in extracting an underlying transitive frame. These include cases involving sentence-initial vocative constructions (e.g., "Vincent, you protected me"), yes–no questions (e.g. "Rygel, are you hearing this?"), and subordinate constructions (e.g. "I know people that swear by it"), where a predicate–argument structure is present but not successfully captured by the extraction pipeline.

B Model Training Details

For each DAM rule described in Section 3, we train a separate GPT-2-small language model from random initialization on the corresponding synthetic corpus. All models are trained under matched architectures and optimization settings across conditions.

Models are trained using a standard causal language modeling objective, without any explicit supervision or rule-specific signals beyond token prediction. Training data consist of the full rule-injected corpus stream, including all sentences from the *Affected*, *Unaffected*, and *Invalid* sets for each condition.

We use the standard GPT-2 tokenizer and extend the vocabulary with two special marker symbols [A] and [P]. The embedding matrix is resized accordingly. No other vocabulary items are added or modified.

All models are trained for a fixed budget of 15,000 optimization steps. Model checkpoints are saved every 500 steps during the first 8,000 training steps and every 1,000 steps thereafter.

Key training hyperparameters shared across all model runs are summarized in Table 7.

Hyperparameter	Value
Model	GPT-2-small
Context length	1024
Training steps	15k
Epochs	∼8.0–8.2[3]
Optimizer	AdamW
Learning rate	3×10^{-4}
LR schedule	Cosine (5% warmup)
Batch size	48×2
Precision	bf16
Checkpointing	500 / 1000 steps
Compute	NVIDIA A100 (∼6h)

Table 7: Training hyperparameters shared across all LM runs.

C Additional Experiments

C.1 Semantic Probing

We conduct a semantic probing experiment to test whether argument-level semantic properties are linearly recoverable from model representations using standard probing techniques (Hupkes et al., 2018; Hewitt and Liang, 2019; Belinkov, 2022). We focus on three binary features underlying our DAM rules: *animacy*, *definiteness*, and *pronominality*.

For each model, probing is performed only on the best training checkpoint based on validation perplexity. Probing sentences are drawn from the annotated test data. For each sentence, we extract the representation of the argument head (subject or object) by selecting the final-layer hidden state of the rightmost token whose character span overlaps the annotated head span.

For each feature, we train a separate binary linear classifier on top of the extracted representations, with probes trained independently for subjects and objects. Balanced datasets are constructed by sampling equal numbers of positive and negative instances, with 2,000 examples per class for training and 1,000 per class for testing. We report classification accuracy on the balanced test sets in Figure 6.

Across all rules, models show high linear separability for all three semantic features, with no systematic degradation across rule conditions. Subject representations consistently yield higher probing accuracy than object representations. This asymmetry aligns with the rule mastery results, where the performance gap between *natural* and *inverse* variants is smaller for *Local-P* rules than for *Local-A* rules.

[3]Models are trained for a fixed number of optimization steps. The effective number of epochs varies slightly across DAM rules due to differences in marker density.

Overall, these results indicate that differences in DAM rule mastery across conditions cannot be attributed to a failure to learn or encode semantic information.

C.2 BLiMP Evaluation

To assess whether DAM perturbations interfere with broader grammatical learning, we evaluate trained models on a subset of BLiMP benchmarks (Warstadt et al., 2020). We select eight BLiMP sub-tasks spanning core syntactic and syntax–semantics phenomena: *Determiner–Noun Agreement*, *Subject–Verb Agreement*, *NPI Licensing*, *Existential There Quantifiers*, *Animate Subject Passive*, *Animate Subject Transitive*, *Transitive*, and *Intransitive*.

To ensure distributional consistency between training and evaluation, we additionally construct DAM-perturbed versions of each selected BLiMP sub-task. In these variants, argument markers are inserted or retained according to the corresponding DAM rule, without altering the grammatical labels of the original BLiMP items.

We evaluate each model (including the *baseline* and *full-perturbation* controls) on DAM-perturbed BLiMP sets using a likelihood-based minimal-pair evaluation, as in Section 4.3. We report sub-task accuracy at the best training checkpoint in Figure 7 and summarize accuracy distributions across rules in Figure 8.

Across the eight BLiMP sub-tasks, DAM-perturbed models show accuracy distributions tightly centered around the unperturbed baseline, with no consistent downward shift [4]. Overall, these results indicate that DAM perturbations do not interfere with broader grammatical learning beyond the targeted argument-marking behavior.

[4]An exception arises in the *Animate Subject Transitive* subtask, which rewards analyses that prefer animate transitive subjects. Rules introducing markers on animate agents, such as *L–A–Ani-inv*, may therefore align more closely with BLiMP's grammatical variants (e.g., *Beth* $\boxed{A}$ *scares Roger*). As a result, observed accuracy differences on this sub-task reflect task–rule alignment effects rather than improvements in broader grammatical competence.

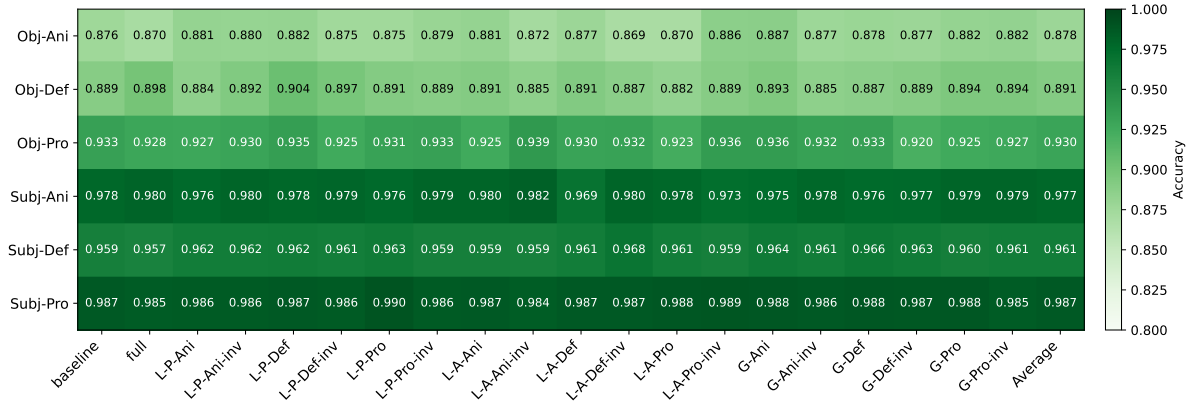


Figure 6: Semantic probing accuracy at the best training checkpoint across DAM conditions, evaluated separately for subject and object representations.

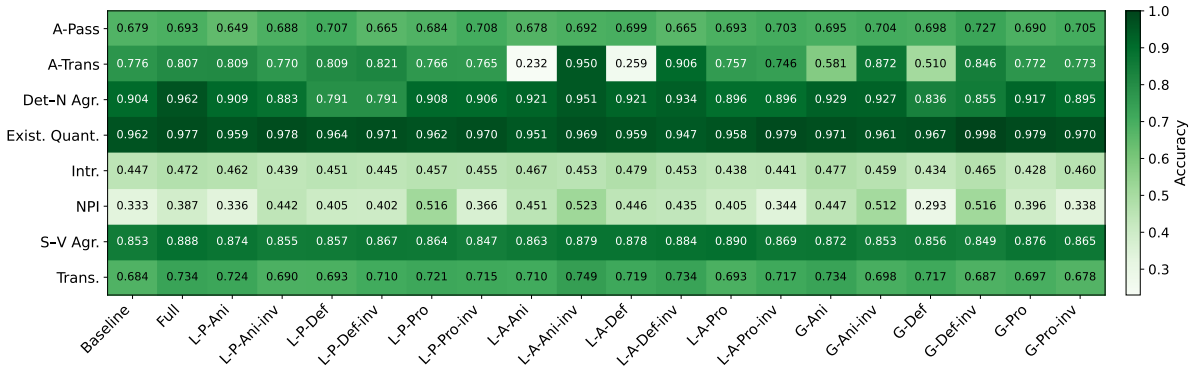


Figure 7: Full BLiMP evaluation results across all DAM rules and the baseline. Column abbreviations: A-Pass = Animate Subject Passive; A-Trans = Animate Subject Transitive; Det-N Agr. = Determiner–Noun Agreement; Exist. Quant. = Existential Quantifiers; Intr. = Intransitive; NPI = Negative Polarity Item Licensing; S–V Agr. = Subject–Verb Agreement; Trans. = Transitive.

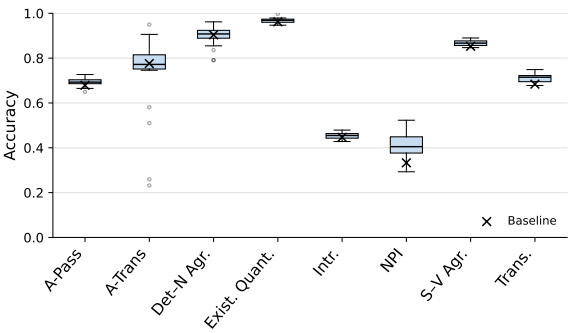


Figure 8: BLiMP accuracies across eight sub-tasks. Boxplots show the distribution across DAM rules; black "×" markers denote the baseline model.