Linguistic Research 43(2): 427-450
DOI: 10.17250/khisli.43.2.202606.004

Binding by structure, distracted by cues: Principle A in Language Models*

Jieun Kim**
(University of Ulsan)

Kim, Jieun. 2026. Binding by structure, distracted by cues: Principle A in Language Models. *Linguistic Research* 43(2): 427-450. Binding Principle A provides a well-studied testing ground for examining how hierarchical syntactic constraints regulate dependency formation (Chomsky 1981, 1986). Psycholinguistic research has long debated whether reflexive interpretation relies on structural filtering that excludes illicit antecedents in advance (Nicol and Swinney 1989; Sturt 2003), or on cue-based retrieval in which structural information competes with graded similarity cues (Lewis and Vasishth 2005; Dillon et al. 2014). This study revisits that debate in transformer-based language models. Using controlled reflexive configurations from the Pythia family (410M, 2.8B, 6.9B), we systematically dissociate structural accessibility from feature-based similarity by independently manipulating c-command, locality, antecedent number, distractor animacy, and model scale. We evaluate models using surprisal at the reflexive themselves. Across models and sentence types, a highly consistent ordering emerges: TRUE ANT < FALSE ANT < CORRUPTED. Structurally licensed plural antecedents yield the lowest surprisal, whereas structurally illicit but number-matching distractors produce only partial facilitation. This asymmetry provides direct evidence that model predictions are guided by structural accessibility rather than surface plurality alone. At the same time, the intermediate status of FALSE ANT reveals reliable distractor interference, indicating that feature-matching competitors are not fully ignored. Additional analyses show that animate distractors generate stronger interference than inanimate distractors, and that smaller models are substantially more susceptible to such interference than larger models. These findings suggest that transformer language models encode a hybrid retrieval profile: reflexive prediction is strongly constrained by structural information, yet remains probabilistically sensitive to competing semantic and featural cues.. **(University of Ulsan)**

* This work was financially supported by the Research Fund of the University of Ulsan (Research Number: 2026-0318)

** The author thanks Dorothy Ahn for insightful discussions on Binding Theory and the two anonymous reviewers for their constructive comments, which substantially improved both the dataset and the methodology presented in this paper.

Keywords reflexive binding, Binding Principle A, transformer-based language models, syntactic structure, computational model processing, cue-based retrieval

1. Introduction

Binding Principle A has long served as a central testing ground for theories of syntactic structure and its interface with interpretation. Within the generative tradition, reflexive anaphors are subject to strict structural constraints, most prominently c-command and locality conditions (Chomsky 1981, 1986). These constraints are taken to reflect the hierarchical organization of grammar rather than linear order, semantic plausibility, or discourse prominence. Theoretical analyses of reflexive interpretation thus provide insight into how syntactic configuration constrains dependency formation.

Beyond formal theory, however, an important question concerns how such structural constraints are implemented in cognitive systems. Psycholinguistic research has debated whether reflexive interpretation is governed by an early structural filter that categorically excludes grammatically illicit antecedents (Nicol and Swinney 1989; Sturt 2003; Gordon et al. 2006), or whether antecedent retrieval proceeds through a cue-based mechanism in which structural position competes with featural similarity (Garrod and Terras 2000; Lewis and Vasishth 2005; Van Dyke and McElree 2006; Van Dyke and McElree 2011; Dillon et al. 2014). This debate raises a broader issue about the architecture of the language faculty: are hierarchical constraints represented as categorical grammatical restrictions, or are they integrated as weighted cues within a memorybased retrieval process?

Recent advances in transformer-based language models introduce a new empirical domain in which this long-standing debate can be re-examined. Although such models are trained solely on distributional input and do not implement temporally staged parsing procedures, they nonetheless exhibit sensitivity to certain hierarchical dependencies. This raises the question of whether structural constraints such as those governing reflexive binding are encoded in a way that functionally excludes illicit antecedents, or whether model behavior reflects graded sensitivity to surface-level similarity cues. The present study investigates this issue by systematically dissociating structural accessibility from feature-based similarity in reflexive configurations. By independently manipulating c-command, locality, and agreement-related features (e.g.,

number and animacy), we test whether structurally inaccessible noun phrases influence the probabilistic prediction of reflexives. In doing so, the study aims to contribute to the theoretical understanding of how hierarchical constraints may be represented in transformer-based language models.

2. Theoretical background

2.1 Structural constraints on reflexive interpretation

Within the Government and Binding framework, Binding Principle A requires that an anaphor be bound within its governing category (Chomsky 1981, 1986). Binding involves both coindexation and a structural relation of c-command, and the dependency must be local. Consequently, two structural conditions determine the set of possible antecedents for a reflexive expression: (i) the binding relation must satisfy locality constraints, and (ii) the antecedent must c-command the anaphor. These requirements are illustrated by the contrast in (1).

(1) a. John$_i$ said that Bill$_j$ criticized himself$_{j/*i}$.
b. The mother$_i$ of Amy$_j$ enjoys taking a picture of herself$_{i/*j}$.

In (1a), only Bill can serve as a licit antecedent of himself, whereas John cannot. Binding Principle A requires that a reflexive expression be bound within the local clause, which defines the relevant domain for the locality condition. Example (1b) illustrates the c-command requirement. Although Amy linearly precedes the reflexive, it does not c-command herself because it is embedded within the possessive phrase the mother of Amy. As a result, Amy cannot serve as a legitimate antecedent of the reflexive. Subsequent theoretical developments have refined these notions (Reinhart and Reuland 1993; Reuland 2011), but the central insight remains: reflexive interpretation is fundamentally structure-sensitive. In configurations where a structurally inaccessible noun phrase is linearly closer to the reflexive than the grammatically licit antecedent, Principle A predicts that only the structurally licensed candidate can serve as the antecedent. Such cases provide clear evidence that hierarchical configuration, rather than linear proximity, determines interpretive

possibilities. The theoretical significance of this observation extends beyond descriptive adequacy. If reflexive dependencies are indeed constrained by hierarchical structure, then any adequate model of linguistic competence must encode structural relations in a way that distinguishes syntactic accessibility from mere feature similarity or discourse salience.

2.2 Structural filtering and cue-based retrieval in sentence processing

Psycholinguistic research has long sought to determine how the structural constraints imposed by binding theory are implemented during real-time sentence comprehension. Early experimental studies (Nicol and Swinney 1989, 2003; Clifton et al. 1997; Sturt 2003) suggested that grammatically illicit antecedents exert little influence on the early stages of reflexive processing. Using cross-modal priming and eye-tracking paradigms, these studies found minimal evidence that structurally inaccessible noun phrases are initially considered as potential antecedents. Such findings have often been interpreted as supporting a structural filtering hypothesis, according to which syntactic constraints restrict the candidate set prior to memory retrieval. Under this view, the parser consults only grammatically licensed antecedents, and structurally inaccessible noun phrases are excluded from consideration at the outset.

In contrast, cue-based retrieval models (Garrod and Terras 2000; Lewis and Vasishth 2005; Van Dyke and McElree 2011; Dillon et al. 2014) propose that antecedent resolution is implemented as a feature-based memory search over representations stored in content-addressable memory. On this view, retrieval is guided by a set of cues associated with the target dependency (e.g., [+NP], [+subject], [+animate], [+plural], structural position). All candidate items in memory are evaluated in parallel with respect to these cues, and the item that best satisfies the cue set is retrieved. Because access depends on cue matching rather than serial search, retrieval latency need not increase monotonically with the number of stored candidates. Evidence from Speed–Accuracy Tradeoff studies supports this assumption (McElree 2000; McElree et al. 2003; Martin and Mcelree 2009). A central consequence of this architecture is the prediction of similarity-based interference. Because retrieval operates over the full memory representation, grammatically illicit noun phrases that partially match the retrieval cues may compete with the correct antecedent. For example, a noun

phrase that matches number or animacy cues may be erroneously activated despite being structurally inaccessible. Such interference effects suggest that structural constraints may interact with graded featural cues during retrieval, rather than functioning solely as categorical pre-retrieval filters. Subsequent work has refined this cue-based perspective by proposing that retrieval cues need not contribute equally. Dillon et al. (2014) argue that cues associated with structural locality may receive greater weight than more distal or weaker cues, yielding a bias toward antecedents within the local syntactic domain. Under this weighted-cue approach, retrieval remains fundamentally cue-based, but cue strength is modulated by the relative diagnosticity of structural information. Importantly, locality effects are therefore treated not as a separate search mechanism, but as emerging from differential cue weighting within the broader content-addressable retrieval architecture.

The resulting debate concerns the representational architecture underlying sentence processing. One possibility is that syntactic constraints operate categorically, filtering out grammatically illicit antecedents prior to memory retrieval. Alternatively, structural information may function as one cue—albeit a highly privileged one—within a unified retrieval mechanism that also incorporates person, number, animacy, and discourse prominence. Although this debate has primarily been framed in terms of temporal dynamics in human sentence processing, its theoretical implications extend more broadly to questions about how hierarchical structure is represented and accessed in linguistic systems.

2.3 Reformulating the debate for Language Models

Language models differ from human comprehenders in an important respect. Human sentence processing unfolds over time through incremental structure building and memory retrieval, whereas transformer-based language models compute a probability distribution over the next token given a preceding context. As a result, psycholinguistic distinctions such as early vs. late processing stages do not transfer directly to language models. Nevertheless, the theoretical debate can be recast in predictive terms. The key question is whether structurally illicit antecedents are effectively excluded from a model's next-token expectations, or whether their featural similarity continues to influence probability assignment despite structural

inaccessibility. A growing body of research shows that neural language models capture some hierarchical dependencies, particularly in agreement and acceptability tasks (Linzen et al. 2016; Gulordava et al. 2018; Marvin and Linzen 2018). However, reflexive binding offers a more precise testing ground because structural accessibility and feature similarity can be directly dissociated.

The present study asks whether language-model behavior is better characterized by (i) a structuralfiltering profile, in which illicit antecedents exert little influence, (ii) a competing-cues profile, in which feature-matching distractors substantially affect prediction, or (iii) a weighted-cue profile, in which multiple cues matter but structural locality receives the strongest weight (Dillon et al. 2014). To test these possibilities, we construct controlled reflexive configurations in which a structurally inaccessible noun phrase may nonetheless match the reflexive in number or animacy. If models rely primarily on structural filtering, such distractors should minimally affect surprisal. If models are strongly cue-driven, distractors should substantially facilitate reflexive prediction. If structural information functions as a privileged cue, distractors should yield intermediate interference while structurally licensed antecedents remain strongly preferred.

3. Experiment

This section examines whether transformer-based language models reflect Binding Principle A in reflexive prediction. Using models from the Pythia family, we evaluate how strongly model expectations for the reflexive *themselves* are shaped by the presence or absence of structurally appropriate antecedents. We first introduce the controlled dataset used in the study, and then describe the surprisal-based evaluation procedure and statistical analyses.

3.1 Dataset construction

The experimental dataset was designed to systematically evaluate how reflexive interpretation is influenced by two factors: structural accessibility and featural similarity. Each item contains two noun phrases preceding the reflexive *themselves*: one structurally licit antecedent (henceforth true antecedent) and one structurally illicit

competitor (henceforth false antecedent). The reflexive is placed in sentence-final position so that all relevant structural information is available prior to the prediction of the target token.

Set. The dataset consists of three structural sets. The three sets differ in the structural relation between the reflexive and the potential antecedents. This manipulation is designed to test whether the model is sensitive to different binding configurations, rather than relying solely on linear proximity or surface number agreement. In all three sets, the False Antecedent condition introduces a plural noun that matches the reflexive in number but does not satisfy the relevant structural constraint for licensing the reflexive. The dataset includes three structural configurations that manipulate the manner in which the false antecedent violates the constraints of Binding Principle A. Recall that Principle A requires a reflexive to be locally bound by an antecedent that c-commands it. Each configuration therefore introduces a structurally inaccessible competitor that violates one or more of these conditions.

Table 1. Representative examples from the Binding Principle A datasets, restricted to N1-animate conditions. Rows are grouped by the animacy of N2. Red marks the structurally licensed antecedent position, blue marks the structurally illicit distractor position, and green marks the reflexive target *themselves*. Coloring is applied consistently across singular and plural forms.

Set	Condition	N2 Animacy	Example sentence
Set 1 [-local, -c-command]	True_Ant	Animate	The authors that the athlete introduced damaged themselves.
	False_Ant	Animate	The author that the athletes introduced damaged themselves.
	Corrupted	Animate	The author that the athlete introduced damaged themselves.
	True_Ant	Inanimate	The authors that the document introduced damaged themselves.
	False_Ant	Inanimate	The author that the documents introduced damaged themselves.
	Corrupted	Inanimate	The author that the document introduced damaged themselves.
Set 2 [-local, +c-command]	True_Ant	Animate	The athlete showed that the authors blamed themselves.
	False_Ant	Animate	The athletes showed that the author blamed themselves.
	Corrupted	Animate	The athlete showed that the author blamed themselves.
	True_Ant	Inanimate	The document showed that the authors blamed themselves.
	False_Ant	Inanimate	The documents showed that the author blamed themselves.
	Corrupted	Inanimate	The document showed that the author blamed themselves.
Set 3 [+local, -c-command]	True_Ant	Animate	The teachers of the musician defended themselves.
	False_Ant	Animate	The teacher of the musicians defended themselves.
	Corrupted	Animate	The teacher of the musician defended themselves.
	True_Ant	Inanimate	The teachers of the report defended themselves.
	False_Ant	Inanimate	The teacher of the reports defended themselves.
	Corrupted	Inanimate	The teacher of the report defended themselves.

Set 1 (Locality and C-command Violation). In Set 1, the false antecedent violates both the locality requirement and the c-command condition. It appears inside a relative clause modifying the true antecedent, and is therefore structurally embedded. As a result, although it is linearly closer to the reflexive, it cannot serve as a licit antecedent. This set serves as the reference condition in the mixed-effects analysis: *The N1 that the N2 V V themselves*.

Set 2 (Locality Violation). In Set 2, the false antecedent violates the locality requirement. The true antecedent (N2) appears in the embedded clause containing the reflexive, whereas the false antecedent (N1) appears in the matrix clause. Thus, the false antecedent c-commands the reflexive but is too distant structurally to license it: *The N1 V that the N2 V themselves*.

Set 3 (C-command Violation). In Set 3, the false antecedent violates the c-command condition but not locality. It appears inside a possessive phrase, which prevents it from c-commanding the reflexive while remaining within the same local domain. Consequently, the false antecedent is structurally local but still unavailable as a licit antecedent: *The N1 of the N2 V themselves*.

Condition. The condition factor is the central manipulation in the dataset. In the TRUE ANTECEDENT condition, the structurally licensed antecedent is plural and therefore matches the plural reflexive *themselves*. In the FALSE ANTECEDENT condition, the structurally illicit noun is plural, while the structurally licensed antecedent is singular. This condition tests whether the model is influenced by a number-matching but structurally inappropriate noun. In the CORRUPTED condition, neither candidate antecedent provides an appropriate plural antecedent for *themselves*. A second manipulation concerns number agreement. Because the reflexive *themselves* requires a plural antecedent, number morphology is used to determine whether a noun phrase can serve as a morphologically compatible antecedent. For each structural configuration, three sentence types are constructed:

1. CORRUPTED: both noun phrases are singular while the reflexive is plural, creating number mismatch for all potential antecedents.
2. TRUE ANT: only the true antecedent is pluralized, yielding a configuration in which the structurally licit antecedent satisfies both structural and agreement constraints.
3. FALSE ANT: only the false antecedent is pluralized, while the true antecedent

remains singular.

N2 Animacy. In addition to the above structural and number-feature manipulations, the dataset also varies featural similarity between the reflexive and the false antecedent. In particular, the feature ANIMACY (Animate vs. Inanimate) is manipulated for the false antecedent to measure the semantic interference effect of the False Ant and thereby, helping to support for the early-filtering approach or for the cue-based retrieval approach. Animacy was chosen because it has been widely reported as a retrieval cue in psycholinguistic studies of dependency resolution (Lewis and Vasishth 2005; Van Dyke and McElree 2006; Dyke 2007)

Measured target. For each sentence, surprisal was measured at the reflexive *themselves*. Lower surprisal indicates that the model assigns a higher probability to the reflexive in the given context. Details of the surprisal computation are provided in the next subsection.

Dataset construction. All experimental sentences were generated using a uniform template-based procedure implemented in Python. Each structural configuration corresponded to a fixed syntactic schema, while lexical items and number morphology were systematically varied across conditions. The dataset contained 100 items per condition, yielding a total of 100 × 3 structural sets ×2 animacy conditions ×3 antecedent conditions = 1,800 sentences.

3.2 Methods

Model behavior was evaluated using surprisal at the reflexive token *themselves*. We compare surprisal across conditions that vary whether a plural antecedent is structurally licensed (TRUE ANT), structurally illicit but feature-matching (FALSE ANT), or absent altogether (CORRUPTED). The relative ordering among these conditions reveals the extent to which model predictions are guided by structural accessibility versus cue-based interference.

Surprisal. The primary measurement used in this study is surprisal, a quantity commonly used in computational psycholinguistics to estimate the predictability of

a word given its preceding context (Shannon 1948; Hale 2001; Levy 2008; Smith and Levy 2013). For a token w_i given preceding context $w_{1:i-1}$, surprisal is defined as:

$$S(w_i) = -\log P(W_i | W_{1:i-1})$$

Lower surprisal indicates that the token is more strongly expected in context. In the present study, we compute surprisal for the reflexive token *themselves* by providing the model with each sentence prefix up to (but excluding) the reflexive and extracting its conditional probability. This provides a direct estimate of how strongly the model expects a plural reflexive under each antecedent configuration.

3.3 Models

All experiments were conducted using models from the PYTHIA family, a suite of decoder-only transformer language models based on the GPT-NeoX architecture. The Pythia models were designed with an explicit focus on transparency and interpretability research: they are trained on a standardized dataset (The Pile) with carefully controlled checkpoints and are released in multiple parameter scales to facilitate systematic analysis of representational and behavioral properties across model sizes.

The interpretability-oriented design of Pythia makes it particularly suitable for investigating questions about internal linguistic representations. Because the models share architecture and training data across parameter scales, differences in behavior can be more plausibly attributed to model capacity rather than confounding architectural or corpus factors. This property is especially relevant for the present study, which aims to examine how hierarchical constraints on reflexive binding are encoded in transformer-based language models.

To assess whether the observed effects depend on model scale, we compared three parameter sizes: 410M, 2.8B, and 6.1B. All models were evaluated on identical experimental materials and with the same inference procedure. This scaling comparison allows us to examine whether sensitivity to structural accessibility and animacy-based interference varies as a function of model capacity, thereby providing insight into the relationship between parameter scale and the representation of

syntactic dependencies.

4. Results and discussions

4.1 Overall surprisal patterns

We begin by examining the mean surprisal of the reflexive *themselves* across structural sets, antecedent conditions, and model sizes. Figure 1 summarizes the sentence-level means with 95% confidence intervals for the three Pythia models (6.9B, 2.8B and 410M). Lower surprisal indicates that the model assigns higher probability to the reflexive in the preceding context. Across all three models, a highly consistent ordering emerges: the TRUE ANT condition yields the lowest surprisal, the FALSE ANT condition yields intermediate surprisal, and the CORRUPTED condition yields the highest surprisal. This pattern is robust across all three structural sets. Since the CORRUPTED condition serves as the reference level in the mixed-effects model (together with Set 1, the 6.9B model, and inanimate distractors), this descriptive pattern already suggests that the availability of a structurally licensed plural antecedent strongly facilitates prediction of the reflexive. The effect is largest in the TRUE ANT condition, where the antecedent satisfies the structural requirements of Binding Principle A and matches the reflexive in number as shown in the following Figure 1 and Table 2.

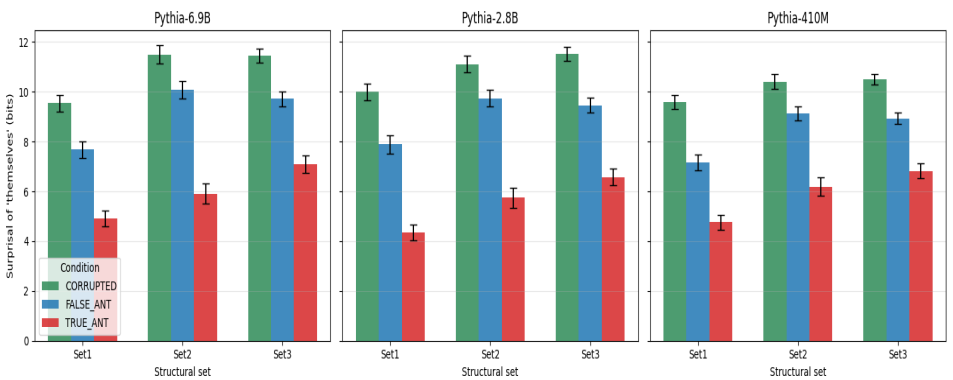


Figure 1. Mean surprisal (bits) of *themselves* across structural sets, antecedent conditions, and model sizes. Error bars indicate 95% confidence intervals. Across all models, the TRUE ANT condition yields the lowest surprisal, followed by FALSE ANT, with CORRUPTED.

In all models, this manipulation substantially lowers surprisal relative to the baseline. Thus, the models systematically prefer *themselves* when a structurally licit antecedent is present. The FALSE ANT condition reveals a more graded pattern. Although the plural distractor is structurally illicit, it nevertheless lowers surprisal relative to the CORRUPTED baseline. This indicates that the models are partially influenced by feature-matching but structurally inappropriate nouns. However, the magnitude of this facilitation is consistently smaller than that observed in the TRUE ANT condition. In other words, structural licensing dominates featural similarity.

A second notable pattern concerns model scaling. While all three models exhibit the same qualitative ordering (TRUE ANT < FALSE ANT < CORRUPTED), the mixed-effects analysis reveals that scaling primarily modulates susceptibility to structurally illicit distractors rather than sensitivity to structurally licensed antecedents.

Table 2. Mean surprisal (bits) of *themselves* across model size, structural set, and antecedent condition. Within each model, the highest value is highlighted in red and the lowest value in blue.

Model	Set	Corrupted	False_Ant	True_Ant
6.9B	Set1	9.54 ± 0.33	7.67 ± 0.34	4.90 ± 0.32
6.9B	Set2	11.50 ± 0.37	10.08 ± 0.36	5.90 ± 0.40
6.9B	Set3	11.46 ± 0.29	9.72 ± 0.30	7.10 ± 0.34
2.8B	Set1	10.00 ± 0.34	7.89 ± 0.36	4.35 ± 0.32
2.8B	Set2	11.12 ± 0.34	9.74 ± 0.34	5.75 ± 0.41
2.8B	Set3	11.53 ± 0.29	9.46 ± 0.30	6.58 ± 0.34
410M	Set1	9.59 ± 0.29	7.18 ± 0.32	4.76 ± 0.31
410M	Set2	10.41 ± 0.30	9.13 ± 0.28	6.19 ± 0.37
410M	Set3	10.49 ± 0.21	8.92 ± 0.23	6.83 ± 0.29

Using the 6.9B model as the reference level, the FALSE ANT facilitation effect becomes progressively larger as model size decreases, particularly in Set 1. In other words, smaller models are more strongly influenced by plural distractors that are linearly available but structurally inaccessible, indicating greater interference from feature-matching competitors. By contrast, the TRUE ANT condition does not display a comparably consistent scaling trend. Across model sizes, the presence of a structurally licit plural antecedent reliably produces a large reduction in surprisal, but this benefit does not systematically increase or decrease with parameter count. This suggests that once the correct antecedent is structurally available, all three models are able to exploit

that cue effectively. Taken together, these findings indicate an asymmetry in the role of scaling. Increasing model size does not fundamentally strengthen the licensing effect of valid antecedents, which is already robust across models. Instead, it mainly improves the filtering of invalid but feature-matching distractors. Larger models therefore appear less prone to interference, rather than categorically better at recognizing licensed antecedents.

Taken together, the descriptive results suggest a hybrid profile of behavior: the models strongly favor structurally licensed antecedents, yet remain susceptible to interference from structurally illicit but feature-matching distractors.

4.2 Mixed effects statistical analysis

The descriptive surprisal patterns reported in the previous subsection provide an informative first overview, but they are not sufficient to capture the full complexity of the experimental design. Our dataset jointly manipulates multiple theoretically relevant factors whose effects may interact with one another, making separate pairwise comparisons inadequate.

First, the set factor (Set 1, Set 2, Set 3) varies the structural configuration of the binding dependency according to which constraint the false antecedent violates. Set 1, which serves as the reference level, combines violations of both locality and c-command. In addition, these structural manipulations also alter the linear proximity of the distractor noun. For example, in Set 2 the false antecedent is linearly more distant from the reflexive than the local true antecedent, potentially reducing interference. Second, the condition factor contrasts CORRUPTED, FALSE ANT, and TRUE ANT. This manipulation directly tests whether the model distinguishes structurally licensed antecedents from structurally illicit but feature-matching distractors. Third, the model size factor allows us to evaluate scaling effects across the three Pythia models (410M, 2.8B, and 6.9B), asking whether larger models exhibit stronger structural sensitivity or weaker distractor interference. Fourth, the N2 animacy factor systematically varies whether the false antecedent is animate or inanimate. As motivated in Section 3.1, this manipulation tests whether semantically prominent distractors generate stronger interference effects.

Because these factors are simultaneously manipulated and potentially interdependent, their contributions cannot be reliably assessed through isolated

comparisons. We therefore employ a linear mixed-effects model, which estimates the independent and interactive contributions of each predictor while controlling for repeated observations over lexicalized items. This approach offers two major advantages: it permits principled inference over the full factorial design, and it generalizes beyond specific lexical items used in the dataset. We fit the following model to all sentence-level observations across the three datasets and three model sizes:

surprisal bits ~ condition * set * model size * N2 animacy + (1|item)

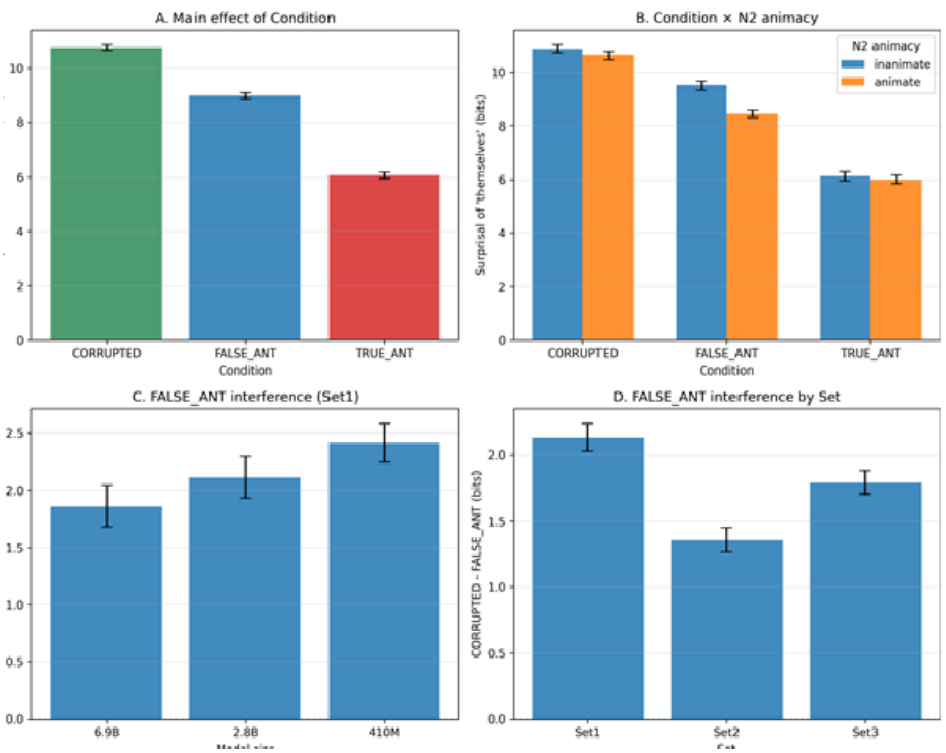
The reference levels were the 6.9B model, Set 1, the CORRUPTED condition, and inanimate distractors. Because the complete mixed-effects output included a large number of fixed-effect terms and interactions, Table 3 presents a selective subset of coefficients that were either statistically robust or theoretically central to our research questions. In particular, we highlight effects bearing on structural set differences, distractor interference, animacy modulation, and model scaling. The following subsections focus on these most informative patterns.

Table 3. Core mixed-effects results corresponding directly to the four summary panels in Figure 2. Reference levels were the 6.9B model, Set 1, the CORRUPTED condition, and inanimate distractors. Negative coefficients indicate lower surprisal (greater predictability of *themselves*) relative to the reference level. Significance codes: *** p < .001, ** p < .01, * p < .05, n.s. = not significant.

Effect	Comparison (vs. reference)	Coef.	SE	p	Sig.
Condition	False_Ant vs. Corrupted	-1.455	0.190	< .001	***
Condition	True_Ant vs. Corrupted	-4.487	0.190	< .001	***
Condition × Animacy	False_Ant × Animate	-0.818	0.268	.002	**
Condition × Model	False_Ant × 410M	-0.732	0.268	.006	**
Condition × Model	False_Ant × 2.8B	-0.324	0.268	.227	n.s.
Condition × Set	False_Ant × Set2	0.558	0.268	.037	*
Condition × Set	False_Ant × Set3	0.352	0.233	.131	n.s.

Table 3 summarizes the subset of mixed-effects coefficients that correspond directly to the four core patterns visualized in Figure 2. This table highlights the theoretically most interpretable effects that motivate the subsequent discussion. First, the significant main effects of condition replicate the basic surprisal ordering shown in Panel A: TRUE ANT produces the strongest reduction in surprisal, FALSE ANT yields an

intermediate reduction, and CORRUPTED remains the least favorable condition. This confirms that models strongly prefer structurally licensed antecedents while still showing partial sensitivity to feature-matching distractors. Second, the significant FALSE ANT × Animate interaction corresponds to Panel B, where animate distractors induce stronger facilitation than inanimate distractors. This suggests that retrieval interference is selectively enhanced when the distractor is semantically prominent. Third, the significant FALSE ANT × 410M interaction captures the pattern shown in Panel C: smaller models exhibit stronger interference effects than larger models. Thus, increasing scale appears to improve filtering of structurally illicit distractors. Fourth, the FALSE ANT × Set coefficients correspond to Panel D, where distractor interference varies across structural configurations. In particular, Set 2 shows attenuated interference relative to Set 1, indicating that the greater linear distance of the interfering distractor facilitates processing in this configuration.



Figur2. Mixed effects core graph

Taken together, Figure 2 provides an intuitive visualization of the main statistical findings, while Table 3 confirms that these patterns are reliable under simultaneous control of all predictors. The following subsections examine each of these effects in greater detail: overall structural sensitivity, model-scaling effects, structural modulation of interference, and animacy-based retrieval effects.

4.3 Overall structural sensitivity: Main effect of condition

We begin with the most robust result of the experiment: the main effect of antecedent condition (Figure 2A). Across all datasets, model sizes, and animacy conditions, surprisal at *themselves* follows a highly consistent ordering: TRUE ANT < FALSE ANT < CORRUPTED

That is, the reflexive is most expected when a structurally licensed plural antecedent is available, least expected when no plural antecedent is present, and intermediate when only a structurally illicit but featurematching plural distractor is available. This pattern provides strong evidence that the models are not relying solely on superficial number agreement. If plurality alone determined reflexive prediction, the FALSE ANT condition should pattern with TRUE ANT. Instead, the clear separation between these two conditions indicates that the models are sensitive to structural constraints that distinguish licensed from unlicensed antecedents. At the same time, the intermediate status of FALSE ANT is equally informative. Although the distractor is structurally inaccessible, it still lowers surprisal relative to the CORRUPTED baseline. Thus, model predictions reflect a hybrid profile: substantial structural sensitivity coexisting with partial cue-based interference. The mixed-effects analysis confirms this interpretation statistically. Relative to CORRUPTED, both FALSE ANT and TRUE ANT significantly reduce surprisal, with the TRUE ANT coefficient substantially larger in magnitude.

Importantly, this three-way ordering bears directly on competing theories of antecedent retrieval. Under a strict symbolic filtering account, structurally illicit distractors should contribute little or no facilitation, leading FALSE ANT to pattern closely with CORRUPTED. Under a purely feature-driven competition account, however, any plural antecedent should facilitate reflexive prediction similarly, leading FALSE ANT to pattern with TRUE ANT. Neither prediction is supported. Instead, the observed intermediate position of FALSE ANT suggests that structural accessibility

strongly constrains retrieval while feature-matching distractors still exert partial influence. This profile closely resembles the structure-sensitive cue-based retrieval account proposed by Dillon et al. (2014). In their analysis of human sentence processing, syntactic constraints do not operate as an absolute early filter that excludes all grammatically illicit candidates from memory access. Nor do all candidates compete on equal terms. Rather, retrieval proceeds through weighted competition in which structurally licensed antecedents receive a substantial advantage, while partially matching distractors may still generate interference.

The present results suggest that the Pythia model family exhibits a strikingly similar computational signature. Reflexive prediction appears neither fully rule-filtered nor purely surface-associative. Instead, it reflects graded competition in which Principle A–relevant structural information functions as a privileged probabilistic cue. In this respect, these models align more closely with human memory-retrieval accounts of binding than with a strictly serial "apply Principle A first, then retrieve" parser architecture.

4.4 Anymiacy-based retrieval effects

Figure 2B shows that animacy selectively modulates the FALSE ANT condition. Animate distractors lower surprisal more than inanimate distractors, whereas the difference is minimal in the CORRUPTED and TRUE ANT conditions. This selective pattern suggests that semantic prominence influences retrieval competition primarily when antecedent resolution is uncertain.

The mixed-effects model confirms this statistically through a significant FALSE ANT × Animate interaction. Thus, distractors are more disruptive when they are animate entities such as athlete or musician than when they are inanimate nouns such as document or report. This finding is broadly consistent with cue-based retrieval accounts in which semantically salient nouns serve as stronger competitors during dependency resolution. However, at the same time, as one reviewer insightfully pointed out, the animate advantage may not be interpreted as a purely semantic animacy effect. Contemporary English increasingly permits singular *they/themselves* with human antecedents, especially when gender is unspecified or non-binary (Arnold et al. 2021; Watson et al. 2025). Consequently, animate singular distractors may function not only as semantically prominent competitors, but also as partially compatible antecedent

candidates for *themselves*. The observed animate advantage may therefore reflect a combination of semantic salience and residual compatibility with singular human antecedents. However, the present results do not suggest that singular-they compatibility alone is sufficient to explain the full pattern. If compatibility with singular animate antecedents were the dominant factor, one might expect substantially larger differences involving the CORRUPTED condition, where no structurally licensed plural antecedent is available. Instead, the effects remain strongest in configurations where structural competition is already relevant, suggesting that antecedent accessibility and semantic prominence continue to play a substantial role. To further decompose this pattern, Figure 3 plots interference magnitude (CORRUPTED − FALSE ANT) separately by model size, animacy, and structural set. Three notable observations emerge.

First, the animate advantage is highly consistent across nearly all model and structural combinations. In every set, animate distractors induce stronger interference than inanimate distractors. This robustness suggests that semantic salience is deeply integrated into antecedent retrieval processes. Second, the animacy effect coexists with model-scaling effects. Even when model size changes, animate distractors remain stronger competitors than inanimate distractors. Thus, scaling reduces overall interference but does not eliminate semantic cue weighting. Third, animacy interacts with structure. The animate boost is strongest in Set 1 and Set 3, but weaker in Set 2, where the distractor is linearly more distant. This again suggests that semantic prominence and structural accessibility jointly determine retrieval strength.

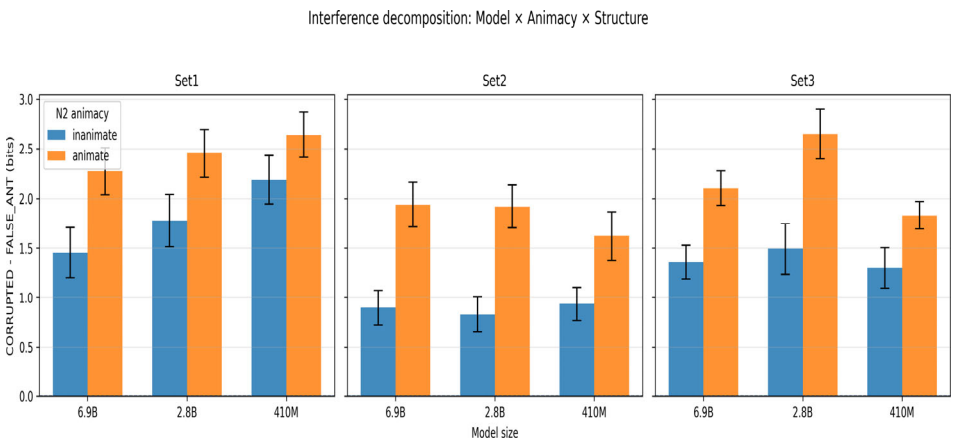


Figure 3. Decomposition of Animacy Effects

These findings parallel psycholinguistic accounts of cue-based retrieval in human sentence processing, where animate nouns often produce stronger interference than less salient inanimate nouns (Van Dyke and McElree 2006; Dyke 2007). More broadly, the present results suggest that language models, like human comprehenders, integrate structural constraints with graded semantic retrieval cues rather than relying on syntax alone.

The current dataset was not specifically designed to isolate the contribution of singular-they compatibility. A cleaner comparison would ideally contrast animate and inanimate singular antecedents while holding semantic plausibility constant. In the present materials, however, many cases with inanimate TRUE ANT antecedents risk semantic infelicity between the subject noun and the predicate containing themselves, making fully matched contrasts difficult to interpret. Future work should therefore construct tightly controlled datasets that independently manipulate animacy, humanness, antecedent number, and semantic plausibility in order to determine whether the present advantage reflects genuine animacy-based retrieval, compatibility with contemporary singular-they usage, or an interaction of both factors.

4.5 Model scaling and distractor interference

Figure 2C isolates the distractor-based facilitation effect in Set 1 by plotting the difference between CORRUPTED and FALSE ANT. Larger values indicate stronger interference from the structurally illicit plural distractor. A clear monotonic trend emerges: 6.9B < 2.8B $\ll$ 410M

Smaller models exhibit greater susceptibility to distractor interference, whereas larger models more effectively discount the grammatically illicit antecedent. This scaling pattern is theoretically important because it suggests that increased parameter count does not primarily improve the ability to recognize structurally licensed antecedents. As shown in the previous subsection, all models already display a strong TRUE ANT advantage. Instead, scaling mainly improves the filtering of competing feature-matching distractors. The mixed-effects model supports this conclusion. The significant FALSE ANT $\times$ 410M interaction indicates that the smallest model shows reliably greater interference than the 6.9B reference model. Thus, larger models appear not categorically more syntactic, but more selective in retrieval under competition.

This result aligns with recent views that scaling enhances robustness of internal representations rather than introducing entirely new grammatical mechanisms. In the present case, the core structural preference is already present in smaller models, while larger models better suppress misleading plural competitors.

4.6 Structural modulation of interference

Figure 2D examines whether distractor interference is constant across sentence types. It is not. The magnitude of the FALSE ANT effect differs systematically across the three structural configurations, with the weakest interference observed in Set 2. This asymmetry is theoretically revealing because the three datasets were designed to vary the structural status of the distractor antecedent. In Set 1, the distractor is both nonlocal and fails to c-command the reflexive. In Set 2, the distractor occurs in the matrix clause and is linearly farther from the reflexive than the local antecedent. In Set 3, the distractor is clause-local but embedded inside a possessive phrase.

The reduced interference in Set 2 suggests that linear distance remains an active factor in model retrieval. When the distractor is farther away from the reflexive, its influence weakens even though its number features remain compatible. This implies that model behavior reflects a combination of structural and recency-based cues. The mixed-effects coefficients confirm this interpretation: the FALSE ANT × Set interaction shows that distractor effects are significantly attenuated in Set 2 relative to Set 1. Thus, interference is not a fixed consequence of feature overlap, but is modulated by the structural and positional accessibility of the competitor. More broadly, these findings indicate that models do not implement an all-or-none structural filter. Instead, multiple cues—including hierarchical structure and linear proximity—jointly determine antecedent competition.

5. General discussion and conclusion

This study investigated whether transformer language models reflect Binding Principle A during reflexive prediction, and if so, in what computational form. Across three structural configurations and three model scales, we found a consistent ordering in surprisal: TRUE ANT < FALSE ANT < CORRUPTED. Structurally licensed

antecedents strongly facilitated prediction of *themselves*, while structurally illicit but feature-matching distractors produced weaker yet reliable interference.

These findings carry an important linguistic implication. The results are incompatible with a purely surface-based account in which plural distractors behave like genuine antecedents, but they are also inconsistent with a strict symbolic filtering account in which inaccessible competitors exert no influence. Instead, the observed profile supports a graded retrieval architecture in which structural accessibility functions as a privileged cue while semantic and similarity-based competitors retain partial influence. This pattern closely parallels psycholinguistic accounts of human sentence processing in which antecedent retrieval is *structure-sensitive* but not immune to interference. In this respect, the present results suggest that transformer-based language models may approximate certain functional properties of human dependency resolution, despite being trained without explicit grammatical rules.

Several limitations should be acknowledged. First, the present study relies on surprisal-based preference measures derived from next-token probabilities. While such measures provide a useful behavioral diagnostic, they do not directly reveal how structural information is represented internally. Second, although the results are consistent with hierarchical sensitivity, they do not by themselves demonstrate that models encode syntactic structure in the same representational format assumed in formal linguistic theory. Future work should therefore investigate the internal mechanisms underlying these effects. Representational probing, causal tracing, and circuit-level intervention methods may help determine where and how structural constraints emerge in model computations. It will also be important to extend the present paradigm to other dependency types, including Principle B/C phenomena, filler-gap dependencies, and negative polarity item licensing.

Overall, the present findings suggest that transformer language models do not implement binding as categorical symbolic rules, nor do they rely solely on shallow feature matching. Rather, reflexive prediction emerges from weighted competition in which structural constraints play a central but non-exclusive role. Understanding how such interactions emerge in these models remains an important challenge for future research.

References

Arnold, Jennifer E., Hannah C. Mayo, and Li Dong. 2021. My pronouns are *they/them*: Talking about pronouns changes how pronouns are understood. *Psychonomic Bulletin & Review* 28(5): 1688–1697. DOI: 10.3758/s13423-021-01905-0.

Chomsky, Noam. 1981. *Lectures on government and binding: The Pisa lectures. Vol.9*. Dordrecht: Foris.

Chomsky, Noam. 1986. *Barriers*: *linguistic inquiry monograph* 13. Cambridge, MA: The MIT Press.

Clifton, Charles, Shelia Kennison, and Jason Albrecht. 1997. Reading the words *her, his, him*: Implications for parsing principles based on frequency and on Structure. *Journal of Memory and Language* 36: 276–292. DOI: 10.1006/jmla.1996.2499.

Dillon Brian, Chow Wing-Yee, Wagers Matthew, Guo Taomei, Liu Fengqin, and Phillips Colin. 2014. The structure-sensitivity of memory access: Evidence from Mandarin Chinese. *Frontiers in Psychology* 5. https://api.semanticscholar.org/ CorpusID:10556465.

Dyke, Julie A. Van. 2007. Interference effects from grammatically unavailable constituents during sentence processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33(2): 407–430.

Garrod, Simon and Melody Terras. 2000. The contribution of lexical and situational knowledge to resolving discourse roles: Bonding and resolution. *Journal of Memory and Language* 42(4): 526–544. https://doi.org/10.1006/jmla.1999.2694.

Gordon C. Peter, Randall Hendrick, Marcus Johnson, and Yoonhyoung Lee. 2006. Similarity-based interference during language comprehension: Evidence from eye tracking during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 32(6):1304–1321.

Gulordava, Kristina, Piotr Bojanowski, Edouard, Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Volume 1 (Long Papers), 1195– 1205. New Orleans, LA: Association for Computational Linguistics. https://aclanthology.org/N18-1108/.

Hale, John. 2001. A probabilistic early parser as a psycholinguistic model. *Second Meeting of the North American Chapter of the Association for Computational Linguistics*. https://aclanthology.org/N01-1021/.

Levy, Roger. 2008. Expectation-based syntactic comprehension. *Cognition* 106(3): 1126–117.

Lewis, Richard and Shravan Vasishth. 2005. An activation-based model of sentence processing

as skilled memory retrieval. *Cognitive Science* 29(3): 375–419.

Linzen, Tal, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics* 4: 521–535. https://aclanthology.org/Q16-1037/.

Martin, Andrea and Brian Mcelree. 2009. Memory operations that support language comprehension: Evidence from verb-phrase ellipsis. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 35(5): 1231–1239.

Marvin, Rebecca and Tal Linzen. 2018. Targeted syntactic evaluation of Language Models. In Ellen Riloff (ed.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,* 1192–1202. Brussels: Belgium: Association for Computational Linguistics, https://aclanthology.org/D18-1151/.

McElree, Brian. 2000. Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguistic Research* 29(2): 111–123. https://api.semanticscholar.org/CorpusID:11290979.

McElree, Brian, Stephani Foraker, and Lisbeth Dyer. 2003. Memory structures that subserve sentence comprehension. *Journal of Memory and Language* 48(1): 67–91. https://www.sciencedirect.com/science/article/pii/S0749596X02005156.

Nicol, Janet and David Swinney. 1989. The role of structure in coreference during sentence comprehension. *Journal of Psycholinguistic Research* 18(1): 5–19.

Nicol, Janet and David Swinney. 2003. The psycholinguistics of anaphora. *Anaphora: A reference guide*. 72–104. Blackwell Publishing

Reinhart, Tanya and Eric Reuland. 1993. Reflexivity. *Linguistic Inquiry* 24(4): 657–720.

Reuland, Eric. 2011. *Anaphora and language design*. Cambridge, MA: The MIT Press.

Shannon, Claude Elwood. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27(3): 379–423.

Smith, Nathaniel J. and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition* 128(3): 302–319. https://www.sciencedirect. com/science/article/pii/S0010027713000413.

Sturt, Patrick. 2003. The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language* 48(3): 542–562. https://www.sciencedirect.com/science/article/pii/S0749596X02005363.

Van Dyke, Julie A. and Brian McElree. 2006. Retrieval interference in sentence comprehension. *Journal of Memory and Language* 55(2): 157–166. https://www.sciencedirect.com/science/article/pii/S0749596X0600043X.

Van Dyke, Julie A. and Brian McElree. 2011. Cue-dependent interference in comprehension. *Journal of Memory and Language* 65(3): 247–263. https://www.sciencedirect.com/science/article/pii/S0749596X11000581.

Jieun Kim
Professor
Department of English Language and Literature
University of Ulsan
93 Deahakro, Nam-gu,
Ulsan, 44920, Korea
E-mail: kimje@ulsan.ac.kr

Received: 2026. 03. 13.
Revised: 2026. 05. 14.
Accepted: 2026. 05. 22.